UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

UNIVERSITAT ROVIRA I VIRGILI
Departament de Bioquímica i Biotecnologia

In silico studies of the effect of phenolic compounds from grape seed extracts on the activity of phosphoinositide 3-kinase (PI3K) and the farnesoid X receptor (FXR)

Memòria presentada per optar al Grau de

Doctora per la Universitat Rovira i Virgili

Tarragona, novembre 2007

Vist i plau del Director de Tesi

Gerard Pujadas Anguiano

Departament de Bioquímica
i Biotecnologia

La interessada

Montserrat Vaqué Marquès

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

Aquest és un dels moments més emotius de la meva tesi. Escriure els agraïments és tan o més complicat que escriure un capítol ja que, en un text breu (ja us avanço que no serà pas tan breu), m'agradaria transmetre el meu agraïment a TOTS. I no en sou pocs els que heu fet possible que aquest projecte hagi arribat fins aquí. A més a més, el procés d'elaboració dels agraïments és una tasca delicada per a mi, perquè m'obliga a assumir que aquesta etapa ja es tanca i cal iniciar-ne una altra.

Han estat quatre anys força intensos i voldria recordar molts moments i anomenar-vos a tots, però ja sé que això és molt difícil. Tot va representar un gran canvi i ara en faig balanç, un balanç molt positiu. Els moments més complicats queden clarament superats pels bons i l'esforç sempre ha valgut la pena. He estat afortunada perquè, durant aquest temps, he tingut la oportunitat de conèixer moltíssima gent i, sense cap mena de dubte, grans persones.

Per tant, en primer lloc, voldria agrair al Gerard Pujadas la confiança, la paciència, el recolzament, l'ajuda i el temps dedicat que han fet possible embarcar-me en aquest projecte i veure'l ja complert. Gràcies per tot el que m'has ensenyat. També vull donar les gràcies a tots els membres del departament, especialment a l'Anna Ardévol (amb qui he pogut compartit llargues converses i rialles camí de casa), a la Cinta, a la Pepa, a la Mayte, al Juan , a la Isabel, al Santi, a l'Anton i al Lluís pel recolzament que sempre m'han donat.

També vull expressar el meu agraïment a la Íngrid Bàrcena i l'Alfred Gil del Servei de Disseny de Fàrmacs del CESCA, que han donat solució i han permès executar una part important d'aquest treball, així com també al Carles Aliagas del Departament d'Enginyeria Informàtica i Matemàtiques per la seva col·laboració en la creació de la interfície gràfica del BDT i al John Bates, per la correcció del treball.

Vull donar les gràcies als que hem van veure començar, sobretot a la Montse Pinent, amb qui he tingut la sort de compartir molts moments i m'ha donat grans consells (alguns dels qual no he sabut seguir), i al Josep, que sempre han estat un referent per a mi. També vull recordar a la Montse Vadillo, la Vanessa, el Cesc i el Nino.

Després ja es va anar configurant el grup dels bioinformàtics (o biocomputing com ens vàrem anomenar nosaltres) el record dels quals resta caricaturitzat en el lab 111 (ja per poc temps). Hi són el Pere i l'Albert, els qui m'han patit més, els qui he turmentat més amb les meves preguntes i als qui vull donar unes molt sinceres GRÀCIES per tot i, sobretot, per ser com sou (crec que us enyoraré). Juntament amb el Pep i l'Eduard, heu fet possible que sempre hi hagués un bon ambient en el laboratori. I si bé en aquest dibuix hi manca la Marina (encara no hi era), aprofito ara per agrair-li tot el que ha fet i continua fent per mi (entre d'altres coses haver estat la meva companyia durant molts caps de setmana). Per a ella tinc la més sincera abraçada. Per la seva amistat i ajuda, vull recordar-me del Gerard i la Safae, també de la Laura, recentment incorporada, i de la Núria. I, entre els bioinformàtics, no m'oblido de l'Esther, però és que volia agrair-li més especialment la seva paciència (sobretot pels darrers mesos de tesis), la seva ajuda i sobretot l'amistat. Gràcies per haver compartit amb mi viatges, conferències, cursos, rialles, nervis i els darrers mesos d'agost a la facultat. Et transmeto els ànims que m'has donat tu dia a dia.

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

A part del grup de bioinformàtics (i sense ser menys importants), vull donar les gràcies a tots els que us poseu la bata blanca (bé, encara que sovint això no és així). Gràcies Sabina, Helena, David (crec que formeu un magnífic equip) i també, Ximena, Isabel, Mario, Niurka i Anna. Com comprendreu a la Gemma vull dedicar-li un especial agraïment per la seva amistat i per voler compartir amb mi molts moments, petits projectes de recerca, viatges, classes d'anglès, mesos d'agost i molts i molts vespres a la facultat. I en aquest grup d'experimental també vull incloure a la Lídia, a qui pels consells químics i per la seva amistat dono les gràcies. I no m'oblido dels companys d'enologia, especialment la Rosana.

A tots us vull agrair els moments passats dins i fora del laboratori. Per a mi ha estat molt important poder compartir la feina, sopars, calçotades, congressos i el dinar de cada dia amb persones com vosaltres. Ànims i mercès per tot. A Falset hi teniu casa.

També vull donar les gràcies a la secretaria de Bioquímica, especialment, a la Cristina (en els primers anys) i a la Maribel per la seva ajuda. També agrair-li a la Mari la paciència de sant que ha tingut amb mi des del primer dia i al Santiago el seu ajut en les pràctiques.

I ja fora de la facultat, vull agrair als amics, a la gent de Falset que s'ha preocupat per mi i també a la Sílvia per les seves mans i atenció.

A tota la meva família us vull agrair la comprensió i el recolzament donat. Tots heu patit la tesi, des dels més petits, la Júlia i el Pau (que poc m'han vist últimament), fins als més grans, l'avi Josep i l'àvia Maria (que ja tenen ganes de veure'm més tranquil·la). A la Padrina Montserrat, per estar sempre pendent de mi, a tots els meus cosins i tiets (que no en sou pocs i comprendreu que no us anomeni), als meus sogres, cunyats i cunyades que durant els quatre anys han estat al meu costat us dono GRÀCIES per tot.

M'agardaria donar les gràcies als meus germans, Jordi, Núria i Maria, per la paciència que tenen amb mi i per estar al meu costat i ajudar-me (entre d'altres coses al disseny de la tesi).

I als meus pares, a qui voldria dedicar el més sincer dels agraïments per fer possible que hagi arribat fins aquí. Gràcies per estar sempre pendents de les meves decisions i recolzar-me. I per ensenyar-me que amb esforç es pot aconseguir el que et proposes.

Per últim, guardo una menció especial per a qui he fet patir més, perquè m'estima i l'estimo, per a qui ha compartit i comparteix amb mi el meu entusiasme i la il·lusió en al meva recerca, per a qui ha col·laborat amb tan d'esforç i és un dels responsables que això comencés i ara s'acabi. Jordi, GRÀCIES.

Diuen que tot esforç té la seva recompensa. Jo crec que és veritat i, per això, m'agradaria compensar-vos tot i saber que és difícil. Compteu amb mi sempre que em necessiteu.

A TOTS, MOLTES GRÀCIES

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

Als meus pares

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

ABBREVIATION LIST

ADT: AutoDockTools

CDCA: Chenodeoxycholic acid

ECDCA: Ethyl-chenodeoxycholic acid

EGCG: Epigallocatechin gallate

FEX: Fexaramine

FXR: Farnesoid X Receptor

GA: Genetic Algorithm

GPCR: Coupled Receptor Activation

GRIP-1: Glucocorticoid Receptor Interacting Protein-1

GSPE: Grape-Seed Procyanidin Extract

Gβγ: βγ subunits of heterotrimeric G-proteins

HTS: High-Throughput Screening

IR: Insulin Receptor

IRS: Insulin Receptor Substrate

LBD: Ligand Binding Domain

LDL: Low Density Lipoprotein

LXR: Liver X Receptor

MAPK: Mitogen-Activated Protein Kinase

MC: Monte Carlo

MVD: Molegro Virtual Docker

NR: Nuclear Receptor

PDB: Protein Data Bank

PDGFR: Platelet-Derived Growth Factor Receptor

PDK-1: Phosphoinositide-Dependent Protein Kinase-1

PI3K: Phosphoinositide 3-kinase

PKA: Protein Kinase A

PKB/Akt: Protein Kinase B

PKC: Protein Kinase C

PLC: Phospholipase C

PLS: Partial Least-Square

PPAR: Peroxisome Proliferator Activated Receptor

PtdIns: Phosphatidylinositol

PTEN: Phosphatase and tensin homolog

PX: Phox Homology

PXR: Pregnane X Receptor

ISBN:978-84-691-1553-4 /DL: T.151-2008

QSAR: Quantitative Structure Relationship

RMS: Root-Mean Square

RMSD: Root-Mean Square Deviation

RMSE: Root Mean-Square Error

ROR: Retinoic related orphan receptor

RXR: Retinoic X Receptor

SAR: Structure Activity Relationship

SD: Standard Deviation

SHIP2: Src Homology 2 Domain-Containing Tyrosine Phosphatase

SHP: Small Heterodimer Partner

TG: Triglycerides

TS: Tabu Search

VHTS: Virtual High-Throughput Screening

VLS: Virtual Ligand Screening.

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

Index

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

Context and goals of the work

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

The influence of nutrition on our daily lives has been one of the main driving forces behind the development of new genomic technologies which are used, among other applications, to improve the processing, the safety, the quality and the health-derived benefits of food. Moreover it is now the age of nutrigenomics (*i.e.* a science whose goal is to understand, at the molecular level, how nutrients improve health conditions and prevent diseases) and nutrigenetics (*i.e.* a science whose goal is to understand how the genetic makeup of individuals coordinates their response to diet). Both sciences have developed out of recent research that has shown: (a) the influence of nutrition on gene expression [1-3]; and (b) the significant effect of genetic variation on food intake, metabolic response to food, individual nutrient requirements and the efficacy of disease-protective dietary factors [4, 5]. Therefore, foods can be developed to be specific to individual human genotypes, can benefit health and can enhance normal physiological processes. For this reason, it has been suggested that food may become the pharmaceutical products of the future. In order to achieve this goal, *in silico* approaches can be useful for analyzing the intermolecular interactions between nutrients and proteins or genes and how these interactions modulate the corresponding target function. Despite recent contributions in this field [6], computational methods developed for molecular design and simulation (*e.g.* pharmacophore generation and protein-ligand docking) which are routinely applied to develop drugs in the pharmaceutical industry [7] are still not frequently used in nutrigenomics or nutrigenetics. From our point of view, the *in silico* tools and classical *in vivo* and *in vitro* methods must be applied together on natural products in order to study their efficacy in the development of functional foods (*i.e.* food that has health benefits beyond the traditional nutrients it contains). The reason is that these computational methodologies will allow us to correlate a compound's biological activity with structural information for deriving 3D quantitative structure-activity relationships (3D-QSAR) and gain insights into the ligand-structural requirements for increased target affinity and/or selectivity. Therefore, the application of such *in silico* approaches is a new event in these nutritional sciences which may help to: (a) understand how bioactive molecules in food improve health conditions and prevent diseases like diabetes, obesity, cardiovascular pathologies and cancer; and (b) predict which non-experimentally tested phytochemistry ligands would be most effective against a pre-defined protein or gene target [6].

The natural products that can be added to food during the production process in order to convert it into functional food include complex extracts and their chemical entities, which are biosynthesized in nature. In general, the effects of these natural extracts have been described as being beneficial for human health [8]. For instance, the phenolic compounds that are commonly found in fruits and vegetables [9-11] are described as cardioprotective, antioxidant, antigenotoxic, anti-inflammatory and anticarcinogenic agents [12-15]. Therefore, there are

ISBN:978-84-691-1553-4 /DL: T.151-2008

several metabolic processes (related to such pathologies as obesity, insulin resistance and diabetes) that are targets for the effects of these phenolic compounds. These processes include the modulation of glucose and cholesterol metabolism, and changes in the lipid plasma profile. Some evidence suggests that the cellular effects of phenolic compounds may be mediated by their interaction with specific proteins that are central to intracellular signalling cascades [16]. Nevertheless, the exact relationship between individual phenolic compounds and diseases such as insulin resistance or diabetes has not been elucidated yet. Thus, in the modulation of the glucose metabolism, several of our *in vivo* and *in vitro* results suggest that grape seed extracts help to stimulate glucose uptake (which is critical point for maintaining glucose homeostasis) [17]. In contrast, it is also well known that whereas some phenolic compounds (*i.e.* catechin and gallic acid) do not affect glucose uptake, others (*i.e.* quercetin, myricetin, catechin-gallate, resveratrol and naringenin) reversibly inhibit it in isolated rat adipocytes, 3T3-L1 adipocytes, rat L6 myotubes and human muscle-derived cell lines [18-20]. Hence, it has been shown that this inhibition in insulin-induced glucose uptake is the result of the competition of these phenolic compounds with ATP for the ATP-binding site of the catalytic subunit of phosphoinositide 3-kinase (PI3K) [18]. Moreover, recent studies with isoform-specific inhibitors have identified p110α (the catalytic subunit of PI3Kα) as the crucial isoform for the insulin-stimulated glucose-uptake in such cell lines as 3T3-L1 adipocytes, L6 myotubes and CHO-IR [21, 22]. For this reason, it seems that the potential of the different phenolic compounds for inhibiting p110α must be predicted to prevent the possible side effects associated with their consumption. Furthermore, it is important to predict the inhibitory activity not only of the phenolic compounds that are frequently found in plant extracts [23], but also of their derived bioactive compounds detected in plasma or urine [9, 24-31]. In order to predict the inhibitory potential of all these phenolic compounds on p110α we introduce and apply *in silico* methodologies (*i.e.* docking and 3D-QSAR). Therefore, this study expects to evaluate and describe for the first time the effect of a huge number of natural compounds and their derived bioactive molecules on the inhibition of p110α and, therefore, its effect on insulin-stimulated glucose-uptake.

Phenolic compounds are also involved in mechanisms that modify transcriptional activities, the control of which is very complex and not well understood. Some of our recent results have demonstrated that at least some of the phenolic compounds present in grape seed procyanidins extract increase the activity of the Farnesoid X Receptor (FXR; a nuclear receptor involved in the bile acid metabolism and in the control of cholesterol and triglyceride metabolism) in a dose-dependent way when the natural ligand of FXR is also present [32]. Therefore, taking into account these experimental results, a computational analysis will be done in an attempt to predict how phenolic compounds activate FXR. In this respect, *in silico* methodologies (*i.e.* docking and 3D-QSAR) together with the biological data of synthetic non-steroidal FXR

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

agonists that bind on a binding site that is close but different to that of the natural ligand will be used to suggest a mechanism of FXR activation by means of most phenolic compounds present in vegetable extracts and their derived bioactive forms.

In summary, the aim of this PhD is to use *in silico* tools to study the effect of the phenolic compounds that are most frequently found in vegetal extracts and their bioactive forms on: (a) insulin-stimulated glucose-uptake; and (b) the activation of the FXR nuclear receptor.

This work has been supported by a fellowship from grant number CO3/O8 of the Fondo de Investigación Sanitaria (FIS) and by grant number AGL2005-04889 from the Ministerio de Educación y Ciencia of the Spanish Government. The research has been performed in the Nutrigenomics Research Group laboratory of the Biochemistry and Biotechnology Departament from the Rovira i Virgili University.

ISBN:978-84-691-1553-4 /DL: T.151-2008

References

[1] H. Al-Hasani, H.G. Joost, Nutrition-/diet-induced changes in gene expression in white adipose tissue., Best Pract Res Clin Endocrinol Metab 19 (2005) 589-603.

[2] B. de Mateo Silleras, A. Miján de la Torre, Nutrition and gene expression, Nutr Hosp 15 Suppl 1 (2000) 128-142.

[3] I.R. Sanderson, Nutrition and gene expression., Nestle Nutr Workshop Ser Clin Perform Programme 2 (1999) 121-137.

[4] L. Afman, M. Müller, Nutrigenomics: from molecular nutrition to prevention of disease., J Am Diet Assoc 106 (2006) 569-576.

[5] M. Müller, S. Kersten, Nutrigenomics: goals and strategies., Nat Rev Genet 4 (2003) 315-322.

[6] J.M. Rollinger, T. Langer, H. Stuppner, Integrated in silico tools for exploiting the natural products' bioactivity., Planta Med 72 (2006) 671-678.

[7] T. Langer, R.D. Hoffmann, Pharmacophores and Pharmacophore Searches, in: W. WILEY-VCH Verlag GmbH & Co. KGaA, Germany (Ed.), vol. 32, 2006.

[8] D.J. Newman, G.M. Cragg, K.M. Snader, The influence of natural products upon drug discovery., Nat Prod Rep 17 (2000) 215-234.

[9] A. Scalbert, G. Williamson, Dietary intake and bioavailability of polyphenols., J Nutr 130 (2000) 2073S-2085S.

[10] A. Fleuriet, J.J. Macheix, Phenolic Acids in Fruits and Vegetables, Flavonoids in Health and Disease, 2003, pp. 1-42.

[11] Documentation for the Update of the USDA Database for Flavonoid Content of Selected foods, Release 2.1 (2007).

[12] M. Pinent, C. Bladé, M.J. Salvadó, M. Blay, G. Pujadas, J. Fernández-Larrea, L. Arola, A. Ardévol, Procyanidin effects on adipocyte-related pathologies., Crit Rev Food Sci Nutr 46 (2006) 543-550.

[13] Y. Yilmaz, R.T. Toledo, Major flavonoids in grape seeds and skins: antioxidant capacity of catechin, epicatechin, and gallic acid., J Agric Food Chem 52 (2004) 255-260.

[14] J.A. Ross, C.M. Kasum, Dietary flavonoids: bioavailability, metabolic effects, and safety., Annu Rev Nutr 22 (2002) 19-34.

[15] P.M. Kris-Etherton, K.D. Hecker, A. Bonanome, S.M. Coval, A.E. Binkoski, K.F. Hilpert, A.E. Griel, T.D. Etherton, Bioactive compounds in foods: their role in the prevention of cardiovascular disease and cancer., Am J Med 113 Suppl 9B (2002) 71S-88S.

[16] H. Schroeter, C. Boyd, J.P. Spencer, R.J. Williams, E. Cadenas, C. Rice-Evans, MAPK signaling in neurodegeneration: influences of flavonoids and of nitric oxide., Neurobiol Aging 23 (2002) 861-880.

[17] M. Pinent, M. Blay, M.C. Bladé, M.J. Salvadó, L. Arola, A. Ardévol, Grape seed-derived procyanidins have an antihyperglycemic effect in streptozotocin-induced diabetic rats and insulinomimetic activity in insulin-sensitive cell lines., Endocrinology 145 (2004) 4985-4990.

[18] S. Fröjdö, D. Cozzone, H. Vidal, L. Pirola, Resveratrol is a class IA phosphoinositide 3-kinase inhibitor., Biochem J (2007).

[19] A.W. Harmon, Y.M. Patel, Naringenin inhibits phosphoinositide 3-kinase activity and glucose uptake in 3T3-L1 adipocytes., Biochem Biophys Res Commun 305 (2003) 229-234.

[20] P. Strobel, C. Allard, T. Perez-Acle, R. Calderon, R. Aldunate, F. Leighton, Myricetin, quercetin and catechin-gallate inhibit glucose uptake in isolated rat adipocytes., Biochem J 386 (2005) 471-478.

[21] C. Chaussade, G.W. Rewcastle, J.D. Kendall, W.A. Denny, K. Cho, L.M. Grønning, M.L. Chong, S.H. Anagnostou, S.P. Jackson, N. Daniele, P.R. Shepherd, Evidence for

functional redundancy of class IA PI3K isoforms in insulin signalling., Biochem J 404 (2007) 449-458.

[22] Z.A. Knight, B. Gonzalez, M.E. Feldman, E.R. Zunder, D.D. Goldenberg, O. Williams, R. Loewith, D. Stokoe, A. Balla, B. Toth, T. Balla, W.A. Weiss, R.L. Williams, K.M. Shokat, A pharmacological map of the PI3-K family defines a role for p110alpha in insulin signaling., Cell 125 (2006) 733-747.

[23] J.J. Macheix, A. Fleuriet, J. Billot, Fruit Phenolics, Boca Raton, FL: CRC Press 1990.

[24] C. Felgines, S. Talavéra, M.P. Gonthier, O. Texier, A. Scalbert, J.L. Lamaison, C. Rémésy, Strawberry anthocyanins are recovered in urine as glucuro- and sulfoconjugates in humans., J Nutr 133 (2003) 1296-1301.

[25] C. Felgines, S. Talavera, O. Texier, A. Gil-Izquierdo, J.L. Lamaison, C. Remesy, Blackberry anthocyanins are mainly recovered from urine as methylated and glucuronidated conjugates in humans., J Agric Food Chem 53 (2005) 7721-7727.

[26] C. Manach, G. Williamson, C. Morand, A. Scalbert, C. Rémésy, Bioavailability and bioefficacy of polyphenols in humans. I. Review of 97 bioavailability studies., Am J Clin Nutr 81 (2005) 230S-242S.

[27] C. Tsang, C. Auger, W. Mullen, A. Bornet, J.M. Rouanet, A. Crozier, P.L. Teissedre, The absorption, metabolism and excretion of flavan-3-ols and procyanidins following the ingestion of a grape seed extract by rats., Br J Nutr 94 (2005) 170-181.

[28] S.A. Aherne, N.M. O'Brien, Dietary flavonols: chemistry, food content, and metabolism., Nutrition 18 (2002) 75-81.

[29] C. Manach, A. Scalbert, C. Morand, C. Rémésy, L. Jiménez, Polyphenols: food sources and bioavailability., Am J Clin Nutr 79 (2004) 727-747.

[30] G. Williamson, C. Manach, Bioavailability and bioefficacy of polyphenols in humans. II. Review of 93 intervention studies., Am J Clin Nutr 81 (2005) 243S-255S.

[31] F. Saura-Calixtoa, J. Serrano, I. Goñi, Intake and bioaccessibility of total polyphenols in a whole diet, Food Chemistry 101 (2007) 492.

[32] J.M. Del Bas, Modulation of hepatic lipoprotein metabolism by dietary procyanidins, Departament de Bioquímica i Biotecnologia, Universitat Rovira i Virgili, Tarragona, 2007, p. 218.

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

Phenolic compounds

I

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

Brief introduction

Molecules with a polyphenol structure (*i.e.* several hydroxyl groups on aromatic rings) are produced by plants to develop physiological functions that are different from those of primary metabolites such as carbohydrates, proteins and lipids. For instance, they can play the role of secondary metabolites that are used by plants to establish symbiotic interactions with other organisms, as well as other types of communication with their environment. They are often produced in response to environmental stresses caused by diseases, insects, climate, ultraviolet radiation, etc. [1, 2]. Thousands of these phenolic molecules have been identified in plants (and several hundreds of them are present in edible ones) and some of them are often characteristic of a plant species or even of a particular organ or tissue of that plant. Moreover, the phenolic component of a plant depends on the cultivar, growing location, agricultural practices, processing method, storage conditions and preparation. Together with vitamins and minerals, phenolic compounds are considered to be the health-promoting factors in fruit and vegetables [3, 4].

In most cases, these phenolic compounds are not found in a free state in plants. For example, phenolic acids are usually found esterified to sugars, organic acids and lipids (except those found in trace) because they are accumulated in vacuoles or linked to cell wall components. In fact, flavonoids (except catechins) are not present in plants as aglycones (*i.e.* non-glycosilated forms). They form, instead, glycoside derivatives to increase the polarity of the flavonoid molecule, which allows them to be stored in plant cell vacuoles. Moreover, when glycosides are formed: (a) there is a preferred glycosylation site (which is on the C-3 position and, less frequently, on the C-7 position of the flavonoid molecule); and (b) the most usual sugar found is D-glucose (although such other carbohydrate substituents as arabinose, galactose, glucorhamnose, lignin, L-rhamnose, and xylose are also found [5, 6]).

A large part of the human diet consists of plant-derived products, such as vegetables, fruits and tea. Therefore, a diet rich in fruits and vegetables contains high levels of polyphenols. The total average intake of polyphenols in a healthy human diet has been estimated to be around 1 g per day [6-8]. However, this is an approximation because these studies are mostly based on the content of only a few phenolic compounds and foods.

ISBN:978-84-691-1553-4 /DL: T.151-2008

Classification

Polyphenols are a wide variety of compounds, which are classified into different groups according to: (a) the number of phenol rings they contain; and (b) the structural elements connecting these rings. Therefore, the main classes of polyphenols are *phenolic acids*, *stilbenes*, *lignans* and *flavonoids* [9]. Figure 1 shows the basic structure of these four classes of polyphenols.

— ***Phenolic acids*** are acidic compounds because their structure contains one carboxylic group. They can be divided into two different subclasses, hydroxybenzoic acids and hydroxycinnamic acids, derived from two non-phenolic compounds, the benzoic and the cinnamic acid, respectively. Phenolic acids are common in plants and their distribution can depend heavily on species, cultivar conditions and physiological stage. They clearly play a role both in the interactions between the plant and its biotic or abiotic environment and in the organoleptic and nutritional qualities of fruits, vegetables and derived products [10].

— ***Stilbenes*** are polyphenolic compounds that are present only in low quantities in the human diet because they are found either in plants that are not routinely consumed as food or in non-edible tissues. The major dietary sources of stilbenes are grapes, grape juices, wine, peanuts and peanut butter. Stilbenes are 1,2-diarylethenes, which contain an A ring that usually carries two hydroxyl groups in the m-position, while the so called B ring is substituted by hydroxy and methoxy groups in the o-, m- and/or p-position (see Figure 1) [11].

— ***Lignans*** are phenolic compounds that are ubiquitously distributed in plants but which are only found in low quantities in food. They consist of two phenylpropane units (see Figure 1) and they may be the source of phytoestrogens in plant-rich diets (*e.g.* vegetarians) [11].

— ***Flavonoids*** are a large class of polyphenolic compounds widely distributed among plants and they are principally found in fruits, vegetables, medicinal plants and popular drinks, such as red wine, tea and beer (http://www.ars.usda.gov/Services/docs.htm?docid=6231 [4]). They are the most abundant polyphenols in the human diet [2, 12-14]. They share a common structure (known as the flavonoid nucleus; see Figure 1) that consists of two aromatic rings (A and B), each of which contains at least one hydroxyl group. These rings are bound together by three carbon atoms and an oxygen atom to form an oxygenated heterocycle (ring C) [6]. They are further divided into subclasses based on: (a) the connection of one of the aromatic rings to the heterocyclic ring; (b) the oxidation

state and functional groups of the heterocyclic ring (their oxidation level on the C-ring); and (c) the degree of polymerization. Therefore, according to their chemical structure, the six main subclasses of flavonoids are *flavonols*, *flavanones*, *flavones*, *anthocyanidins*, *isoflavones* and *flavanols* (*i.e.* catechins) (see Figure 1). In addition to this diversity, they can be hydroxylated, methoxylated, glycosylated and acetylated [6]. Thus, the flavonoids with sugar moieties are called flavonoids glycoside and the form in which there is no attached sugar is called aglygone [15]. So, the biological activities of flavonoids and their metabolites depend on their chemical structure and relative orientation of various moieties on the molecule. *Flavanols* are one of the most abundant subclasses of flavonoids, often referred to as flavan-3-ol or catechins. They are present as monomers, oligomers and polymers and are often esterified with gallic acid [6]. The structure of flavanols is based on the flavonoid nucleus where ring C is a heterocyclic pyran. Procyanidins are flavanols that contain dimers and oligomers of up to 10 units whereas larger polymer structures are classified as condensed tannins [6, 16]. Figure 1 shows the monomeric structure and how oligomeric structures are built.

Bioavailability and metabolism

In some cases, the term *bioavailability* refers to the proportion of a nutrient or bioactive component that is absorbed from the gastrointestinal tract. In others, it refers to the metabolism, excretion and utilization of these metabolites and the measure of their efficacy [8]. In the past few years, our knowledge of polyphenol absorption, metabolism and excretion has increased considerably (see Figure 2). Thus, it has been shown that the bioactive form of the polyphenols absorbed by humans differs from that found in the vegetables in the diet, such as the glycoside derivatives [6-9, 17, 18]. Therefore, knowing how the phenolic compounds are present in a food and in what quantities is not as important as knowing how much of the compound is bioavailable.

Although the exact mechanisms involved in flavonoid absorption are not clear, studies with humans have shown that most of them are absorbed in the small intestine [7, 19, 20]. Thus, it has been suggested that aglicone flavonoids are absorbed via passive difusion across membranes in the small intestine because of their hydrophobic nature [21]. However, most polyphenols are present in food in the form of esters, glycosides or polymers, which cannot be absorbed in their native form. Therefore, to explain how flavonoid glycosides move into the enterocyte, two mechanisms have been proposed. In the first, the flavonoid glycosides penetrate as an intact structure by means of a sodium-glucose co-transporter (active transport). If this is

the case, further evidence is required to determine the possible role of sugar transporters in flavonoid glycoside uptake. Furthermore, glycosilated polyphenols are generally not found in plasma, urine and tissues. The second mechanism proposed is an extracellular hydrolysis of the glycoside. According to this theory, phenolic compounds are first hydrolized by intestinal enzymes or by the colonic microflora so that they can then be absorbed. Therefore, deglycosylation is required for uptake in the small intestine by means of enzymes with glycosidase activity that enable the aglycone structure to be passively diffused by means of the previously described mechanism [17, 20]. Finally, the remaining fraction of flavonoids that has not been absorbed in the small intestine is metabolized (*i.e.* deglycosylated) under bacterial metabolism in the lower intestine and subjected to transport or further metabolization [7, 17]. This deconjugation and degradation can produce phenolic acids that are found in urine and plasma after exposure to oral polyphenols [17]. Recently, a new pathway of polyphenol absorption has been proposed that involves transportation through the lymph system [22].

The bioavailability of polyphenol compounds also depends on the metabolism during and after their absorption. The metabolization of most phenolic compounds occurs intracellularly either in the small intestine (in the enterocytes by the oxidative and P450-related metabolism) or in the liver (in the hepatocytes by phase II enzymes, UDP-glucuronosyl transferases, sulfotransferases and catechol-O-methyl transferases). In the enterocytes and hepatocytes, flavonoids and their derivatives may undergo reactions such as hydroxylations, methylations, and conjugation reactions with glucuronic acid and/or sulfate. Some *in vitro* studies have shown that glucuronidation can increase or decrease the biological activity of individual polyphenols [23]. Moreover, the resulting plasma metabolites of polyphenols are more hydrophilic than polyphenol aglycones and, therefore, are easily eliminated through bile and urine.

Another point to take into account when evaluating the bioavailability of polyphenols is the level of absorption in the tissues. Although this is conditioned by their hydrophilic nature (*i.e.* they are water soluble due to the glucuronide and sulfate moieties), which makes it difficult for them to pass through membranes, several studies have found these molecules in such tissues as the liver, kidneys, lungs, brain, pancreas and bladder. In this respect, determining the bioavailability of these metabolites in the different tissues may be much more important than knowing their plasma concentrations [9, 24].

Furthermore, in biovailability studies determining the rate of absorption is also important. It can be estimated *in vitro* by using intestinal cell monolayers that simulate the digestive processes. These rates are very diverse and can be important to explain the effect exerted by polyphenolic compounds [25].

The polar structure of these compounds makes it possible for them to be excreted in the urine and bile. Thus, when they are excreted in the bile, they enter the duodenum and are then metabolized by the intestinal microflora. The metabolites resulting from this degradation (fragmentation, hydrolisis of glucurono- or sulfoconjugates) can be reabsorbed and enter the enterohepatic circulation. Finally, it is worth mentioning that the degree of biliary excretion depends on the substitution on the phenolic molecule, degree of polarity, and molecular weight [26].

In conclusion, the phenolic compounds that can cross the intestinal barrier and which are the result of digestive and hepatic activity, are present in blood with a structure that usually differs from that of the native compounds. These glucoronidated, methylated or sulfated structures are usually the active metabolites. The most abundant polyphenols in our diet, then, are not necessarily those that have the best bioavailability profile, so it is essential to know the bioavailability of polyphenols relative to tissue targets because it will help us: (a) to identify those that are most likely to exert protective health effects; and (b) to understand their potential actions *in vivo*. Table 1 shows the bioavailable structures from phenolic compounds that have been detected in plasma or urine. It also shows the metabolites produced by the intestinal microflora, which can also contribute to the biological effects of the phenolic compounds in humans. According to the data, the polyphenols that are absorbed best in humans are isoflavones and gallic acid, followed by catechins, flavanones, and quercetin glucosides. The polyphenols that are least well absorbed are procyanidins, galloylated tea catechins and anthocyanins [17]. At this point, it is worth mentioning that the analysis of the bioavailability of the active forms of phenolic compounds is severely hampered by the technical limitations of their determination.

Effects of phenolic compounds on health

Polyphenols are defined as bioactive compounds because they influence physiological or cellular activities and, as a consequence, have a beneficial effect on health. They have been considered to be cardioprotective, antiinflammatory, anticarcinogenic and antimutagenic, among others [27-29]. These protective effects are related to their capacity to: (a) act as free radical scavengers; (b) chelate metals; (c) activate antioxidant enzymes; (d) reduce α-tocopherol radicals; and (e) inhibit oxidases in biological systems [16]. In this respect, the most beneficial health effects attributed to flavonoids are their antioxidant and chelating abilities [28]. When phenolic compounds act as free radical scavengers, they protect the cell against oxidative stress and free radical-induced damage in membranes and nucleic acids [6]. In fact, antioxidant activity is probably the most extensively studied aspect of the bioactivity of phenolic

ISBN:978-84-691-1553-4 /DL: T.151-2008

compounds. It depends on the configuration and the total number of hydroxyl groups in their structures, which make them highly susceptible to oxidation [16, 30]. Several studies have attributed the cardiovascular protection to mechanisms that are not related to their antioxidant capacity. These mechanisms are related to alterations in cell membrane receptors, interactions with specific proteins of intracellular signalling cascades and modulation of gene expression [31, 32]. Several of the effects of phenolic compounds may also be related to their ability to bind to the ATP-binding site of various proteins (*e.g.* transport ATPases and protein kinases). Thus, they can inhibit or stimulate specific pathways by modifing the phosphorylation state of their target molecules [33].

Epidemiological studies have suggested that a high dietary intake of selected polyphenols can protect against the development of such human diseases as cardiovascular diseases and cancer [30]. For this reason, animal and *in vitro* models have been used to identify the mechanism of action of these compounds in these pathologies. Despite the evidence for the protective effect of phenolic compound extracts (*e.g.* protection against cardiovascular disease and oxidative damage or amelioration of insulin resistance states), the exact polyphenol molecules which are responsible for their beneficial effects *in vivo* are still unknown. In this sense, the wide structural variety of these compounds (*e.g.* the diversity in the degree of polymerization, the number of isomers, the conjugation pattern, etc.) and the limitations of the techniques used to study their bioavailability make it difficult to ascribe the observed effects to a specific molecule. This lack of information can explain the contradictions found in several studies about the effects of phenolic compounds and, for instance, their pro-oxidant activity [34]. What is more, procyanidins may not need to be efficiently absorbed through the gut to have a beneficial effect on health. They may have direct effects on the intestinal mucosa and protect it against oxidative stress or the actions of carcinogens [17].

The protective properties of these compounds mean that it is important to study the principal mechanisms of action of selected polyphenols. Furthermore, advances in this research may lead to the development of nutritional products (*i.e.* food supplements) and semisynthetic analogs that retain substantial protective capacity but produce minimal adverse side effects.

Effects of polyphenols on intracellular signalling pathways

Several metabolic processes (related to such pathologies as obesity, insulin resistance and diabetes) are targets for the effects of phenolic compounds. These processes include the modulation of glucose and cholesterol metabolism, and changes in the lipid plasma profile.

As already mentioned, there is evidence to suggest that the cellular effects of phenolic compounds may be mediated by their interaction with specific proteins that are central to intracellular signalling cascades [35]. In fact, the structure of flavonoids favours their binding to the ATP binding site of a large number of proteins. Previous studies have shown that phenolic compounds can interact such components of signalling pathways as protein kinases [*e.g.* phosphatidylinositol 3-kinase (PI3K), Akt/PKB, tyrosine kinase C (PKC), and MAP kinases] [32, 36-38]. Therefore, the effect of the flavonoids can be achieved by modulating the phosphorylation state of target molecules (which, among other effects, can result in the modulation of the gene expression) and, in consequence, inhibiting or stimulating pathways/mechanisms that affect the cellular function. Therefore, this modulation can have a beneficial effect on such pathologies as cancer, proliferative diseases and inflammation [31].

Phenolic compounds have also been described as antihyperglycemic agents in diabetic rats [39]. In insulin-sensitive cell lines (*i.e*. 3T3-L1 adipocytes and L6E9 myotubes), grape seed procyanidins have an insulin-like effect. They have been observed to enhance the glucose uptake through mediators of the insulin-signalling pathways, such as PI3K and p38 MAPK activation and GLUT-4 translocation. The reduction in glycemia (blood glucose levels) caused by phenolic compounds has been attributed to such actions as a reduction in the absorption of nutrients (*i.e.* tea catechins inhibit intestinal glucose absorption [40]), reduction in food intake (*i.e.* green tea epigallocatechin gallate significantly reduces food intake [41]), induction of β cell regeneration [42] and a direct action on adipose cells that enhances insulin activity [43]. In this respect, the mechanism of action depends on the structure of the compounds. Although *in vivo* physiological changes induced by polyphenols are well described, the molecular mechanisms used by these phenolic compounds to exert these changes are not absolutly clear. In conclusion, the exact relationship between individual phenolic compounds and insulin resistance or diabetes has yet to be elucidated.

Polyphenols not only have antioxidant activity that has cardioprotective effects, they are also involved in processes such as the inhibition of platelet aggregation, antinflammatory mechanisms, vasorelaxing activity and modulation of lipid metabolism and plasma lipid profile. In this respect, some phenolic compounds have been reported to reduce plasma lipids and

atherogenic lipoproteins (mainly LDL and chylomicron remnants) in animal models [44]. *In vivo* studies have shown that red wine polyphenols can reduce plasma LDL concentrations, apolipoprotein B, triglycerides and cholesterol [45, 46]. The reduction in plasma lipids may be related to the reduction of fat intake by the inhibition of lipases (*i.e.* pancreatic lipase, the enzyme for dietary triacylglycerols digestion) or the reduction in lipoprotein lipase [47]. Furthermore, the reduction in the absorption of cholesterol has been related to the ability of polyphenols to interact with the ATP binding site in enterocyte cholesterol transporters [48]. Therefore, the LDL cholesterol reduction and hypocholesterolemic effect have been shown to be a preventive atherosclerosis mechanism of phenolic compounds [44, 49].

Phenolic compounds also modify enzymatic and transcriptional activities. The control mechanisms at the transcriptional level are very complex and not well understood. However, the ability of nuclear receptors to modulate a wide battery of genes reveals them to be targets in the treatment of such disorders as diabetes or dyslipemia. Nuclear receptors are implicated in the control of lipid homeostasis. They establish a coordinated net of metabolic sensors which integrates lipid metabolism, inflammation, drug metabolism, bile acid synthesis and glucose homeostasis among other processes [50, 51]. For instance, Farnesoid X Receptor (FXR) is implicated in the metabolism of bile acids and the control of the metabolism of cholesterol and triglycerides. LXR, PPARs, Pregnane X receptor (PXR) and RORs also participate in the control of lipid and lipoprotein metabolism. [52, 53]. It has recently been demonstrated that at least some of the phenolic compounds present in a grape seed procyanidin extract: (a) are RORα agonists; and (b) increase FXR activity in a dose-dependent way when the natural ligand of the nuclear receptor is present [54].

In summary, phenolic compounds have been reported to modulate intracellular signalling by: (a) interfering in platelet-derived growth factor receptor (βPDGFR) signalling (through PI3K and p38 MAPK pathways); (b) modulating the activity of target enzymes (*e.g.* nitric oxide synthase); and (c) modulating gene expression.

References

[1] M. Wink, Evolution of secondary metabolites from an ecological and molecular phylogenetic perspective., Phytochemistry 64 (2003) 3-19.
[2] J.B. Harborne, C.A. Williams, Advances in flavonoid research since 1992., Phytochemistry 55 (2000) 481-504.
[3] J.J. Macheix, A. Fleuriet, J. Billot, Fruit Phenolics, Boca Raton, FL: CRC Press 1990.
[4] Documentation for the Update of the USDA Database for Flavonoid Content of Selected foods, Release 2.1 (2007).
[5] M.S. DuPont, Z. Mondin, G. Williamson, K.R. Price, Effect of variety, processing, and storage on the flavonoid glycoside content and composition of lettuce and endive., J Agric Food Chem 48 (2000) 3957-3964.
[6] S.A. Aherne, N.M. O'Brien, Dietary flavonols: chemistry, food content, and metabolism., Nutrition 18 (2002) 75-81.
[7] A. Scalbert, G. Williamson, Dietary intake and bioavailability of polyphenols., J Nutr 130 (2000) 2073S-2085S.
[8] F. Saura-Calixtoa, J. Serrano, I. Goñi, Intake and bioaccessibility of total polyphenols in a whole diet, Food Chemistry 101 (2007) 492.
[9] C. Manach, A. Scalbert, C. Morand, C. Rémésy, L. Jiménez, Polyphenols: food sources and bioavailability., Am J Clin Nutr 79 (2004) 727-747.
[10] A. Fleuriet, J.J. Macheix, Phenolic Acids in Fruits and Vegetables, Flavonoids in Health and Disease, 2003, pp. 1-42.
[11] A. Cassidy, B. Hanley, R.M. Lamuela-Raventos, Isoflavones, lignans and stilbenes - origins, metabolism and potential importance to human health, J Sci Food Agric 80 (2000) 1044-1062.
[12] G.R. Beecher, Overview of dietary flavonoids: nomenclature, occurrence and intake., J Nutr 133 (2003) 3248S-3254S.
[13] J.M. Harnly, R.F. Doherty, G.R. Beecher, J.M. Holden, D.B. Haytowitz, S. Bhagwat, S. Gebhardt, Flavonoid content of U.S. fruits, vegetables, and nuts., J Agric Food Chem 54 (2006) 9966-9977.
[14] W. Mullen, S.C. Marks, A. Crozier, Evaluation of phenolic compounds in commercial fruit juices and fruit drinks., J Agric Food Chem 55 (2007) 3148-3157.
[15] J.B. Harborne, Nature, distribution, and function of plant flavonoids, in: V. Cody, E. Middleton, J.B. Harborne (Eds.), Plant flavonoids in biology and medicine: biochemical, pharmacological, and structure–activity relationships, vol. 213, Alan R. Liss, New York, 1986, pp. 1-15.
[16] K.E. Heim, A.R. Tagliaferro, D.J. Bobilya, Flavonoid antioxidants: chemistry, metabolism and structure-activity relationships., J Nutr Biochem 13 (2002) 572-584.
[17] C. Manach, G. Williamson, C. Morand, A. Scalbert, C. Rémésy, Bioavailability and bioefficacy of polyphenols in humans. I. Review of 97 bioavailability studies., Am J Clin Nutr 81 (2005) 230S-242S.
[18] G. Williamson, C. Manach, Bioavailability and bioefficacy of polyphenols in humans. II. Review of 93 intervention studies., Am J Clin Nutr 81 (2005) 243S-255S.
[19] P.C. Hollman, J.H. de Vries, S.D. van Leeuwen, M.J. Mengelers, M.B. Katan, Absorption of dietary quercetin glycosides and quercetin in healthy ileostomy volunteers., Am J Clin Nutr 62 (1995) 1276-1282.
[20] G. Williamson, A.J. Day, G.W. Plumb, D. Couteau, Human metabolic pathways of dietary flavonoids and cinnamates., Biochem Soc Trans 28 (2000) 16-22.
[21] V. Crespy, C. Morand, C. Besson, N. Cotelle, H. Vézin, C. Demigné, C. Rémésy, The splanchnic metabolism of flavonoids highly differed according to the nature of the compound., Am J Physiol Gastrointest Liver Physiol 284 (2003) G980-988.
[22] K. Murota, J. Terao, Quercetin appears in the lymph of unanesthetized rats as its phase II metabolites after administered into the stomach., FEBS Lett 579 (2005) 5343-5346.

ISBN:978-84-691-1553-4 /DL: T.151-2008

[23] J.P. Spencer, H. Schroeter, A.J. Crossthwaithe, G. Kuhnle, R.J. Williams, C. Rice-Evans, Contrasting influences of glucuronidation and O-methylation of epicatechin on hydrogen peroxide-induced cell death in neurons and fibroblasts., Free Radic Biol Med 31 (2001) 1139-1146.

[24] V.C.J. de Boer, Towards functional effects of polyphenols, Wageningen University, 2007.

[25] Y. Konishi, S. Kobayashi, Microbial metabolites of ingested caffeic acid are absorbed by the monocarboxylic acid transporter (MCT) in intestinal Caco-2 cell monolayers., J Agric Food Chem 52 (2004) 6418-6424.

[26] L.A. Griffiths, Mammalian metabolism of flavonoids, in: J.H. Harborne, T.J. Mabry (Eds.), The flavonoids: advances in research, Chapman and Hall, London, 1982, pp. 681-718.

[27] S.E. Rasmussen, H. Frederiksen, K. Struntze Krogholm, L. Poulsen, Dietary proanthocyanidins: occurrence, dietary intake, bioavailability, and protection against cardiovascular disease., Mol Nutr Food Res 49 (2005) 159-174.

[28] W.G. Li, X.Y. Zhang, Y.J. Wu, X. Tian, Anti-inflammatory effect and mechanism of proanthocyanidins from grape seeds., Acta Pharmacol Sin 22 (2001) 1117-1120.

[29] X. Terra, J. Valls, X. Vitrac, J.M. Mérrillon, L. Arola, A. Ardèvol, C. Bladé, J. Fernandez-Larrea, G. Pujadas, J. Salvadó, M. Blay, Grape-seed procyanidins act as antiinflammatory agents in endotoxin-stimulated RAW 264.7 macrophages by inhibiting NFkB signaling pathway., J Agric Food Chem 55 (2007) 4357-4365.

[30] I.C. Arts, P.C. Hollman, Polyphenols and disease risk in epidemiologic studies., Am J Clin Nutr 81 (2005) 317S-325S.

[31] R.J. Williams, J.P. Spencer, C. Rice-Evans, Flavonoids: antioxidants or signalling molecules?, Free Radic Biol Med 36 (2004) 838-849.

[32] J.P. Spencer, C. Rice-Evans, R.J. Williams, Modulation of pro-survival Akt/protein kinase B and ERK1/2 signaling cascades by quercetin and its in vivo metabolites underlie their action on neuronal viability., J Biol Chem 278 (2003) 34783-34793.

[33] S. Fröjdö, D. Cozzone, H. Vidal, L. Pirola, Resveratrol is a class IA phosphoinositide 3-kinase inhibitor., Biochem J (2007).

[34] J.D. Lambert, S. Sang, C.S. Yang, Possible controversy over dietary polyphenols: benefits vs risks., Chem Res Toxicol 20 (2007) 583-585.

[35] H. Schroeter, C. Boyd, J.P. Spencer, R.J. Williams, E. Cadenas, C. Rice-Evans, MAPK signaling in neurodegeneration: influences of flavonoids and of nitric oxide., Neurobiol Aging 23 (2002) 861-880.

[36] G. Agullo, L. Gamet-Payrastre, S. Manenti, C. Viala, C. Rémésy, H. Chap, B. Payrastre, Relationship between flavonoid structure and inhibition of phosphatidylinositol 3-kinase: a comparison with tyrosine kinase and protein kinase C inhibition., Biochem Pharmacol 53 (1997) 1649-1657.

[37] C.J. Vlahos, W.F. Matter, K.Y. Hui, R.F. Brown, A specific inhibitor of phosphatidylinositol 3-kinase, 2-(4-morpholinyl)-8-phenyl-4H-1-benzopyran-4-one (LY294002). J Biol Chem 269 (1994) 5241-5248.

[38] L. Gamet-Payrastre, S. Manenti, M.P. Gratacap, J. Tulliez, H. Chap, B. Payrastre, Flavonoids and the inhibition of PKC and PI 3-kinase., Gen Pharmacol 32 (1999) 279-286.

[39] M. Pinent, M. Blay, M.C. Bladé, M.J. Salvadó, L. Arola, A. Ardévol, Grape seed-derived procyanidins have an antihyperglycemic effect in streptozotocin-induced diabetic rats and insulinomimetic activity in insulin-sensitive cell lines., Endocrinology 145 (2004) 4985-4990.

[40] M. Shimizu, Y. Kobayashi, M. Suzuki, H. Satsu, Y. Miyamoto, Regulation of intestinal glucose transport by tea catechins., Biofactors 13 (2000) 61-65.

[41] Y.H. Kao, R.A. Hiipakka, S. Liao, Modulation of endocrine systems and food intake by green tea epigallocatechin gallate., Endocrinology 141 (2000) 980-987.

[42] M.J. Kim, G.R. Ryu, J.S. Chung, S.S. Sim, D.S. Min, D.J. Rhie, S.H. Yoon, S.J. Hahn, M.S. Kim, Y.H. Jo, Protective effects of epicatechin against the toxic effects of streptozotocin on rat pancreatic islets: in vivo and in vitro., Pancreas 26 (2003) 292-299.

[43] R.A. Anderson, M.M. Polansky, Tea enhances insulin activity., J Agric Food Chem 50 (2002) 7182-7186.

[44] E. Cascón, R. Roig, A. Ardèvol, M.J. Salvadó, L. Arola, C. Bladé, Nonalcoholic components in wine reduce low density lipoprotein cholesterol in normocholesterolemic rats., Lipids 36 (2001) 383-388.

[45] J.A. Vinson, K. Teufel, N. Wu, Red wine, dealcoholized red wine, and especially grape juice, inhibit atherosclerosis in a hamster model., Atherosclerosis 156 (2001) 67-72.

[46] C. Auger, B. Caporiccio, N. Landrault, P.L. Teissedre, C. Laurent, G. Cros, P. Besançon, J.M. Rouanet, Red wine phenolic compounds reduce plasma lipids and apolipoprotein B and prevent early aortic atherosclerosis in hypercholesterolemic golden Syrian hamsters (Mesocricetus auratus). J Nutr 132 (2002) 1207-1213.

[47] D.A. Moreno, N. Ilic, A. Poulev, D.L. Brasaemle, S.K. Fried, I. Raskin, Inhibitory effects of grape seed extract on lipases., Nutrition 19 (2003) 876-879.

[48] G. Conseil, H. Baubichon-Cortay, G. Dayan, J.M. Jault, D. Barron, A. Di Pietro, Flavonoids: a class of modulators with bifunctional interactions at vicinal ATP- and steroid-binding sites on mouse P-glycoprotein., Proc Natl Acad Sci U S A 95 (1998) 9831-9836.

[49] C. Auger, P. Gérain, F. Laurent-Bichon, K. Portet, A. Bornet, B. Caporiccio, G. Cros, P.L. Teissédre, J.M. Rouanet, Phenolics from commercialized grape extracts prevent early atherosclerotic lesions in hamsters by mechanisms other than antioxidant effect., J Agric Food Chem 52 (2004) 5297-5302.

[50] J.J. Eloranta, G.A. Kullak-Ublick, Coordinate transcriptional regulation of bile acid homeostasis and drug metabolism., Arch Biochem Biophys 433 (2005) 397-412.

[51] S.W. Beaven, P. Tontonoz, Nuclear receptors in lipid metabolism: targeting the heart of dyslipidemia., Annu Rev Med 57 (2006) 313-329.

[52] A.C. Li, C.K. Glass, PPAR- and LXR-dependent pathways controlling lipid metabolism and the development of atherosclerosis., J Lipid Res 45 (2004) 2161-2173.

[53] M. Makishima, Nuclear receptors as targets for drug development: regulation of cholesterol and bile acid metabolism by nuclear receptors., J Pharmacol Sci 97 (2005) 177-183.

[54] J.M. Del Bas, Modulation of hepatic lipoprotein metabolism by dietary procyanidins, Departament de Bioquímica i Biotecnologia, Universitat Rovira i Virgili, Tarragona, 2007, p. 218.

[55] C. Felgines, S. Talavera, O. Texier, A. Gil-Izquierdo, J.L. Lamaison, C. Remesy, Blackberry anthocyanins are mainly recovered from urine as methylated and glucuronidated conjugates in humans., J Agric Food Chem 53 (2005) 7721-7727.

[56] X. Wu, G. Cao, R.L. Prior, Absorption and metabolism of anthocyanins in elderly women after consumption of elderberry or blueberry., J Nutr 132 (2002) 1865-1871.

[57] T. Frank, M. Netzel, G. Strass, R. Bitsch, I. Bitsch, Bioavailability of anthocyanidin-3-glucosides following consumption of red wine and red grape juice., Can J Physiol Pharmacol 81 (2003) 423-435.

[58] C. Felgines, S. Talavéra, M.P. Gonthier, O. Texier, A. Scalbert, J.L. Lamaison, C. Rémésy, Strawberry anthocyanins are recovered in urine as glucuro- and sulfoconjugates in humans., J Nutr 133 (2003) 1296-1301.

[59] M.G. Boersma, H. van der Woude, J. Bogaards, S. Boeren, J. Vervoort, N.H. Cnubben, M.L. van Iersel, P.J. van Bladeren, I.M. Rietjens, Regioselectivity of phase II metabolism of luteolin and quercetin by UDP-glucuronosyl transferases., Chem Res Toxicol 15 (2002) 662-670.

[60] L.I. Mennen, D. Sapinho, H. Ito, P. Galan, S. Hercberg, A. Scalbert, Urinary excretion of 13 dietary flavonoids and phenolic acids in free-living healthy subjects - variability and possible use as biomarkers of polyphenol intake., Eur J Clin Nutr (2007).

ISBN:978-84-691-1553-4 /DL: T.151-2008

[61] W. Mullen, C.A. Edwards, A. Crozier, Absorption, excretion and metabolite profiling of methyl-, glucuronyl-, glucosyl- and sulpho-conjugates of quercetin in human plasma and urine after ingestion of onions., Br J Nutr 96 (2006) 107-116.

[62] M.S. DuPont, A.J. Day, R.N. Bennett, F.A. Mellon, P.A. Kroon, Absorption of kaempferol from endive, a source of kaempferol-3-glucuronide, in humans., Eur J Clin Nutr 58 (2004) 947-954.

[63] C. Manach, C. Morand, A. Gil-Izquierdo, C. Bouteloup-Demange, C. Rémésy, Bioavailability in humans of the flavanones hesperidin and narirutin after the ingestion of two doses of orange juice., Eur J Clin Nutr 57 (2003) 235-242.

[64] I. Erlund, M.L. Silaste, G. Alfthan, M. Rantala, Y.A. Kesäniemi, A. Aro, Plasma concentrations of the flavonoids hesperetin, naringenin and quercetin in human subjects following their habitual diets, and diets high or low in fruit and vegetables., Eur J Clin Nutr 56 (2002) 891-898.

[65] M.J. Brunet, C. Bladé, M.J. Salvadó, L. Arola, Human apo A-I and rat transferrin are the principal plasma proteins that bind wine catechins., J Agric Food Chem 50 (2002) 2708-2712.

[66] C. Tsang, C. Auger, W. Mullen, A. Bornet, J.M. Rouanet, A. Crozier, P.L. Teissedre, The absorption, metabolism and excretion of flavan-3-ols and procyanidins following the ingestion of a grape seed extract by rats., Br J Nutr 94 (2005) 170-181.

[67] M.P. Gonthier, J.L. Donovan, O. Texier, C. Felgines, C. Remesy, A. Scalbert, Metabolism of dietary procyanidins in rats., Free Radic Biol Med 35 (2003) 837-844.

[68] L.Y. Rios, M.P. Gonthier, C. Rémésy, I. Mila, C. Lapierre, S.A. Lazarus, G. Williamson, A. Scalbert, Chocolate intake increases urinary excretion of polyphenol-derived phenolic acids in healthy human subjects., Am J Clin Nutr 77 (2003) 912-918.

Tables and Figures

Table 1. List of phenolic compounds that have been detected in plasma or urine

Anthocyanins	**References**
Cyanidin	55
Cyanidin 3-glucoside	55, 56, 57, 17
Cyanidin 3-xyloside	55
Cyanidin diglucuronide [(1)]	55
Cyanidin glucuronide [(1)]	55
Cyanidin–3-sambubioside	56
Cyanidin–3-glucoside monoglucuronide	56
Peonidin	55
Peonidin 3-glucoside	54, 56, 57
Peonidin monoglucuronide	55, 56
Peonidin–3-sambubioside	56
Pelargonidin	58
Pelargonidin glucuronide [(2)]	58
Pelargonidin-3-glucoside	58, 17
Pelargonidin-sulfate	58
Malvidin 3-glucoside	57, 17
Petunidin 3-glucoside	57
Delphinidin 3-glucoside	57

The preferred glycosylation site on the flavonol molecule is the C-3 position and, less frequently, the C-7 position. D-glucose is the most usual sugar residue, but other substitutions include arabinose, galactose, glucofamnose, lignin, L-rhamnose and xylose [6].

(1) The analytical techniques used could not determine the exact sites of glucuronidation. However, glucuronidation of flavonoids occurred at different hydroxyl groups within the structure (with preferences for the 7-, 3-, 3' or 4'-hydroxyls) [59, 58]

(2) Pelargonidin does not have a hydroxyl group in 3'. So, the three putative glucuronides are 7-, 3- and 4'-monoglucuronides of pelargonidin [58].

Flavonols	**References**
Quercetin	17, 60
Quercetin diglucuronide	61
Isoharmentin 3 glucuronide	17
Isoharmentin (quercetin methylated in 3'-position)	17, 60
Quercetin-3-*O*-glucuronide	17
3'-*O*-methylquercetin-3-*O*-glucuronide	17
Quercetin-3'-*O*-sulfate	17
Kaempferol	62
Kaempferol-3-glucuronide	62

ISBN:978-84-691-1553-4 /DL: T.151-2008

The presence of intact glycosides of quercetin in plasma was debated a few years ago, but it is currently accepted that these compounds are absent from plasma after nutritional doses [17, 61].

Flavanones	**References**
Hesperetin monoglucuronides [(1)]	17
Hesperetin sulphoglucuronides	63
Naringenin	17, 60
Hesperetin	64, 60

[(1)] The positions of glucuronidation are still unknown

Flavanols (monomers or catechins)	**References**
(+)-catechin [(1)]	65
(+)-catechin glucuronide* [(2)]	66
Methyl-(+)-catechin glucuronide*	66
(+)-catechin sulphate*	66
Methyl-(+)-catechin sulphate*	66
3'-*O*-methyl-(+)-catechin*	66
3'-*O*-methyl-(+)-catechin glucuronide*	66
Epicatechin	17
Epicatechin-3'-*O*-glucuronide	17
(-)-epicatechin glucuronide *	66
[(-)-epicatechin-7-*O*-glucuronide or 3'-*O*-methyl-(-)-epicatechin-7-*O*-glucuronide]	
Methyl-(-)-epicatechin glucuronide*	66
4'-*O*-methylepicatechin-3'-*O*-glucuronide	17
4'-*O*-methylepicatechin-5-*O*-glucuronide	17
4'-*O*-methylepicatechin-7-*O*-glucuronide	17
4'-*O*-methylepicatechin	17
(-)-epicatechin sulphate*	66
Methyl-(+)-epicatechin sulphate*	66
Epigallocatechin (EGC) [(3)]	17
4'-*O*-methyl-epigallocatechin	17
Epigallocatechin gallate (EGCG) [(4)]	17
4',4''-di-*O*-methyl-EGCG	17

* Molecules found in rats

[(1)] Catechin can also be methylated (preferentially at the 3'-position [17].
[(2)] The location of the glucuronization has not been elucidated.
[(3)] Galloylated catechins have never been recovered in urine, because they are preferentially excreted in bile.
[(4)] EGCG is the only known polyphenol present in plasma in large proportions in a free form. The other catechins are highly conjugated with glucuronic acid and/or sulfate groups.

Procyanidins	References
B1 (1)	17, 66
B2 (1)	17, 66
Tetramethylated dimeric *	66
B3*	66
B4*	66
C2*	66

Gonthier et al. [67] showed that the extent of degradation into aromatic acids decreased as the degree of polymerization increased (*i.e.* it is 21 times lower for polymers than for the catechin monomer). This is, probably, the result of the antimicrobial properties and protein-binding capacity that have been frequently described for proanthocyanidins.

* The tetramethylated dimeric was detected in the liver of rats; and the dimers (B3 and B4) and trimer (C2) were detected in rat urine.
(1) Polymeric proanthocyanidins are not absorbed as such. In fact, the absorption of these dimers was less (aprox 100-fold lower) than that of the flavanol monomers. *In vitro* and animal studies confirmed that the polymerization degree greatly impairs intestinal absorption [17].

Isoflavones	References
Daidzen	17
Genistein	17
Glycitein	17
Daidzin	17
Genistin	17
Glycitin	17

Acids	References
Caffeic acid	17
Chlorogenic acid (1)	17
Ferulic acid	17
Gallic acid	17
Gallic glucuronidated	17

(1)It is not clear whether its acid is present, as such or in a conjugated form, in human plasma.

ISBN:978-84-691-1553-4 /DL: T.151-2008

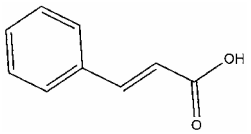
Metabolites produced by microflora (acids)	**Precursor**	**References**
Protocatechuic acid [(1)]	Cyanidin-3-*O*-glucoside	17
3,4-dihydroxyphenylacetic	Quercetin	17
3-methoxy-4-hydroxyphenylacetic (homovanillic acid)	Quercetin	17
3-hydroxyphenylacetic propionic acid [(2)]	Naringenin/Hesperetin	17
p-hydroxybenzoic acid [(2)]	Naringenin/Hesperetin	17
p-coumaric acid [(2)]	Naringenin/Hesperetin	17
Phenylpropionic acid [(2)]	Naringenin/Hesperetin	17
5-(3',4',5'-trihydroxyphenyl)valerolactone	Epigallocatechin and epicatechin	17
5-(3',4'-dihydroxyphenyl)valerolactone	Epigallocatechin and epicatechin	17
5-(3', 5'-dihydroxyphenyl)valerolactone	Epigallocatechin and epicatechin	17
m-hydroxyphenylpropionic acid [(3)]	proanthocyanidins	17
p-hydroxyphenylpropionic acid [(3)]	proanthocyanidins	17
m-hydroxyphenylacetic acid	proanthocyanidins	17
p-hydroxyphenylacetic acid	proanthocyanidins	17
m-hydroxyphenylvaleric acid	proanthocyanidins	17
Phenylpropionic acid	proanthocyanidins	17
Phenylacetic acid	proanthocyanidins	17
m-hydroxybenzoic acid [(3)]	proanthocyanidins	17
Equol	isoflavones	17
7-*O*-glucuronides	isoflavones	17
4'-*O*-glucuronides	isoflavones	17
Sulfate esters	isoflavones	17
Dihydrodaidzein	isoflavones	17
Dihydrogenistein	isoflavones	17
Dihydroequol	isoflavones	17
O-desmethylangolensin	isoflavones	17
6-hydroxy-*O*-desmethylangolensin	isoflavones	17
4-*O*-methylgallic acid	Gallic acid	17

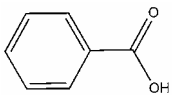
(1) Detected in rats.
(2) The microbial metabolites were obtained by *in vitro* incubation of naringenin with human microflora and further detected in rat urine.
(3) These compounds were shown to increase in human urine after consumption of procyanidin-rich chocolate [68]. However, the microbial metabolism of proanthocyanidins has never been studied in humans after consumption of purified proanthocyanidin polymers.

Figure 1. Molecular structures of the four main classes of phenolic compounds

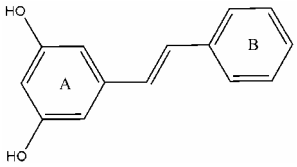
Phenolic acids

(a) Hydroxycinnamic acids (b) Hydroxybenzoic acids

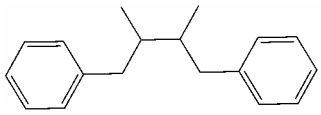




Stilbenes

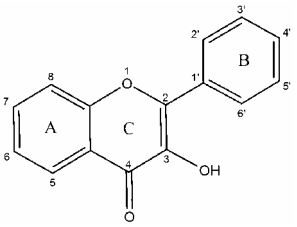


Lignans

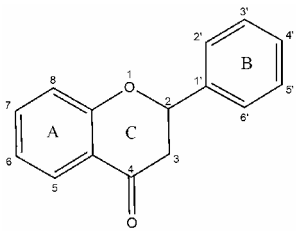


Flavonoids

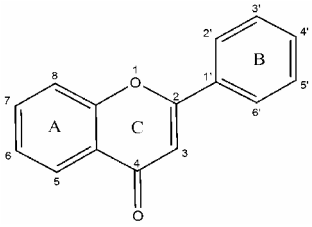
(a) Flavonols



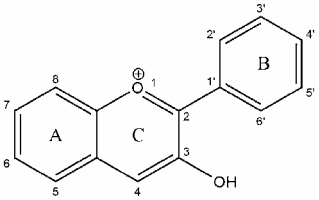
(b) Flavanones



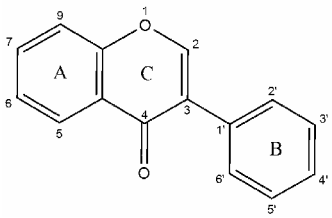
(c) Flavones



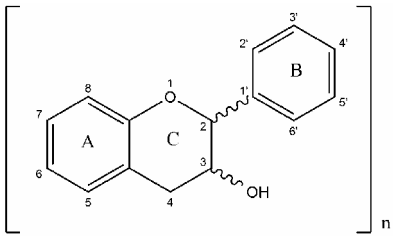
(d) Anthocyanidins



(e) Isoflavones

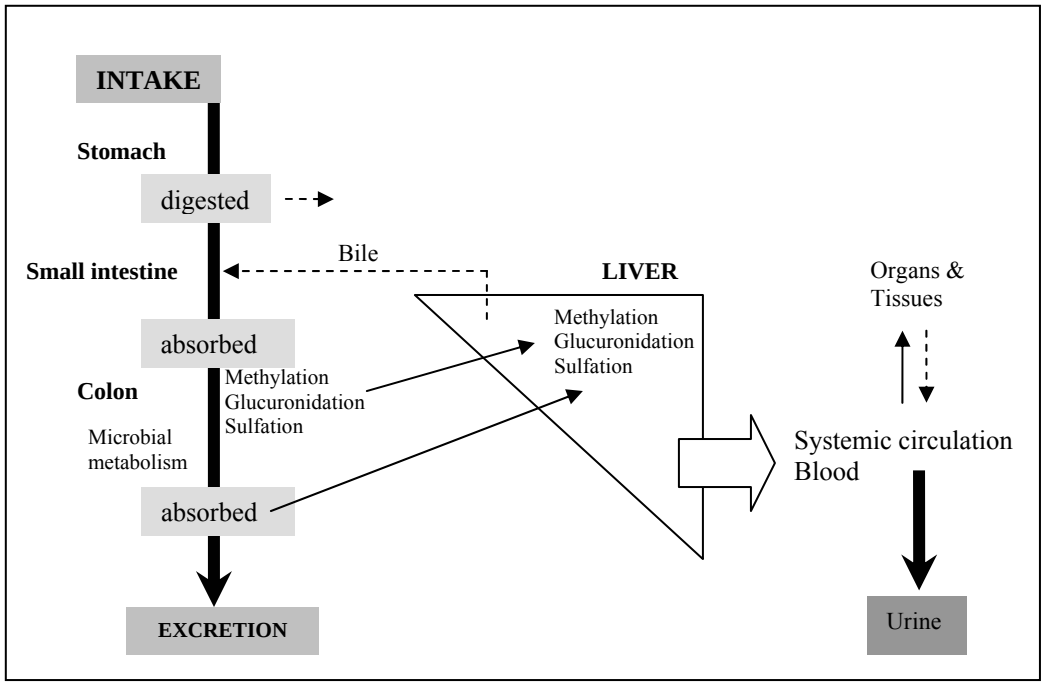


f) Flavanol monomers, dimers, trimers, etc.



ISBN:978-84-691-1553-4 /DL: T.151-2008

Figure 2. Schematic diagram of the putative pathways of digestion, absorption and metabolization of phenolic compounds



The absorption of phenolic compounds in the stomach has been reported in rats but in humans the main site of absorption is the small intestine, where some polyphenols can be either directly absorbed or hydrolyzed by glycosidases thus allowing the absorption of the aglycon form (*i.e.* the non-glycosylated form). The phenolic compounds that are not absorbed in the small intestine can be metabolized by microbial flora and absorbed in the colon. Recently, a new absorption pathway has been proposed that involves the lymphatic system (Murota & Terao, 2005). After absorption, flavonoids are bound to albumin and transported to the liver via the portal vein (the liver seems to be the main organ involved in flavonoid metabolism although the intestinal mucosa and/or kidneys also have to be considered). In the liver, the phenolic compounds and their derivates may undergo further reactions such as methylation, glucuronidation and sulfation. Then, it is thought that phenolic compounds are delivered to the blood because some have been detected in plasma (although technical limitations have only enable a few of these bioactive molecules to be detected). Since some flavonoid conjugates are polar compounds, they can be excreted in the urine and bile. In the latter case, phenolic compounds can be metabolized again by the intestinal bacteria and the resulting metabolites (or fragmentation products) can be reabsorbed and enter the enterohepatic cycle.

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

Docking

II

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

Protein-ligand docking: a review of recent advances and future perspectives

Montserrat Vaqué, Anna Ardèvol, Cinta Bladé, M. Josepa Salvadó, Mayte Blay, Juan Fernández-Larrea, Lluís Arola, Gerard Pujadas*

Departament de Bioquímica i Biotecnologia. Universitat Rovira i Virgili, C/ Marcel·lí Domingo s/n, Campus de Sant Pere Sescelades. Tarragona 43007, Catalonia (Spain)

Current Pharmaceutical Analysis (in press)

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

ABSTRACT

Understanding the interactions between proteins and ligands is crucial for the pharmaceutical and functional food industries. The experimental structures of these protein/ligand complexes are usually obtained, under highly expert control, by time-consuming techniques such as X-ray crystallography or NMR. These techniques are therefore not suitable for routinely screening the possible interaction between one receptor and thousands of ligands. To overcome this limitation, computational algorithms (*i.e.* docking algorithms) have been developed that use the individual structures of the receptor and ligand to predict the structure of their complex. The present review, then, summarizes: (a) the fundamentals of the algorithms of the most commonly used docking programmes (with particular emphasis on their strengths and limitations); (b) how the results from different docking algorithms compare (*i.e.* which software gives the best predictions); and (c) the future perspectives and challenges for docking techniques.

ABBREVIATIONS

GA: Genetic Algorithm; HTS: High-Throughput Screening; MC: Monte Carlo; MVD: Molegro Virtual Docker; RMSD: root-mean square deviation; TS: tabu search; VHTS: virtual high-throughput screening; VLS: virtual ligand screening.

KEYWORDS

eHiTS, GOLD, Molegro Virtual Docker, AutoDock, Glide

ISBN:978-84-691-1553-4 /DL: T.151-2008

INTRODUCTION

Understanding how small ligands bind to different kinds of proteins (*e.g.* enzymes, nuclear receptors, transport proteins, etc.) in order to inhibit or activate them is of capital importance not only for pharmaceutical companies but also for those industries that produce functional foods [*i.e.* those foods that, as well as containing the traditional nutrients also contain natural extracts with bioactive molecules (so called *phytochemicals* or *phytonutrients*) that can potentially provide a health benefit]. In order to understand this process, the ligands that can bind to a specific protein target (usually referred as *receptor* in the argot) must first be identified. This can be done either through expensive High-Throughput Screening (HTS) experiments in which libraries of small chemical compounds are screened against the target receptor in order to find those ligands that have a high binding affinity for it [1] or through a cheaper alternative called virtual high-throughput screening (VHTS) in which software programs try to predict if the different ligand-receptor complexes are possible or not by doing what is known as *protein-ligand docking* or also *small-molecule–protein docking*.

Therefore, the present review aims to address the *state-of-art* of protein-ligand docking and associated issues by focusing on: (a) where to find the protein and the ligand structures that are needed; (b) where to dock ligands (*i.e.* where the receptor ligand-binding site is); (c) how a docking program works (with special emphasis on how the program accounts for ligand flexibility and how the results are scored); (d) the importance of considering receptor flexibility during docking; (e) a description of some popular docking programs with special emphasis on their strengths and limitations; (f) how docking programs compare (*i.e.* which one gives the best results); and (g) the future perspectives and challenges for docking techniques.

The raw matherials for docking. Where to find protein and ligand structures.

The first question that has to be answered when planning a docking experiment is *Is there an experimental structure for the protein I want to use as a target during the docking?* (see Figure 1). To answer this question, it is necessary to go to the PDB database [2] (http://www.pdb.org) and determine whether the corresponding target has been deposited or not. One way of doing this is to determine whether there is a protein in the PDB with a sequence that is 100% identical to that of the protein target. At present, the PDB contains the experimental structures for 40511 proteins obtained either by X-ray crystallography (35105) or by NMR (5406). Of these, only 6700 polypeptide chains are below the 90% similarity threshold and have enough crystallographic quality (resolution cutoff of 2.0 Å and R-factor cutoff of 0.250) to be used as targets in docking experiments according to the latest download list from the PISCES server [3] (http://dunbrack.fccc.edu/Guoli/pisces_download.php). Therefore, considering that there are currently

289252 proteins with a known sequence in the UniProt Knowledgebase [4] (http://www.expasy.org/sprot/), it is clear that the PDB only covers a small part of the *protein space* that is made up of the potential targets of a docking experiment (although the rate of 3D macromolecular structure determination increases every year with the development of techniques such as high-throughput X-ray crystallography [5]). If the target protein has not yet been deposited in the PDB but the database contains a protein with a similar sequence, then homology modelling can be used to predict the 3D structure of the former protein. In this respect, as well as programs or webservers that allow users to build these models by strictly controlling the values of several parameters [6-8], some databases of protein structures modelled by homology have been automatically created by these programs or webservers by using default parameter values when generating the homology model. Examples of such databases are ModBase [9] (http://modbase.compbio.ucsf.edu/modbase-cgi/search_form.cgi) and the SWISS-MODEL Repository [10] (http://swissmodel.expasy.org/repository/) whose models can be easily accessed from links within appropiate UniProt Knowledgebase entries. At this point, it is worth pointing out that some years ago, docking into protein active sites built by sequence homology was considered to be a source of unreliable results. Nevertheless, the improvement in homology modelling methods has radically changed this point of view and they have now been successfully used in docking experiments [11-18].

The second question to answer is *Where can the structures of the ligands that will be used during the docking be found?* (see Figure 1). There are two main sources of ligand structures. The first source is computer programs that use graphic tools or languages such as SMILES [19] (*i.e.* **S**implified **M**olecular **I**nput **L**ine **E**ntry **S**ystem) to build these structures. Examples of such programs are: (a) ChemDraw Ultra™ (CambridgeSoft Corporation, Cambridge MA, USA, http://www.cambridgesoft.com/); (b) KnowItAll ChemWindow™ (Bio-Rad Laboratories Inc., Hercules CA, USA, http://www.bio-rad.com/); (c) ISIS/Draw™ (MDL Information Systems, Inc., San Ramon CA, USA, http://www.mdli.com/); and (d) ACD/ChemSketch™ (Advanced Chemistry Development, Inc., Toronto, Canada, http://www.acdlabs.com/). These programmes are useful only if the set of ligands is limited or if the molecules have not been previously reported. The second source is databases of, sometimes, purchasable compounds. Small molecular structures can be searched for and the ones that match the search criteria can be retrieved to the user computer to perform the virtual ligand screening (VLS) [although in some cases, some curation of the structures (*e.g.* assigning the correct protonation state and the partial charges, obtaining the tautomeric forms, converting 2D structures to 3D, etc.) is necessary before they can be used in docking experiments]. Examples of such chemical databases and their current status in May 2007 are: (a) the Available Chemicals Directory™ (http://www.mdl.com/products/experiment/available_chem_dir), which contains 571000 unique chemicals

ISBN:978-84-691-1553-4 /DL: T.151-2008

that can be purchased from over 719 suppliers worldwide; (b) the MDL® screening compounds directory (formerly ACD-SC™; http://www.mdl.com//products/experiment/screening_compounds/) where 3.4 million structures from 46 international chemical suppliers can be found; (c) the iResearch™ Library (http://www.chemnavigator.com/cnc/products/iRL.asp), which contains over 37.2 million chemical structures (around 20.6 million unique) from more than 252 chemistry suppliers; (d) the National Cancer Institute database [20] (http://129.43.27.140/ncidb2/download.html) with 250251 compounds; (e) the Ligand.Info database [21] (http://Ligand.info/), which is a compilation of various public databases of small molecules [*i.e.* ChemBank [22] (http://chembank.broad.harvard.edu/), ChemPDB (http://www.ebi.ac.uk/msd-srv/msdchem/cgi-bin/cgi.pl), KEGG [23] (http://www.genome.ad.jp/ligand/), NCI [20] (http://dtp.nci.nih.gov/), AKos GmbH (http://www.akosgmbh.eu/), Asinex Ltd (http://www.asinex.com/), and TimTec (http://www.timtec.net/)] and which contains 1159274 entries; (f) the GDB database [24] (*i.e.* the **G**enerated **D**ata**b**ase of Chemical Space of Small Molecules; http://www.dcb.unibe.ch/groups/reymond/), which contains 26.4 million compounds, the vast majority of which have never been synthesized; and (g) the ZINC database [25] (http://blaster. docking.org/zinc/), which contains over 4.6 million compounds in ready-to-dock format from the catalogues of 46 different chemical suppliers. This last database can be freely accessed, searched and downloaded and is of special importance for those scientist interested in VLS because it has been designed with docking, substructure searching, and compound purchasing in mind. In this respect, each molecule in the library is ready to be downloaded and used by a number of popular docking programs and it is also annotated with the following properties: (a) molecular weight; (b) vendor and original catalog number; (c) calculated LogP; (d) number of rotatable bonds; (e) number of hydrogen-bond donors; (f) number of hydrogen-bond acceptors; (g) number of chiral centers; (h) number of chiral double bonds (E/Z isomerism); (i) polar and apolar desolvation energy (in kcal/mol); (i) net charge; (j) number of rigid fragments; and (k) function or activity (when available). Each molecule in ZINC is also prepared in the protonation states that correspond to pH values between 5 and 9.5 and in their different tautomeric forms (this is very important because molecules that are not in the biologically appropriate protonation state, tautomeric, stereo-, or regio-isomeric form often fail to dock and score correctly). A web-based query tool can be used to search ZINC with several criteria: (a) limits on such molecular properties as net charge, molecular weight, functional groups, polar and apolar desolvation, polar surface area and calculated logP; (b) constraints on the compound vendor; and (c) substructure searching by means of a molecular drawing interface or SMILES strings. The molecular structures obtained from a database search or the entire ZINC database may be downloaded for in-house use. At this point, it is worth mentioning that ZINC molecules can be downloaded in mol2 and SDF formats and, therefore, can be directly imported into most popular docking programs (and, if necessary, the Open Babel

software [26] can be used to translate them into other popular formats; http://openbabel.sourceforge.net).

Where can ligands dock?

Before starting a docking experiment, scientists must decide on the location of the potential ligand-binding site in the receptor's surface area (see Figure 1). This is not as trivial as it seems, specially when: (a) no 3-D structure is known for a complex of the receptor under study (in particular when the activity data about the mutation of residues that are expected to be important for the receptor function are also unknown); and (b) experimental data support a certain degree of cooperativity between the natural ligand of the receptor and some of the molecules that will be docked onto this receptor (*i.e.* thus suggesting that both ligands can be bound at the same time). In this situations, there are two alternatives: (a) make an initial identification of the potential ligand-binding sites and then dock the ligands on them; and (b) make no *a priori* ligand-binding site assumption and dock all the ligands directly onto the complete receptor structure (which is known as a *blind docking* experiment). Each alternative has its strengths and weaknesses. Whereas the first alternative needs fewer computational resources and a shorter execution time than the second one (an important consideration when a large library of ligands has to be docked) the reliability of its results relies heavily on the correct prediction of the ligand-binding site (a limitation that is overcome by a blind docking experiment). Therefore, it is not always easy to find a correct balance between speed and accuracy. So, it is very important to use tools that can accurately predict putative ligand-binding sites and achieve this equilibrium. Examples of these tools are: (a) Fuzzy-oil-drop [27] (http://www.bioinformatics.cm-uj.krakow.pl/activesite); (b) PLB [28]; (c) LigProf [29] (http://www.cropnet.pl/ligprof); (d) Q-SiteFinder [30] (http://www.bioinformatics.leeds.ac.uk/qsitefinder/); (e) Protemot [31] (http://protemot.csbb.ntu.edu.tw/); (f) CASTp [32] (http://sts.bioengr.uic.edu/castp/); (g) MEDock [33] (http://medock.csbb.ntu.edu.tw/); (h) PASS [34] (http://www.ccl.net/cca/software/UNIX/pass/overview.shtml); (i) SURFNET-ConSurf [35] (http://consurf-hssp.tau.ac.il); and (j) the pocket detection algorithm that comes with the eHiTS® (http://www.simbiosys.ca/ehits/) [36] and Molegro Virtual Docker™ (MVD) [37] docking programs (see below for a more detailed description). For those readers interested in ligand-binding site prediction, we recommend some excellent reviews of the field [38-41].

How does a docking program work?

In a protein-ligand docking experiment, the coordinates from the individual molecular entities (*i.e.* the receptor and the ligand) are used to predict the coordinates for the resulting complex. To do so, all docking programs have the following in common: (a) an algorithm that makes a search that is as exhaustive as possible of all the possible coordinates for the complex and, therefore, suggests its candidate 3D structures; and (b) a scoring function that scores all these

ISBN:978-84-691-1553-4 /DL: T.151-2008

candidates and ranks them according to the intermolecular interaction energy (*i.e.* the more negative this energy is, the higher the candidate's score). Despite this apparent coincidence, each docking program differs from the others in the way the searching algorithm works and/or the weight it gives to the different types of intermolecular interactions and steric overlaps that it considers to get the score (*i.e.* the self scoring function).

When looking for the coordinates of the resulting protein-ligand complex, different degrees of molecular flexibility can be considered. The first degree assumes that the ligand and the receptor are rigid bodies and, therefore, that there is no degree of freedom around any rotable bond. This is the basis of the simplest searching algorithms and, in fact, this was the way that the first docking programs worked [42]. Currently, this approach is only used for protein-protein docking. A more recent approach to flexibility in protein-ligand docking considers that the receptor is rigid whereas the ligand is flexible. In fact, this is the approach used by most of the currently available docking programs and, in the following review sections, a variety of approaches for dealing with ligand flexibility will be described. Nevertheless, it has also been argued that, in many cases, the receptor cannot be considered to be a rigid body and that its flexibility (or at least, its binding site flexibility) should also be taken into account (particularly when the receptor changes its ligand binding site geometry upon ligand binding). Nowadays, some kind of receptor flexibility is included in the most recent versions of some protein-ligand docking programs (see below).

How do docking programs account for ligand flexibility? Methods for finding the ligand poses

The ability to produce a large and diverse set of ligand poses (*i.e.* candidate coordinates for conformations, positions and orientations of the ligand within the protein-ligand complex) is a prerequisite for a docking tool to be useful [36]. There are two main types of algorithms that allow docking programs to search the conformational space of the ligand in order to find its poses: (a) systematic or directed approaches; and (b) random or stochastic methods (see Figure 1). Simulation methods such as molecular dynamics and quantum mechanics are also able to do this but, at present, they are too computationally expensive to be applied during VLS.

There are three subtypes of systematic or directed search algorithms (see Figure 1): (a) conformational search methods; (b) fragmentation or incremental construction methods; and (c) database methods. They all have in common that the algorithms try to explore all the degrees of freedom of the ligand; the way they carry out the search, however, is different. In this respect, the conformational search algorithms try to obtain all possible ligand conformations by subjecting all the ligand bonds that can be rotated to a 360º turn by using a fixed increment. One of the main problems of this methodology is that the number of ligand conformations that can

be generated increases exponentially with the number of rotatable bonds and, therefore, its application in its purest form is very limited (*i.e.* usually several constraints and restraints on the ligand are needed to reduce the combinatorial explosion problem). On the other hand, fragmentation or incremental construction methods are currently used by docking programs such as eHiTS® [36], LUDI™ (http://www.accelrys.com/products/insight/sbd_modules.html) [43], FlexX™ (http://www.biosolveit.de/FlexX/) [44], DOCK (http://dock.compbio.ucsf.edu/) [45] and Hammerhead [46] to search for available ligand conformations. In these methods, the ligands are incrementally grown in the binding site either by dividing the ligand into several rigid fragments, docking them and finally trying to rebuild the ligand structure by joining the rigid fragments with the flexible segments that join them (this is the so called *place-and-join* approach) or by dividing the ligand into a rigid core fragment that is first docked and the rest of the ligand segments are then successively added (this is the so called *incremental* approach). The last subtype of systematic search algorithms are the database methods that use libraries of pregenerated conformations (so called *conformational ensembles*) that are subsequently subject to a rigid body docking. One example of a docking program that uses this database methodology to deal with the ligand flexibility issue is FLOG [47].

There are three subtypes of random or stochastic search algorithms: (a) Monte Carlo (MC) methods; (b) Genetic Algorithm (GA) methods; and (c) tabu search (TS) methods (see Figure 1). They all try to sample the conformational space by performing a random conformational change to the ligand followed by acceptance or rejection of the resulting conformer by using a predefined probability function. Then, if the ligand conformation is finally accepted, it is used as the starting point for a new random conformational change. The main limitation of such methods is that the pose that matches the experimenal conformation of the ligand in the complex may not be achieved and, therefore, will not be evaluated by the corresponding scoring function. In MC methods, the ligand is randomly placed in the receptor binding site, it is scored and a new conformation is generated by random changes that are applied to: (a) the ligand rotable bonds; and (b) the ligand position (*i.e.* the ligand is randomly rotated and translated). After each change, the ligand is typically minimized and scored [48]. Then, if the new solution scores better than the previous one, it is immediately accepted. On the other hand, if the latter conformation is not a new minimum, a Boltzmann-based probability function is applied. If the pose succeeds in this probability function test, it is accepted; if not, it is rejected (this is what is called a *Metropolis criterion*). The previous steps are repeated until the desired number of poses are obtained. Moreover, to improve convergence, the simulation usually occurs in several cycles: the first at high temperatures and the subsequent ones at decreasingly lower temperatures (this approach is known as MC simulated annealing). Docking programs that can deal with ligand conformational searches using MC-based algorithms are AutoDock

(http://autodock.scripps.edu/) [49] (in fact, it was the first docking program in which a MC simulated annealing algorithm was implemented), PRODOCK [50], ICM™ (http://www.molsoft.com/docking.html) [51], MCDOCK [52], DockVision™ (http://www.dockvision.com/) [53] and QXP [54]. GAs use concepts derived from genetics and the theory of biological evolution to explore the conformational space of the ligands. Unlike MC methods, GAs start from an initial population of different conformations of the ligand and each one is defined by a set of state variables or *genes* that describe the translation, the orientation, and the conformation of the ligand relative to the receptor. The complete set of these ligand *genes* is the *genotype*, whereas the resulting atomic coordinates are the *phenotype*. Then, genetic operators like mutations, crossovers, and migrations are applied to the population to sample the conformational space, until a final population that optimizes a predefined fitness function is reached. Docking programs that can use GA or GA-like algorithms to deal with ligand conformational searches are GOLD™ (http://www.ccdc.cam.ac.uk/products/life_sciences/gold/) [55, 56], AutoDock v3.0 and v4.0 [57], DIVALI [58], MVD™ [37] and DARWIN [59]. The last subtype of random search algorithms are the TS methods that work by imposing restrictions that prevent already explored areas of the ligand conformational space from being visited again and, therefore, favor the analysis of new conformations. To do so, when a new ligand conformation is available, its root-mean square deviation (*i.e.* RMSD) relative to the previously visited conformations is calculated. When this calculation is finished, the lowest RMSD is compared with a certain threshold value and, if it is higher, then the analyzed conformation for the ligand is accepted and its coordinates are stored and used to accept or reject new conformations. PRO_LEADS [60] and MVD™ [37] are the most popular docking programs that can use a TS algorithm.

How does the program score the results? The scoring functions

Once the candidate conformations for the ligand in the complex have been predicted, their binding affinity for the receptor must be scored [61-63]. This is done by means of a scoring function that evaluates the search results and then gives, ideally, the highest score to the right pose. In fact, if the search algorithm can find the correct pose but the scoring function cannot recognize it, the program will make an invalid and useless suggestion to the scientist. Therefore, the role of the scoring function is critical in every docking protocol. Nevertheless, using a very accurate scoring function is not possible because of its huge computational cost. Consequently, some assumptions and simplifications must be made to reduce their complexity and reach a balance between speed and accuracy. In fact, the lack of a suitable scoring function, both in terms of speed and accuracy, is the major bottleneck in docking [64]. Some recent studies have shown that the scoring function's performance on a particular target protein is largely case-dependent [65-74]. Therefore, any docking study should start either with an objective evaluation

of available scoring functions on the target protein so that the most suitable one can be chosen or with a scoring function that has been tuned according to the binding site characteristics [36, 75]. The scoring functions normally used in protein-ligand docking can be divided into four major classes: (a) force field-based; (b) empirical-based; (c) knowledge-based, and (d) consensus-based.

A force field is a function that expresses the energy of a system as a sum of diverse molecular mechanics (or other) terms. The use of force fields for scoring results in docking is based on the pioneering work of Prof. Irwin Kuntz at the University of California in Sant Francisco (USA) [42], which was followed up among others, by Shoichet [76] and Abagyan's [77] groups. Force field-based scoring functions are similar to empirical-based functions (see below) because they both predict the binding free energy of a protein-ligand complex by adding individual contributions from different types of interactions. Nevertheless, the interaction terms of the former are derived from the theoretical physics that underlie molecular mechanics force fields as opposed to the experimental affinities used to derive the latter. Thus, force field scoring functions use energy functions from classical molecular mechanics [78] and, in general, quantify the sum of two energies: (a) the interaction energy between the receptor and the ligand; and (b) the internal energy of the ligand. The binding free energy of the intermolecular interaction is often approximated as the result of the sum of a van der Waals energy term (described by means of the Lennard-Jones potential function) and an electrostatic energy term (described by means of a Coulombic formulation with a distance-dependent dielectric function that reduces the contribution from charge–charge interactions) [64, 65]. Moreover, AutoDock v3.0 and v4.0 [57] and G-Score [44] also consider a hydrogen-bonding term (although with different functional forms) in an attempt to increase the potential of specific molecular recognition [64]. On the other hand, the functions that describe the internal energy of the ligand are usually very similar to the protein–ligand interaction energy because they also contain van der Waals and/or electrostatic contributions (although AutoDock [57] –and optionally GoldScore [56]– also considers a hydrogen bond term and G-Score [44] includes a torsional entropy term). Force field-based scoring functions have the following limitations: (a) they have difficulties in considering solvation and entropic terms and either ignore them or deal with them in a cursory fashion [64, 79, 80] (because they were originally formulated to model enthalpic gas-phase contributions to structure and energetics); (b) non-bonded interactions can only be dealt with by introducing, in a more or less arbitrary way, cut-off distances that complicate the accurate treatment of the long-range effects involved in binding [64, 79]; (c) polar interactions are overemphasized [81]; and (d) the calculation of atomic partial charges relies on fast but inaccurate methods based on electronegativity indices instead of quantum mechanics methods such as those used in force-field development [82, 83]. Some of the force field scoring functions

that are most commonly used by docking programs are D-Score [44], G-Score (based on the Tripos force field) [44], GoldScore [56, 84], the ones used by AutoDock v3.0 and v4.0 [57] [based on the AMBER (**A**ssisted **M**odel **B**uilding and **E**nergy **R**efinement) force field [85-87] but modified with empirical parameters], the full AMBER molecular mechanics scoring function with implicit solvent used by DOCK v6.1 [45] and the one used by single ligand docking in DockVision [53, 88]. Obviously, more rigorous molecular mechanical force-fields such as AMBER [85-87], CHARMM (**C**hemistry at **HAR**vard **M**acromolecular **M**echanics) [89, 90], GROMOS (**GRO**ningen **MO**lecular **S**imulation System) [91] and OPLS (**O**ptimized **P**otentials for **L**iquid **S**imulations) [92] can also be used although at the expense of a huge increment in the computational cost. Table 1 shows the mathematical formulation for some selected force field scoring functions.

Empirical-based scoring functions are based on the idea that the binding energy (*i.e.* ΔG_{bind}) can be obtained by adding several individual and uncorrelated terms [64] and on the pioneer work of Hans-Joachim Böhm at the BASF AG Central Research at Ludwigshafen (Germany) [93]. Many of the terms in the empirical scoring functions have equivalences in the force-field scoring functions but the former are usually simpler in form than the latter. Empirical scoring functions contain terms that account for the net contribution of different types of non-bonded interactions (*i.e.* hydrogen bonds, ionic and hydrophobic interactions with the receptor) to the overall binding energy. However, some of these scoring functions also contain terms that, to some extent [64], account for: (a) non-enthalpic contributions such as the so-called rotor term, which approximates entropy penalties on binding from a weighted sum of the number of rotatable bonds in ligands [94]; and (b) solvation and desolvation effects [95-97]. Moreover, the terms that correspond to the contribution of a specific type of non-bonded interaction can be function-specific (such as the extra additional term for aromatic interactions in F-Score [98]) or implemented depending on the scoring function. For instance, the hydrogen-bonding term can be unique (as it is in ChemScore [94]) or divided into two: one for the neutral and another for the ionic hydrogen bonds (as in the original LUDI scoring function [93]). Another example is to take the hydrophobic contributions into account by considering the molecular surface area (as in LUDI [99]) or by evaluating contacts between hydrophobic atom pairs (as in ChemScore [94]). The main strength of empirical functions is that their terms are often simple to compute. Thus, experimental data such as binding energies or a set of X-ray receptor-ligand complexes are used as the input for a regression analysis that best fits the scoring function to the experimental data. Their main drawbacks, however, are that: (a) they are strongly dependent on the experimental data used in the parameterization process and, therefore, terms from differently fitted scoring functions cannot easily be recombined into a new scoring function; and (b) it is unclear whether they are able to predict the binding affinity of ligands that are structurally different from those

used in the training set. Examples of empirical scoring functions are Böhm's scoring function (which comes with LUDI [93, 99, 100]), F-Score [98], ChemScore [94], SCORE [101, 102], Fresno [95], X-SCORE [103], PLP [104] and Slide [105]. The scoring function used in Hammerhead [46] and Surflex [106-108] are fundamentally based on Böhm's approach [93] whereas the empirical-based scoring functions used by GOLD™ [55, 56] and Glide™ (http://www.schrodinger.com/) [109, 110] are modifications of the Eldridge's ChemScore function [94]. FlexX [44], on the other hand, implements modifications of both the Böhm and Eldridge functions. Finally, the "MolDock Score" scoring function used by MVD™ [37] is derived from PLP [104] but adds a new hydrogen bonding term and uses new charge schemes. It has been claimed that the empirical scoring functions are the ones that give better results [72]. Table 2 shows the mathematical formulation for some selected empirical scoring functions.

Knowledge-based scoring functions are based on knowing contact preferences and rely on the classical statistical physics idea that observed distributions of geometries can be used to deduce the potential that gives rise to the observed distribution. They were first proposed for studying protein folding [111] but they have also been used in protein ligand docking and to predict protein structures [112] and protein–protein complexes [113, 114]. In the protein-ligand docking field, they are used to score ligand binding poses by means of relatively simple atomic interaction-pair potentials that are built from statistical analyses of experimentally determined protein-ligand structures. These functions, then, are a direct consequence of the exponential number of protein-ligand complexes that have been deposited in the PDB database in recent years (20901 of the 27224 complexes that are now in the PDB were deposited after the first of January 2000) [115, 116]. In this respect, knowledge scoring functions try to extract rules by capturing such information as the frequency of occurrence and non-occurrence (*i.e.* negative data) of different atom–atom pair contacts and other typical interactions in the experimental structures of protein-ligand complexes because it is assumed that: (a) interatomic distances occurring more often than some average value should represent favorable contacts, and vice *vice versa* [81]; and (b) the observed distribution of distances between pairs of different atom types reflects their interaction energy [82]. In practice, large training sets of protein–ligand structures are analyzed to provide sets of distribution functions that are then converted to sets of atom-pair potentials by means of the inverse Boltzmann law (which provides an energy value for a given state based on observed probabilities) [82, 117]. Thus, in common with empirical methods, knowledge-based scoring functions attempt to implicitly capture those binding effects that are difficult to model explicitly [64]. Among the strengths of these functions are: (a) their simple atomic interactions-pair potentials have a low computational cost which means that large compound databases can be efficiently screened [118]; and (b) that they require no experimental binding affinities to be derived. In contrast, some limitations are that: (a) it is difficult to predict

their behaviour (*i.e* they should only be used for VLS when enough ligand and structural information is available to validate the setup of the calculation [81]); (b) they are designed to reproduce experimental structures rather than binding energies [36, 64]; (c) they are used to identify decoys or in combination with other scoring functions during VLS but not during the optimization phase [82]; and (d) they are not general or accurate enough because of the limited number of interactions that can be inferred from crystal structures and the inadequate description of repulsive forces [119]. The most popular knowledge-based scoring functions are Muegges's **P**otential of **M**ean **F**orce (PMF) [115, 120-122], BLEEP [123, 124], DrugScore [125, 126] and the **SM**all **M**olecule **G**rowth (SMoG) [127, 128] but others, such as the M-score, have recently been published [116]. Hybrid functions have also been developed that combine either the empirical and knowledge-based [36, 129] or 3D-QSAR and knowledge-based approaches [130-132]. The most important differences among knowledge-based scoring functions are: (a) the collection of experimental complexes by which they are obtained; (b) the expression of their energy functions; (c) the definition of the protein and the ligand atom types; (d) the definition of the reference states; and (e) the distance cutoffs. Table 3 shows the mathematical formulation for some selected knowledge-based scoring functions.

Finally, consensus scoring functions [67, 133] combine the information obtained from different scores to compensate for errors from individual scoring functions, thus improving the probability of finding the correct solution. Several studies have shown that these scoring functions perfom better than the individual ones [65, 71, 133]. For instance, a study covering thirteen scoring functions, two docking methods, three target receptors, several hundred active ligands and 10000 random compounds found that consensus scoring can dramatically reduce the number of false positives identified by individual scoring functions [133]. One example of a consensus scoring function is X-CSCORE [103], which combines PMF [115, 120-122], ChemScore [94] and FlexX [44] scoring functions.

The importance of considering receptor flexibility during docking

Handling the flexibility of the protein receptor is currently considered to be one of the major challenges in the field of protein-ligand docking. In fact it is well known that not all the receptors act as if they were the lock in the lock-and-key model of Emil Fischer. Thus, the coordinates for a protein that have been obtained by X-ray diffraction and that are available from the PDB correspond to the averages of the coordinates of that protein: (1) in the unit cells that are part of the diffracted crystal; and (2) during the time that the diffraction experiment is carried out. Nevertheless, the lines that contain atom coordinates in the PDB files also contain a parameter value (*i.e.* the temperature factor or B-factor) that can be thought of as a measure of how much the corresponding atom oscillates or vibrates around the position specified by the

corresponding coordinates (*i.e.* the higher the value of the B-factor, the less precise the atom coordinates are). All of this is coherent with the fact that proteins are in constant motion between different conformational states with similar energies and, in consequence, when only the receptor coordinates for one conformation are taken into account, only a partial picture of the receptor structure is used during docking. Moreover, the most populated conformations of the receptor in the unbound state are not necessarily the most populated in the protein–ligand complex and it is also important to consider that swapping ligands from two different complexes of the same protein in re-docking experiments may fail to reproduce the experimental complexes [69, 134-136]. It is also known that some proteins undergo conformational changes upon ligand binding, which range from local motions of side chains to large domain movements, and these allow the receptor to closely conform to the shape and binding mode of the ligand, reduce adverse intermolecular interactions and maximize the total binding free energy [137]. In this respect, it has been found that ligand-binding sites in proteins frequently have a dual character with some areas of low conformational stability (*e.g.* mobile loops that close over the ligand upon binding) and others of high conformational stability (*e.g.* catalytic residues in enzymes) [138]. This dual character of the active-site environment is important for optimum binding [139]. Finally, known 3D structures of protein–ligand complexes sometimes have ligands with a buried surface area between 70–100% and this can only be achieved as a direct consequence of protein flexibility [140]. Therefore, it is clear that there are good reasons to consider that using the receptor as a rigid body can give misleading results during protein-ligand docking because: (1) known active ligands cannot be docked in the receptor; and (2) true binders can be poorly scored (*i.e.* they become false negatives). In fact, it has been shown that even small changes in the receptor conformation can be important in computing binding affinities [136].

Therefore, some of the most popular approaches for incorporating receptor flexibility during docking are: (a) the *soft docking* approach which matches the surface structure of the ligand with the receptor on the basis of complementarity in size and shape, close packing, and the absence of steric hindrance and allows minor conformational changes implicitly [141]; (b) the *hinge-bending* concept that allows hinge induced motions of protein domains upon ligand binding [142, 143]; (c) the *relaxed complex* method that recognizes that a ligand may bind to conformations that occur only rarely in the dynamics of the receptor [144, 145]; (d) *rotamer-library* methods that try to represent the protein conformational space as a set of experimentally observed and preferred rotameric states for each side chain but without considering the possibility of changes in the backbone of the protein [146-151]; and (e) the combination of the information about protein motion that can be obtained from an ensemble of conformations from the target receptor [152-156] (where doubts exists about whether to use crystallized complexes,

ISBN:978-84-691-1553-4 /DL: T.151-2008

NMR or calculations as the best conformational source and about how to combine the information derived from all these conformations [157]). For a more detailed description of all these methodologies, the interested reader is recommended to take a look at the recent review by Alonso and coworkers [139].

Finally, it is worth pointing out that the latest versions of some of the more common protein-ligand docking programs (*e.g.* FlexX™/FlexE™ [156], Glide™/Prime™ [109, 110], GOLD™ [55, 56, 84], AutoDock v3.0/BDT [57, 158], AutoDock v4.0 and MVD™ [37]) allow some kind of receptor flexibility (see below for more details).

Main characteristics of selected docking software

eHiTS® (**e**lectronic **Hi**gh **T**hroughput **S**creening; http://www.simbiosys.ca/ehits/; see Table 4)

eHiTS® [36, 159] is being developed by SymBioSys Inc. in Toronto (Canada) and its main strengths are: (a) it is easy to use; (b) it performs very well (it is both quick and accurate); and (c) it has a lot of automated features that simplify the drug design workflow and provide innovative solutions to common docking problems (*e.g.* the protonation state of the ligand/receptor pair; an exhaustive search of the ligand poses; the speed-up of the calculations in VLS; automatic identification of probable binding sites, the capacity to tailor the scoring function to the characteristics of the receptor binding site, etc.). On the other hand, eHiTS® is not perfect and also has some limitations such as: (a) the receptor flexibility is limited to rotations of the -OH or $-NH3^+$ groups from some amino acids (*i.e.* Ser, Thr, Tyr and Lys); (b) all ring systems in the ligands are considered as rigid and therefore, their conformations are not changed during docking (hence, for a complete conformational sampling it is necessary to use multiple ring conformers); and (c) no knowledge-based constraints can be imposed on the docking (*e.g.* a specific ligand atom cannot be forced to be in a specific location in the poses; certain interactions cannot be prevented from occurring; etc.).

Comments on the main strengths of eHiTS®:

- It is easy to use, largely because it can directly use receptors and ligands in PDB format without previous user handling (although other formats are also available; see Table 4) and this allows, for instance, automatic: (a) identification of the ligand's rotable bonds; (b) assignment of atomic partial charges; (c) correction of the most common format errors in the PDB files; and (d) addition of hydrogens. Moreover, a graphic interface for eHiTS® has been recently released by SymBioSys Inc. [CheVi® (*i.e.* ***Che****mical* ***Vi****sualizer*); http://www.simbiosys.ca/chevi/index.html]. It makes it very easy to set up eHiTS® with full control over the run characteristics: (a) input selection; (b) output selection; (c) parameter selection; (d) database options; (e) parallel processing or distribution options;

and (f) filtering options. Moreover, CheVi® can be used to make a straightforward analysis of the docking results from eHiTS® runs, with a particular focus on the intermolecular interactions between the ligand and the receptor. Finally, both eHiTS® and CheVi® are also easy to install.

— The issue of protonation state is very important to the docking problem because ligands and receptors with different protonation states can have dramatically different binding poses. Thus, eHiTS® deals with this issue by using ambiguous property flags for positions that can be either protonated or deprotonated (*i.e.* they have a lone pair). This not only has practical consequences (*i.e.* it does not require that the user defines, *a priori*, the protonation states of the molecules involved in the docking) but also influences the correctness of the final result (*i.e.* if a docking program were to pre-set the protonation state of the ligand/receptor pair, then possible intermolecular interactions could be lost). Hence, with this approach, eHiTS® can systematically and automatically evaluate all possible protonation states for every receptor-ligand pair in a single run.

— The exhaustive search of the ligand poses made by eHiTS ® follows a five-step process (the so called *divide and conquer* approach): (1) the ligand that has to be docked is automatically divided into rigid (*i.e.* non rotable bonds) and connecting fragments; (2) each rigid fragment is independently docked into the receptor binding site to obtain the corresponding rigid fragment poses; (3) pose-sets are built with all possible combinations of rigid fragment poses (where each rigid fragment contributes with a single pose to each set) but only those sets from which the complete ligand structure can be rebuilt by adding the connecting fragments are kept for further use (*i.e.* the rest are discarded); (4) the rigid fragments of the remaining pose-sets are joined with the connecting fragments; and (5) the complete ligand poses are refined by a local energy minimization in the active site of the receptor that is driven by the self scoring function. This process has the advantage that rigid ligand fragments that by themselves have poor interaction scores with the receptor but, in contrast, are part of a pose set that scores very well are not discarded in the initial steps (because decisions on which poses are retained for further processing and optimization are made on the basis of an overall score of the full ligand, and not on partial structures). In fact, it is very important to keep fragment poses that do not get good scores, because even for high affinity ligands some fragments may be acting simply as linkers that have a minor contribution to the binding. Another advantage of the way in which eHiTS® searches for ligand poses is that when two rigid fragment poses are connected, any dihedral angle in the connecting segment can be virtually analyzed (although the one with the lowest energy is the one

that is finally selected). Therefore, the dihedral angle sampling of eHiTS® is, in practice, equivalent to a continuous sampling. This reflects what is found in the experimental complexes deposited in the PDB where: (a) many ligand fragments have no interactions with the protein, or even interactions that are clearly repulsive (obviously, in both cases, the energy loss due to these "bad" interactions must be compensated for by strong attractive interactions formed by other fragments in the same ligand); and (b) the coordinates of the ligands do not always correspond to their most stable conformations.

— A VLS experiment is speeded up by reusing ligand fragments that have previously docked onto the same receptor site. This takes advantage of the fact that eHiTS® docks all rigid fragments in a ligand independently and anywhere in the receptor binding site. Therefore, the results from ligand-rigid fragments that have previously docked onto the same binding site can be reused when they are also present in a new ligand. This saves a lot of calculation time in VLS because there are many rigid fragments in "drug-like" molecules that are repeated many times in ligand databases. In this respect, it has been shown that by using this feature to dock around 3000 ligands, eHiTS® can reduce the time to dock a ligand from 2.2 to 1 minute [36]. Moreover, eHiTS® comes with built-in support for distributed architectures and grid computing, which allows parallel execution.

— As mentioned, the program also has an algorithm that automatically identifies probable binding sites in PDB structures. This algorithm searches for the independent pockets located in the receptor and calculates their depth and volume. Then, this information is used to select which of the pockets is the correct binding site. The performance and reliability of this method has been tested in over 300 experimental complexes and the results show that the correct pocket is identified in 99% of the test cases [36]. This feature allows the eHiTS® software to be used at the leading edge of drug discovery science by docking potential drug candidates against new therapeutic targets whose function and active site is not yet known [36].

— The scoring function used by eHiTS® (*i.e.* eHiTS_Score) is a hybrid between empirical and knowledge-based functions that is based on the interactions between points of chemical interest on the surface of the ligand and the receptor [36]. Thus, complementary ligand-receptor surface points receive a positive score and repulsive ligand-receptor surface points receive a penalty score. Moreover, as well as this ligand-receptor complementarity, other components are also used in the final scoring function that are known to be important in the intermolecular interactions: for example, (a) steric clashes; (b) depth within the binding pocket; (c) the covering of the receptor surface area (exposed/buried hydrophobic area); (d) the conformational strain energy of the

ligand, (e) the intra-molecular interactions within the ligand; (f) the family-coverage; and (g) the entropy loss of the ligand due to the *freezing* of its rotatable bonds when the complex is formed with the receptor. Interestingly, when statitistics were collected from pairwise interaction geometries (*e.g.* hydrogen bonds) to build eHiTS_Score, the *quality* of the coordinates in the crystal structures was seen to be not homogeneous (*i.e.* coordinates with low temperature factors are more reliable than those with high ones; where the temperature factor is a measure of how much the atom position varies relative to its coordinates in the PDB file). Thus, eHiTS® uses the probability of the atom position during the collection of statistics to create a statistically derived empirical scoring function. Then, reliable interactions statistics were collected from a PDB set of 1420 protein-ligand complexes obtained at a resolution equal or better than 2.5 Å. In this set, all ligand-protein atom pairs within 5.6Å of each other were used to gather the interaction information (giving a total of about 10 million interacting atom pairs).

— Even the best scoring function currently available is clearly biased towards the receptors that were used to derive it. Therefore, it is nearly impossible to obtain a scoring function that represents all the proteins equally well. The terms of the eHiTS® scoring function are combined using weights that can be adjusted according to the protein family (a concept that is based only on the shape of the active site and on the types of residues that make it up). Hence, eHiTS® is shipped with 72 weight sets (71 family-based and 1 default weight set to be used for unrecognized receptors) that were built by classifying 1315 protein-ligand complexes downloaded from the PDB into protein families. Therefore, when eHiTS® starts to dock ligands onto a receptor, the program initially (and automatically) identifies the protein family to which the receptor belongs and selects its corresponding weight set (or the default weight set is used when a matching family is not found). eHiTS® also has training scripts that allow users to fine-tune the scoring function to better suit their systems of interest and exploit the available experimental data. Training utilities are provided that allow both validation and enrichment training. Validation training aims to improve the accuracy of the binding mode prediction capabilities of eHiTS® and it uses several co-crystallized complexes of the same protein family that are provided by the user [an optional set of decoys (*i.e.* ligands that do not bind to the receptor) can also be provided]. Then, eHiTS® will adjust the weights in the scoring function to better replicate the experimental ligand poses. Therefore, when members of this protein family are used as the receptors of a docking experiment, eHiTS® will have a more suitable scoring function with which to perform the calculations. On the other hand, enrichment training focuses on separating active and inactive ligands and uses known binders that have been determined experimentally

(*i.e.* ligands that bind to the receptor) and decoys to adjust the weights in the scoring function to help the active ligands to outscore the decoys (*i.e.* to rank the active ligands higher than the inactives ones). It has been shown that this tailoring of the scoring function can dramatically improve the ability of eHiTS® to identify the correct binding pose or to distinguish active ligands from decoy ones [36].

— VHTS can also be done in a very effective way with eHiTS LASSO® (where *LASSO* stands for ***L**igand **A**ctivity by **S**urface **S**imilarity **O**rder*), which is part of the eHiTS® docking package. Thus, LASSO® provides a very easy-to-use training utility that can capture the chemical features of a limited set of known active molecules on the target of interest. Further, it uses this information to quickly screen for other potentially active molecules with similar features in large databases (*i.e.* it can screen 1 million ligands per minute). Interestingly, LASSO® is conformation-independent because it only uses the surface properties of the potential ligands (and not their 2D topology or their 3D conformation) to do the screening.

GOLD™ (**G**enetic **O**ptimisation for **L**igand **D**ocking; see Table 4)

GOLD™ [55, 56, 84] was originally developed by the University of Sheffield (http://www.shef.ac.uk/), GlaxoSmithKline plc (http://www.gsk.com/) and the CCDC (**C**ambridge **C**rystallographic **D**ata **C**entre; http://www.ccdc.cam.ac.uk/). At present, CCDC is developing and maintaining GOLD™ in an ongoing project with Astex Therapeutics (http://www.astex-therapeutics.com/) and Syngenta (http://www.syngenta.com). GOLD's main strengths are that: (a) some backbone and side chain flexibility (for the moment, up to ten user-defined residues) can be included in the calculations; (b) the program can use user-defined scoring functions and not only the two default ones provided with GOLD™ (and even modify them); (c) the energy functions are partly based on conformational and non-bonded contact information derived from the Cambridge Structural Database (http://www.ccdc.cam.ac.uk/products/csd/) [160]; (d) a variety of constraint options can defined for the docking; (e) crystallographic water molecules in the ligand binding site can be considered during the docking; (f) it is optimized for parallel execution on processor networks and a distributed version of GOLD™ is available for use on commercial grid systems; (g) it can handles metal atoms automatically if they are set up correctly in the protein input file; and (h) VHTS results can be analyzed and post processed easily with the companion programs SILVER™ or GoldMine™. On the other hand, GOLD's main limitations are that: (a) it uses a random method (*i.e.* a GA) instead of a systematic one to search for the ligand poses in the protein-ligand complex; (b) the default scoring functions have been optimized to predict the ligand poses not the binding affinities; (c) systematic problems in ranking very polar ligands and general ligands in large cavities have been reported [161]; (d) the

protein and ligand require a previous set up by the user that includes, for instance, the addition of all the hydrogen atoms; and (e) GOLD™ does not vary tautomeric or ionization states during docking and, therefore, if the user is not sure about the correct tautomeric state of a particular residue, then separate GOLD™ runs must be performed with all the possibilities.

Comments on the main strengths of GOLD™:

- When the GoldScore scoring function is used (*i.e.* the default one), GOLD™ can allow side chain flexibility or localized backbone movements for up to 10 user-defined active site residues. Thus, each selected flexible side chain will be allowed to rotate around one or more of its acyclic bonds during docking by using exact or certain tolerance torsions. A rotamer library that compiles the most commonly observed side-chain conformations for the naturally occurring amino acids can also be used [162]. The flexible side chain movements produced during the docking can be viewed with either SILVER™ or GoldMine™.
- GOLD™ also rotates the torsion angles of the Ser, Thr and Tyr hydroxyl groups and the positions of the lysine NH_3^+ groups during docking in order to optimize the H-bonding interactions of these residues with the ligand. Therefore, the starting positions of Ser, Thr and Tyr OH groups and Lys NH_3^+ groups do not matter.
- Extensive options for customising or implementing new scoring functions are available through the *Scoring Function Application Programming Interface* that allows users to implement their own scoring function or to enhance existing ones (*i.e.* by calculating and writing additional data after each docking and/or by adding extra terms to the scoring function).
- GOLD™ supports a variety of constraints which allow the user to bias the solution towards certain intermolecular interaction requirements (and one or more constraints can be simultaneously considered). These constraints are on: (a) the distance between a pair of specific protein and ligand atoms; (b) the required presence of hydrogen bonds between protein and ligand; (c) the bias of the docking towards solutions in which particular regions of the binding site are occupied by specific ligand atoms or types of ligand atoms (*e.g.* hydrophobic atoms); or (d) the bias of the conformation of docked ligands towards a given solution or template. It can also be decided not to dock a ligand when one of the selected constraints is physically impossible.
- Crystallographic water molecules in the ligand binding site may be set up with the following options, which will be applied during the docking: (a) they can be present (*i.e.* used); (b) they can be absent (*i.e.* not used); and (c) GOLD™ can decide whether

ISBN:978-84-691-1553-4 /DL: T.151-2008

the water should be present or absent. GOLD™ can also rotate the water hydrogen atoms around their three principal axes in order to optimize hydrogen bonding during docking.

— SILVER™ or GoldMine™ enables GOLD's docking results to be visually inspected and descriptors to be computed from the self ligand poses that characterize and analyze how each pose interacts with the receptor binding site. The descriptors are the following: (a) general contact descriptors (for detecting whether contacts or clashes occur between specified sets of protein atoms and selected types of ligand atoms); (b) hydrogen-bonding descriptors (for monitoring whether particular atoms or groups of atoms in the protein binding site are H-bonded to the ligand; where the geometric hydrogen-bond criteria can be tailored, if needed); (c) accessible surface area descriptors (for determining the loss of the ligand's solvent-accessible area upon binding to the protein; either for the ligand as a whole or for a specific selection of atoms); and (d) simple property descriptors (like the molecular weight; the number of rotatable bonds; the ligand H-bond donor/acceptor count; the number of ligand atoms H-bonded to proteins; the number of ligand atoms clashing with proteins; the number of ligand/protein atoms not fulfilling their H-bonding capacity; a count of solvent-exposed hydrophobic ligand atoms, etc.). All these descriptors can be easily set up and combined through a graphical front-end and may be used, for instance, to filter out unpromising ligands and to facilitate the selection of suitable drug leads from docking results.

Molegro Virtual Docker (**Mol**ecule & All**egro** Virtual Docker; see Table 4)

MVD™ (formerly called MolDock [37]) is being developed by Molegro ApS (Aarhus, Denmark). MVD's main strengths are that it: (a) can automatically set up the input structures by assigning charges, bond orders and hybridization and by adding hydrogens; (b) is able to automatically predict potential binding sites in the receptor; (c) can deal with receptor sidechain flexibility by taking into account induced fit interactions; (d) is able to dock in precalculated energy grids (which speeds up the calculations); (e) can deal with user-defined constraints during docking; (f) can benefit from the use of templates (*i.e.* pharmacophores) during docking; and (g) is able to distribute the calculations on multiple computers. On the other hand, MVD's main limitations are that: (a) the stochastic nature of the docking engine implies that more than one docking run may be needed to identify the correct binding mode; and (b) the protonation state of the ligand/receptor pair is not exhaustively searched and pre-fixed protonation states are used during the docking.

Comments on the main strengths of MVD™:

— It does not require user intervention to import and prepare the input molecules so that they can be used during docking. Hence, MVD™ automatically: (a) assigns bonds [two atoms are connected if their distance is more than 0.4Å and less than the sum of their covalent radii plus a threshold of 0.45Å (this threshold is set to 0.485Å if one of the atoms is phosphorus)]; (b) assigns bond orders (*i.e.* whether bonds are single, double or triple) and hybridization (*i.e.* sp, sp2 and sp3); (c) detects aromatic rings; (d) assigns partial charges; (e) adds explicit hydrogens; (f) detects flexible torsions in ligands; and (g) identifies cofactors [a molecule is considered a cofactor if it has less than five heavy atoms or its name is included in a list of common cofactor names (like 'HEM', 'SO4', 'PO4', etc.); although cofactor recognition can be overridden if this is not desired]

— Constraints are limitations imposed on the molecular system and based on chemical insight or knowledge. MVD™ considers two kinds of constraints: (a) hard constraints (that the docking engine tries to fully satisfy and which can be used to force a specific ligand atom to be in a given region); and (b) soft constraints (that act as extra energy terms contributing to the overall energy of the system and which can be more or less satisfied). In this respect, MVD™ will enforce hard constraints by moving or modifying the poses during docking. If several hard constraints exist and they are not simultaneously satisfied, then MVD™ will choose to satisfy an arbitrary one (when a hard constraint is not satisfied, it will add 100 units to the soft constraint energy penalty). On the other hand, soft constraints can be used, for instance, to favour ligands in certain regions. If there are several soft constraints, they will all be taken into account (*i.e.* several extra terms are added to the overall docking energy function while docking). Therefore, by using these two kinds of constraints, the energy landscape can be altered either by rewarding certain regions of the search space or by forcing some interactions or preventing others from occurring. The constraints can be defined either on the basis of general chemical properties (*e.g.* hydrogen bond donors, hydrogen bond acceptors or ring atoms) or by specifying individual atoms for each ligand.

— Sidechain conformational changes can be worked with in two ways: (a) by softening the steric, hydrogen bonding and electrostatic-force potentials used during the docking simulation and, therefore, simulating flexibility in the ligand binding site (a kind of *induced fit*); and (b) by defining the residues whose sidechain torsional angles can change during the docking simulation (the backbone is kept rigid). After setting up sidechain flexibility, the next sequential steps are taken during the docking simulation: (1) the ligands are docked to the receptor as if it were rigid (but using the softened potentials); and (2) the sidechains that have been set up to be flexible are minimized relative to the ligand pose (this is done with standard non-softened potentials). It should

ISBN:978-84-691-1553-4 /DL: T.151-2008

also be pointed out that when this MVD™ feature is used, the user has to work with the "MolDock Score [GRID]" scoring function because the other one (*i.e.* "MolDock Score") always uses unsoftened potentials (which take into account the hydrogen bond directionality).

— Template docking can be used if the 3D structures of complexes between the receptor of interest and some ligands are known. Hence, from the conformations of the ligands in these complexes, a template can be created with features expected to be relevant for the binding. This allows the search algorithm to focus on those poses that are similar to the docking template. By using templates, it is also possible to align molecules and then, from the overlap in each individual template point, extract detailed information about the similarity. Then, this information can be used by the Data Analyzer window to create a regression model against some known empirical quantity (*i.e.* 3D-QSAR).

AutoDock (see Table 4)

AutoDock (formerly called AutoDoq since it used quaternions to perform rotations) [49, 57, 163, 164] was the first docking package to model the ligand with full conformational flexibility. It is being developed by Prof. Arthur J. Olson at the Department of Molecular Biology of the Scripps Research Institute in La Jolla (California, USA) and, at present, it is the most cited docking software in the bibliography [79]. The package consists of two programs (*i.e.* AutoGrid and AutoDock), which are sequentially applied. AutoGrid is initially used to calculate the non-covalent energy of interaction between the rigid part of the receptor and a probe atom that is located at the various grid points of the lattice (where the probe atoms are those that are present in the ligands that will be docked onto the receptor). Therefore, AutoGrid builds as many files (*i.e.* the so-called *grid-map files*) as the number of probe atoms used. In addition to these atomic affinity grid maps, AutoGrid also generates: (a) an electrostatic potential grid map by using either a point charge of +1 as the probe or a Poisson-Boltzmann finite difference method such as DELPHI [165]; and (b) a desolvation map. Then the full set of grid maps and the flexible part of the receptor are used by AutoDock to guide the docking process of the selected ligands through the lattice volume. AutoDock's main strengths are that: (a) it can deal with receptor flexibility during docking; (b) it can be used in blind-docking when the exact location of the ligand binding site is unknown; (c) it uses pre-calculated grid maps on a binding site in order to dock ligands with the consequent saving of computational time; (d) the atomic grid maps can also be used as a guide to design new ligands that bind with more affinity to the receptor; (e) it has a free-energy scoring function that is based on a linear regression analysis, the AMBER force field, and a large set of protein-ligand complexes with known inhibiton constants; and (f) it provides good correlations between predicted inhibition constants and experimental ones. In

constrast, AutoDock's main limitations are that: (a) the protein and the ligand need to be set up before being used by AutoGrid/AutoDock [*e.g.* hydrogens and partial charges have to be added, non-polar hydrogens and their charges have to be merged with their parent carbon atom, rotatable bonds in the ligand have to be set up, etc.; most of this work can be done with the AutoDockTools (http://autodock.scripps.edu/downloads/resources/adt)]; (b) its different pose search algorithms are stochastic and not systematic; and (c) the protonation state of the ligand/receptor pair is not exahustively searched and pre-fixed protonation states are used during the docking.

Comments on the main strengths of AutoDock:

- AutoDock can consider the receptor as flexible during docking by means of two different strategies. The first one is based on the work of Fredrik Österberg [152] and consists of: (a) obtaining the coordinates from different *snapshots* of the receptor that are the product of its intrinsic flexibility [where the *snapshots* can be obtained from the PDB files of the same protein that have been crystallized in different conditions, obtained from FlexWeb simulations (http://flexweb.asu.edu/) [166] or retrieved from the MODEL database (http://mmb.pcb.ub.es/MODEL)]; (b) then, using each receptor *snapshot* with AutoGrid to obtain its grid maps; (c) next, combining all the equivalent grid maps that correspond to different receptor conformations to obtain a single grid-map file [where: (1) this step is repeated for each set of equivalent grid maps; and (2) the energy of each grid point is obtained from a weighted average of the energies of the same point in all the original conformational-dependent grid maps and the weights are calculated using either a clamped grid or a Boltzmann assumption based on the interaction energy]; and (d) finally, using these consensus grid maps with AutoDock. This methodology has been shown to be consistent and accurate for protein-ligand docking [152] but, unfortunately, it is not easy to apply from the self AutoGrid/AutoDock package without strong computational skills. In order to avoid, among others, this limitation, we have designed BDT (http://www.quimica.urv.cat/~pujadas/BDT) [158], a front-end application to AutoGrid/AutoDock v3.05. The second strategy considers the receptor as flexible during docking and is only possible with v4.0 of AutoGrid/AutoDock. It consists of: (1) dividing the receptor into two parts (*i.e.* one that is considered to be rigid and another that contains all the residues that are known to move during the ligand binding); (2) then, running AutoGrid with the rigid receptor part; and (3) finally, using the flexible receptor residues during the AutoDock run.
- Blind docking is used when there is no information about the possible ligand binding site in the receptor and so nothing can be assumed about its location (*i.e.* the complete receptor surface is considered to be, potentially, part of the ligand binding site). Several

ISBN:978-84-691-1553-4 /DL: T.151-2008

authors have succeeded in using AutoGrid/AutoDock to perform blind docking [167-172]. There are two ways of doing a blind docking experiment with AutoGrid/AutoDock. The first one is to use AutoGrid to make a single grid that covers the entire receptor surface. This involves using: (a) the maximum number of points allowed per dimension that the program can deal with; and (b) enough grid-point separation to ensure that the whole receptor surface is completely covered by the grids. Then, the resulting grid maps are used by AutoDock to find, roughly, the preferred ligand-binding areas (the ones around the ligand conformations with the highest affinity for the receptor). Subsequently, new grid maps are built around these preferred ligand-binding site locations but with a shorter grid-point separation. Finally, the thinner grid maps obtained around these potential ligand-binding areas are used by AutoDock and results should be better because the energy maps better reproduce the continuum of energy that corresponds to the non-covalent interaction between receptor and ligand [169]. The second way of doing a blind docking experiment with AutoGrid/AutoDock is to use BDT [158]. The advantages of using BDT over the former blind-docking strategy are that: (a) the user can use any value for the grid-point distance; and (b) there is no need of human intervention during the process (*i.e.* no preliminary result is discarded during the run). Moreover, by using BDT, as many ligands as wanted can be docked with AutoGrid/AutoDock in a single set up step (either in a blind-docking experiment or when a well-defined ligand binding site is used). At present, BDT works only with AutoGrid/AutoDock v3.0.5. More information about how BDT deals with blind-docking can be found on its website (http://www.quimica.urv.cat/~pujadas/BDT).

Glide™ (**G**rid-based **Li**gand **D**ocking with **E**nergetics; see Table 4)

Glide™ [109, 110] is being developed by Schrödinger (Portland, Oregon, USA). Its main strengths are that: (a) it not only considers that the receptor is flexible but also that the receptor can change its conformation upon ligand binding due to an induced fit mechanism; (b) it can use user-defined constraints for restricting a ligand atom to lie within a certain region that is defined in relation to the features of the receptor that are responsible for ligand binding; (c) it can be used together with two more of Schrödinger's products (*i.e.* Liaison™ or Qsite™) to obtain more accurate binding energies for ligand-receptor pairs; (d) it can consider water molecules in the receptor's active site during docking or not; and (e) it has parallel processing or distribution options. On the other hand, Glide's main limitations are that: (a) PDB receptors have to be manually prepared (*i.e.* atom types and bond orders must be assigned, the charge and protonation states must be corrected, side chains reoriented if necessary, incomplete side chains in the active site have to be rebuilt, and steric clashes relieved); and (b) the protonation state of

the ligand/receptor pair is not exahustively searched and pre-fixed protonation states have to be used during the docking.

Comments on the main strengths of Glide™:

— In standard virtual docking studies, ligands are docked into the binding site of a receptor by holding the receptor rigid and leaving the ligand free to move. However, the assumption of a rigid receptor can give misleading results, since in fact many proteins undergo side chain or backbone movements, or both, upon ligand binding. These changes allow the receptor to alter its binding site structure so that it can conform more closely to the shape and binding mode of the ligand. This is often referred to as the "induced fit" mechanism [173] and is one of the main complicating factors in structure-based drug design. In order to take an induced fit process into account, Glide™ has to be used together with the refinement module of another of Schrödinger's product: Prime™ in the so called *Induced Fit* protocol [174]. This protocol simulates the receptor's changes upon ligand binding by using Glide™ to dock the ligand into the ligand-binding site. During this docking process it is important to generate a large and varied set of ligand poses by means of a softened potential. This can be done by: (1) reducing the van der Waals radii; (2) using an increased Coulomb-vdW cutoff; and (3) temporarily removing highly flexible side chains. Thus, by default, a maximum 20 poses per ligand are retained, and the poses to be retained must have a Coulomb-vdW score less than 100 and an H-bond score less than -0.05 (default values). Then, for each resulting pose, Prime™ uses its structure prediction capabilities to accommodate the ligand by reorienting nearby residue side chains (by default the residues affected by this Prime™ side-chain prediction are located 5.0 Å from any ligand pose). Subsequently, Prime™ minimizes the corresponding set of residues and ligands from each protein/ligand complex pose and therefore, the resulting receptor structure now reflects an induced fit with the ligand structure and conformation. Finally, Glide™ redocks each ligand into the induced-fit receptor structure by using a more rigorous methodology (*i.e.* the default Glide™ settings in non-induced fit dockings) and the resulting complexes are ranked. The ability to model the induced fit mechanism during docking has had two main applications: (1) to generate the structure of the protein-ligand complex when it is known that the ligand is active but cannot be docked in the receptor if this is considered to be rigid; and (2) to recover, for further use in VHTS experiments, the false negative ligands that are obtained when docking in a single conformation of the receptor but which have higher scores if the docking is performed against a set of receptor conformations that are obtained from the induced fit protocol. A Python script

ISBN:978-84-691-1553-4 /DL: T.151-2008

automates the complete induced fit docking process and it can be accessed either through Maestro™ (*i.e.* the Schrödinger's suite graphic interface) or from the command line. Some examples of the succesful use of this induced fit protocol for predicting ligand binding modes and the concomitant changes in the receptor structure have recently been published [174-179].

— A *Glide constraint* is a ligand-receptor interaction requirement. This means that a ligand atom must lie within a certain region that is defined in relation to the features of the receptor that are responsible for ligand binding. There are four types of constraints in the current version of Glide™: (1) positional constraints (*i.e.* a requirement that one or more ligand atoms occupy a spherical volume that is centered at a particular position); (2) hydrogen bond constraints (*i.e.* a requirement that a particular receptor-ligand hydrogen bond be formed); (3) metal constraints (a requirement that a particular metal-ligand interaction be present when the ligand is docked); and (4) hydrophobic constraints (*i.e.* a requirement that a user-defined number of hydrophobic heavy atoms in the ligand occupy a hydrophobic region in the active site). To use constraints within Glide™, before the grids are generated in the receptor binding site, it is necessary to specify the sites for possible ligand interactions (where there is a maximum of ten constraints that can be defined for a given grid, distributed among the above mentioned four constraint types). Thus, when a further docking run is started with a set of ligands, it is possible to select which constraints will be applied from the list of receptor constraint sites that are previously defined in this receptor. Then, Glide™ takes into account whether the selected constraints have been complied with by using several hierarchical filters during pose selection so that those docked poses that fail to meet the selected requirements can be promptly rejected. For instance, the first constraint filter simply takes into account whether ligand atoms can comply with the imposed requirement in terms of intermolecular interaction. If it cannot (*e.g.* if a selected receptor constraint atom is a polar hydrogen and the ligand has no hydrogen-bond acceptors), then Glide™ simply leaves that ligand out of the docking run. However, if some of the atoms in the ligand qualify, Glide™ keeps a list of the possible "partner atoms" for each constraint, so that they can be used in subsequent filters.

Which docking software gives better results?

The comparison of the predictive power and the performance of protein-ligand docking programs is currently an active area of research [180-184]. This predictive power is usually evaluated by studying how the studied software performs relative to other docking programs in: (1) re-docking assays in which crystallographic protein-ligand complexes are split and the

resulting ligand is docked into the resulting protein [44, 46, 54, 56, 60, 68, 77, 84, 105, 106, 161, 185-190]; (2) enrichment studies in which the software has to distinguish between true binders and decoys for a protein target in virtual screening studies [65, 70, 84, 105, 106, 189, 191-193]; and (3) predicting binding free energies from the best-scored pose [60, 65, 70, 71, 84, 125, 191, 194, 195]. Nevertheless, comparing the performance of docking programs is surprisingly complex and many factors have to be considered before the *best one* can be chosen [196]. No program has yet been found that offers robust and accurate solutions of general applicability to a majority of the various docking problems [78].

For instance, the results of re-docking assays are only meaningful for comparison if all the programs are run in equivalent conditions [196]. Thus, (a) all the programs must be validated with the same set of protein complexes (*i.e.* the so called *test set*); (b) all the programs must search the same 3D space for the ligand pose (this is difficult because not all the programs define in the same way the search space where the ligand docking will be assayed and this is important for the success of the docking [197]); (c) the correctness of the top-one poses must be carefully compared with the experimental solutions and not based on *blind* RMSD measures; (d) all ligands must be prepared in the same way [196, 198]; and (e) the time allowed to dock a ligand must be the same for all the programs (because one program may be better than another when fast settings are used but worse with slow settings). As far as this last point is concerned, it is not unusual for comparisons to allow one algorithm much more time to find the top-one pose than the rest [135] (although other studies carefully select settings that give comparable time per ligand for each program [65, 199]). Moreover, it has been well established that docking programs can better predict the complexes in some proteins than in others and that, therefore, they are sensitive to the target structure [65, 136, 182, 197, 200]. Hence, any comparison of the predictive power of docking programs by means of re-docking assays has to be made with a test set of proteins that is large enough to cover as many different families as possible. This will make it possible to determine whether the evaluated program can or cannot be of general use for a varied set of protein targets. Furthermore, the experimental complexes that are part of this test set have to be reliable (*i.e.* the experimental pose of the ligand has to be unambiguously defined within the complex) because it has been shown that some docking programs apparently fail because of re-docking experiments with low quality structures [201]. In this respect, it is important that the structures in the test set do not include the ones that comply with any of the following exclusion criteria: (a) low resolution; (b) high temperature factors (*i.e.* larger than 50 $Å^2$) and occupancies lower than 1 for the ligand atoms [202]; (c) incorrect features (*e.g.* contacts too close between protein side chains and ligand); (d) an incorrect fit of the ligand structure in the corresponding electron density; and (e) the presence of neighboring chains or ligands in the crystal-packing environment that are close enough to the binding site to influence the ligand-

binding geometry. Additionally, it is advisable that the test set does not include complexes on which any of the docking programs to be evaluated were trained (where "training" accounts for any change made to the program to improve the success rate [196]). At present, a test set built with these criteria is freely available from the Cambridge Crystallographic Data Centre and Astex (http://www.ccdc.cam.ac.uk/products/life_sciences/validate/astex/). This curated set contains 305 high-quality complexes that can be used to validate docking programs by means of re-docking assays [201]. Another additional problem, as already mentioned, is the *blind* use of RMSD values for evaluating the success of a docking program. Thus, in re-docking assays, the predictive power of the sofware is usually evaluated by: (1) calculating, for each re-docked ligand, the RMSD for the coordinates of its heavy atoms in the experimentally-observed pose and in the top-scored predicted one; and (2) summarizing the results for all the experimental re-docked complexes by indicating the percent of comparisons where the above RMSD value is below a user-defined threshold (usually, 2.0 Å [44, 125, 182]). Therefore, the higher this percent is, the better the predictive power of the docking software should be. This methodology is widely used because RMSD values are very easy to calculate. Nevertheless, it has been reported that, in some situations, it can be misleading because: (1) two very different poses from small ligands can give low RMSD values; and, (2) in contrast, RMSD can be high even when the essential intermolecular binding features of the poses with the protein are correct. Therefore, it has been suggested that RMSD is not always a proper measure of the docking accuracy and that it should be measured with parameters that take into account if the top-one predicted pose keeps the key intermolecular interactions that are found in the experimental complexes [203]. One example of this kind of parameters is IBAC (*i.e.* **i**nteractions-**b**ased **a**ccuracy **c**lassification) [203]. Unfortunately, this kind of interaction-based measures are not error free because they are more *subjective* than RMSD. For instance they are very dependent on the intermolecular interactions being correctly defined (which is not always easy to do, specially for those interactions that are weak and have no directional preference). Therefore, their advantages over RMSD-validation are not always clear [196]. Consequently, the following methodology has been recommended for evaluating docking programs based on re-docking tests [196]: (1) start the evaluation by counting as a success any re-docked complex whose RMSD is lower than 1.5 (for small ligands) or 2 Å (for medium or large ligands); (2) continue by refining the previously found set by means of interaction-based measures and/or visual inspection of the results (in order to ensure that solutions are only counted as successes if all the key intermolecular interactions are reproduced); and (3) use the resulting set to calculate the percent of success of the corresponding docking program.

Therefore, it can be concluded that there is no docking method nowadays that clearly outperforms the others and the results reported so far depend on the target on which the methods are applied, and on the experimental structure used to perform the test [65, 70, 78, 133, 191, 197].

Future goals of docking

In this review, we have highlighted some of the main limitations of current docking programs. Basically, none of them is perfect but, collectively, it is true that most of the key challenges in protein-ligand docking seem to be correctly addressed. For instance, some programs (*e.g.* Glide™/Prime™ [109, 110]) can now succesfully address the induced fit changes in protein conformation that are the result of ligand binding in some proteic targets. Other programs (*e.g.* eHiTS® [36, 159]) can fine-tune the scoring function to better suit the targets of interest and exploit the experimental data that is made available. Others (*e.g.* eHiTS® [36, 159]) can exahustively search the conformational space for the ligand poses. Finally, another one (*i.e.* eHiTS® [36, 159]) can evaluate all possible protonation states of the ligand/receptor pair during docking without any previous user set-up of either the receptor or the ligand. Therefore, while waiting for the *perfect* program, we strongly recommend the use of Glide™/Prime™ [109, 110] and eHiTS® [36, 159].

Furthermore, we encourage additional efforts to extend the use of docking tools beyond the confines of bioinformatic groups. This involves more than the availability of a graphic interface. For instance, the receptor and the ligand set-up may need to be done without any user-intervention (*i.e.* the whole set-up occurs internally and the user only needs to provide the files with the coordinates for the targets and the ligands). In this respect, of all the programs that have been evaluated in this review, only eHiTS® [36, 159] and MVD™ [37] can do the receptor and ligand set-up internally (see Table 4).

ACKNOWLEDGMENTS

We thank John Bates of our University's Language Service for correcting the manuscript. This study was supported by grant number CO3/O8 from the Fondo de Investigación Sanitaria (FIS) and AGL2005-04889 from the Comisión Interministerial de Ciencia y Tecnología (CICYT) of the Spanish Government. Montserrat Vaqué is the recipient of a fellowship from grant number CO3/O8.

ISBN:978-84-691-1553-4 /DL: T.151-2008

REFERENCES

[1] R.A. Kumar, D.S. Clark, High-throughput screening of biocatalytic activity: applications in drug discovery., Curr Opin Chem Biol 10 (2006) 162-168.

[2] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The Protein Data Bank., Nucleic Acids Res 28 (2000) 235-242.

[3] G. Wang, R.L.Jr. Dunbrack, PISCES: recent improvements to a PDB sequence culling server., Nucleic Acids Res 33 (2005) W94-98.

[4] C.H. Wu, R. Apweiler, A. Bairoch, D.A. Natale, W.C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M.J. Martin, R. Mazumder, C. O'Donovan, N. Redaschi, B. Suzek, The Universal Protein Resource (UniProt): an expanding universe of protein information., Nucleic Acids Res 34 (2006) D187-191.

[5] T.L. Blundell, H. Jhoti, C. Abell, High-throughput crystallography for lead discovery in drug design., Nat Rev Drug Discov 1 (2002) 45-54.

[6] K. Arnold, L. Bordoli, J. Kopp, T. Schwede, The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling., Bioinformatics 22 (2006) 195-201.

[7] A. Fiser, A. Sali, Modeller: generation and refinement of homology-based protein structure models., Methods Enzymol 374 (2003) 461-491.

[8] N. Guex, M.C. Peitsch, SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling., Electrophoresis 18 (1997) 2714-2723.

[9] U. Pieper, N. Eswar, F.P. Davis, H. Braberg, M.S. Madhusudhan, A. Rossi, M. Marti-Renom, R. Karchin, B.M. Webb, D. Eramian, M.Y. Shen, L. Kelly, F. Melo, A. Sali, MODBASE: a database of annotated comparative protein structure models and associated resources., Nucleic Acids Res 34 (2006) D291-295.

[10] J. Kopp, T. Schwede, The SWISS-MODEL Repository: new features and functionalities., Nucleic Acids Res 34 (2006) D315-318.

[11] C. Bissantz, P. Bernard, M. Hibert, D. Rognan, Protein-based virtual screening of chemical databases. II. Are homology models of G-Protein Coupled Receptors suitable targets?, Proteins 50 (2003) 5-25.

[12] A. Hillisch, L.F. Pineda, R. Hilgenfeld, Utility of homology models in the drug discovery process., Drug Discov Today 9 (2004) 659-669.

[13] C. Oshiro, E.K. Bradley, J. Eksterowicz, E. Evensen, M.L. Lamb, J.K. Lanctot, S. Putta, R. Stanton, P.D. Grootenhuis, Performance of 3D-database molecular docking studies into homology models., J Med Chem 47 (2004) 764-767.

[14] W.M. Rockey, A.H. Elcock, Structure selection for protein kinase docking and virtual screening: homology models or crystal structures?, Curr Protein Pept Sci 7 (2006) 437-457.

[15] V. Kairys, M.X. Fernandes, M.K. Gilson, Screening drug-like compounds by docking to homology models: a systematic study., J Chem Inf Model 46 (2006) 365-379.

[16] J.V. Hobrath, S. Wang, Computational elucidation of the structural basis of ligand binding to the dopamine 3 receptor through docking and homology modeling., J Med Chem 49 (2006) 4470-4476.

[17] V. Kairys, M.K. Gilson, M.X. Fernandes, Using protein homology models for structure-based studies: approaches to model refinement., ScientificWorldJournal 6 (2006) 1542-1554.

[18] P. Ferrara, E. Jacoby, Evaluation of the utility of homology models in high throughput docking., J Mol Model 13 (2007) 897-905.

[19] D. Weininger, SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, J. Chem. Inf. Comput. Sci. 28 (1988) 31-36.

[20] J.H. Voigt, B. Bienfait, S. Wang, M.C. Nicklaus, Comparison of the NCI open database with seven large chemical structural databases., J Chem Inf Comput Sci 41 (2001) 702-712.

[21] M. von Grotthuss, G. Koczyk, J. Pas, L.S. Wyrwicz, L. Rychlewski, Ligand.Info small-molecule Meta-Database., Comb Chem High Throughput Screen 7 (2004) 757-761.

[22] R.L. Strausberg, S.L. Schreiber, From knowing to controlling: a path from genomics to drugs using small molecule probes., Science 300 (2003) 294-295.

[23] M. Kanehisa, S. Goto, M. Hattori, K.F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, M. Hirakawa, From genomics to chemical genomics: new developments in KEGG., Nucleic Acids Res 34 (2006) D354-357.

[24] T. Fink, J.L. Reymond, Virtual exploration of the chemical universe up to 11 atoms of C, N, O, F: assembly of 26.4 million structures (110.9 million stereoisomers) and analysis for new ring

systems, stereochemistry, physicochemical properties, compound classes, and drug discovery., J Chem Inf Model 47 (2007) 342-353.

[25] J.J. Irwin, B.K. Shoichet, ZINC--a free database of commercially available compounds for virtual screening., J Chem Inf Model 45 (2005) 177-182.

[26] R. Guha, M.T. Howard, G.R. Hutchison, P. Murray-Rust, H. Rzepa, C. Steinbeck, J. Wegner, E.L. Willighagen, The Blue Obelisk-interoperability in chemical informatics., J Chem Inf Model 46 (2006) 991-998.

[27] M. Brylinski, M. Kochanczyk, E. Broniatowska, I. Roterman, Localization of ligand binding site in proteins identified in silico., J Mol Model 13 (2007) 665-675.

[28] S. Soga, H. Shirai, M. Kobori, N. Hirayama, Use of Amino Acid Composition to Predict Ligand-Binding Sites., J Chem Inf Model 47 (2007) 400-406.

[29] G. Koczyk, L.S. Wyrwicz, L. Rychlewski, LigProf: a simple tool for in silico prediction of ligand-binding sites., J Mol Model 13 (2007) 445-455.

[30] A.T. Laurie, R.M. Jackson, Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites., Bioinformatics 21 (2005) 1908-1916.

[31] D.T. Chang, Y.Z. Weng, J.H. Lin, M.J. Hwang, Y.J. Oyang, Protemot: prediction of protein binding sites with automatically extracted geometrical templates., Nucleic Acids Res 34 (2006) W303-309.

[32] J. Dundas, Z. Ouyang, J. Tseng, A. Binkowski, Y. Turpaz, J. Liang, CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues., Nucleic Acids Res 34 (2006) W116-118.

[33] D.T. Chang, Y.J. Oyang, J.H. Lin, MEDock: a web server for efficient prediction of ligand binding sites based on a novel optimization algorithm., Nucleic Acids Res 33 (2005) W233-238.

[34] G.P.J. Brady, P.F. Stouten, Fast prediction and visualization of protein binding pockets with PASS., J Comput Aided Mol Des 14 (2000) 383-401.

[35] F. Glaser, R.J. Morris, R.J. Najmanovich, R.A. Laskowski, J.M. Thornton, A method for localizing ligand binding pockets in protein structures., Proteins 62 (2006) 479-488.

[36] Z. Zsoldos, D. Reid, A. Simon, B.S. Sadjad, A.P. Johnson, eHiTS: an innovative approach to the docking and scoring function problems., Curr Protein Pept Sci 7 (2006) 421-435.

[37] R. Thomsen, M.H. Christensen, MolDock: a new technique for high-accuracy molecular docking., J Med Chem 49 (2006) 3315-3321.

[38] A.T. Laurie, R.M. Jackson, Methods for the prediction of protein-ligand binding sites for structure-based drug design and virtual ligand screening., Curr Protein Pept Sci 7 (2006) 395-406.

[39] J. An, M. Totrov, R. Abagyan, Comprehensive identification of "druggable" protein ligand binding sites., Genome Inform 15 (2004) 31-41.

[40] S.J. Campbell, N.D. Gold, R.M. Jackson, D.R. Westhead, Ligand binding: functional site location, similarity and docking., Curr Opin Struct Biol 13 (2003) 389-395.

[41] C. Sotriffer, G. Klebe, Identification and mapping of small-molecule binding sites in proteins: computational tools for structure-based drug design., Farmaco 57 (2002) 243-251.

[42] I.D. Kuntz, J.M. Blaney, S.J. Oatley, R. Langridge, T.E. Ferrin, A geometric approach to macromolecule-ligand interactions., J Mol Biol 161 (1982) 269-288.

[43] H.J. Böhm, The computer program LUDI: a new method for the de novo design of enzyme inhibitors., J Comput Aided Mol Des 6 (1992) 61-78.

[44] B. Kramer, M. Rarey, T. Lengauer, Evaluation of the FLEXX incremental construction algorithm for protein-ligand docking., Proteins 37 (1999) 228-241.

[45] D.T. Moustakas, P.T. Lang, S. Pegg, E. Pettersen, I.D. Kuntz, N. Brooijmans, R.C. Rizzo, Development and validation of a modular, extensible docking program: DOCK 5., J Comput Aided Mol Des 20 (2006) 601-619.

[46] W. Welch, J. Ruppert, A.N. Jain, Hammerhead: fast, fully automated docking of flexible ligands to protein binding sites., Chem Biol 3 (1996) 449-462.

[47] M.D. Miller, S.K. Kearsley, D.J. Underwood, R.P. Sheridan, FLOG: a system to select 'quasi-flexible' ligands complementary to a receptor of known three-dimensional structure., J Comput Aided Mol Des 8 (1994) 153-174.

[48] J.Y. Trosset, H.A. Scheraga, Reaching the global minimum in docking simulations: a Monte Carlo energy minimization approach using Bezier splines., Proc Natl Acad Sci USA 95 (1998) 8011-8015.

[49] D.S. Goodsell, G.M. Morris, A.J. Olson, Automated docking of flexible ligands: applications of AutoDock., J Mol Recognit 9 (1996) 1-5.

[50] J.Y. Trosset, H.A. Scheraga, PRODOCK: Software package for protein modeling and docking, J Comput Chem 20 (1999) 412-427.
[51] R. Abagyan, M. Totrov, D. Kuznetzov, ICM—a new method for protein modeling and design: applications to docking and structure prediction from the distorted native conformation., J Comput Chem 15 (1994) 488-506.
[52] M. Liu, S. Wang, MCDOCK: a Monte Carlo simulation approach to the molecular docking problem., J Comput Aided Mol Des 13 (1999) 435-451.
[53] T.N. Hart, R.J. Read, A multiple-start Monte Carlo docking method., Proteins 13 (1992) 206-222.
[54] C. McMartin, R.S. Bohacek, QXP: powerful, rapid computer algorithms for structure-based drug design., J Comput Aided Mol Des 11 (1997) 333-344.
[55] G. Jones, P. Willett, R.C. Glen, Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation., J Mol Biol 245 (1995) 43-53.
[56] G. Jones, P. Willett, R.C. Glen, A.R. Leach, R. Taylor, Development and validation of a genetic algorithm for flexible docking., J Mol Biol 267 (1997) 727-748.
[57] G.M. Morris, D.S. Goodsell, R.S. Halliday, R. Huey, W.E. Hart, R.K. Belew, A.J. Olson, Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function, J Comput Chem 19 (1998) 1639-1662.
[58] K.P. Clark, Flexible ligand docking without parameter adjustment across four ligand–receptor complexes., J Comput Chem 16 (1995) 1210-1226.
[59] J.S. Taylor, R.M. Burnett, DARWIN: a program for docking flexible molecules., Proteins 41 (2000) 173-191.
[60] C.A. Baxter, C.W. Murray, D.E. Clark, D.R. Westhead, M.D. Eldridge, Flexible docking using Tabu search and an empirical estimate of binding affinity., Proteins 33 (1998) 367-382.
[61] M.A. Ajay, J. Murcko, Computational methods to predict binding free energy in ligand-receptor complexes., J Med Chem 38 (1995) 4953-4967.
[62] J.D. Hirst, Predicting ligand binding energies, Curr Opin Drug Discovery Dev 1 (1998) 28-33.
[63] H.J. Böhm, M. Stahl, Rapid empirical scoring functions in virtual screening applications., Med Chem Res 9 (1999) 445-462.
[64] D.B. Kitchen, H. Decornez, J.R. Furr, J. Bajorath, Docking and scoring in virtual screening for drug discovery: Methods and applications, Nat Rev Drug Discov 3 (2004) 935-949.
[65] C. Bissantz, G. Folkers, D. Rognan, Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations., J Med Chem 43 (2000) 4759-4767.
[66] P. Ferrara, H. Gohlke, D.J. Price, G. Klebe, C.L.3rd. Brooks, Assessing scoring functions for protein-ligand interactions., J Med Chem 47 (2004) 3032-3047.
[67] I. Halperin, B. Ma, H. Wolfson, R. Nussinov, Principles of docking: An overview of search algorithms and a guide to scoring functions., Proteins 47 (2002) 409-443.
[68] N. Paul, D. Rognan, ConsDock: A new program for the consensus analysis of protein-ligand interactions., Proteins 47 (2002) 521-533.
[69] C. Pérez, A.R. Ortiz, Evaluation of docking functions for protein-ligand docking., J Med Chem 44 (2001) 3768-3785.
[70] M. Stahl, M. Rarey, Detailed analysis of scoring functions for virtual screening., J Med Chem 44 (2001) 1035-1042.
[71] G.E. Terp, B.N. Johansen, I.T. Christensen, F.S. Jørgensen, A new concept for multidimensional selection of ligand conformations (MultiSelect) and multidimensional scoring (MultiScore) of protein-ligand binding affinities., J Med Chem 44 (2001) 2333-2343.
[72] R. Wang, Y. Lu, S. Wang, Comparative evaluation of 11 scoring functions for molecular docking., J Med Chem 46 (2003) 2287-2303.
[73] R. Wang, Y. Lu, X. Fang, S. Wang, An extensive test of 14 scoring functions using the PDBbind refined set of 800 protein-ligand complexes., J Chem Inf Comput Sci 44 (2004) 2114-2125.
[74] D. Wilton, P. Willett, K. Lawson, G. Mullier, Comparison of ranking methods for virtual screening in lead-discovery programs., J Chem Inf Comput Sci 43 (2003) 469-474.
[75] M.A. Miteva, W.H. Lee, M.O. Montes, B.O. Villoutreix, Fast structure-based virtual ligand screening combining FRED, DOCK, and Surflex., J Med Chem 48 (2005) 6012-6022.
[76] B.Q. Wei, L.H. Weaver, A.M. Ferrari, B.W. Matthews, B.K. Shoichet, Testing a flexible-receptor docking algorithm in a model binding site., J Mol Biol 337 (2004) 1161-1182.
[77] M. Totrov, R. Abagyan, Flexible protein-ligand docking by global energy optimization in internal coordinates., Proteins Suppl 1 (1997) 215-220.
[78] U. Rester, Dock around the clock - Current status of small molecule docking and scoring, Qsar & Combinatorial Science 25 (2006) 605-615.

[79] S.F. Sousa, P.A. Fernandes, M.J. Ramos, Protein-ligand docking: Current status and future challenges, Proteins 65 (2006) 15-26.
[80] A.N. Jain, Scoring functions for protein-ligand docking., Curr Protein Pept Sci 7 (2006) 407-420.
[81] T. Schulz-Gasch, M. Stahl, Scoring functions for protein–ligand interactions: a critical perspective., Drug Discovery Today, Technol 1 (2004) 231-239.
[82] X. Barril, R. Soliva, Molecular modelling., Mol Biosyst 2 (2006) 660-681.
[83] J. Gasteiger, M. Marsili, Iterative partial equalization of orbital electronegativity – a rapid access to atomic charges., Tetrahedron 36 (1980) 3219-3228.
[84] M.L. Verdonk, J.C. Cole, M.J. Hartshorn, C.W. Murray, R.D. Taylor, Improved protein-ligand docking using GOLD., Proteins 52 (2003) 609-623.
[85] P.K. Weiner, P.A. Kollman, AMBER—assisted model building with energy refinement—a general program for modeling molecules and their interactions., J Comput Chem 2 (1981) 287-303.
[86] S.J. Weiner, P.A. Kollman, D.A. Case, U.C. Singh, C. Ghio, G. Alagona, S. Profeta, P. Weiner A new force-field for molecular mechanical simulation of nucleic-acids and proteins., J AmChem Soc 106 (1984) 765-784.
[87] S.J. Weiner, P.A. Kollman, D.T. Nguyen, D.A. Case, An all-atom force field for simulations of protein and nucleic acids., J Comput Chem 7 (1986) 230-252.
[88] T.N. Hart, S.R. Ness, R.J. Read, Critical evaluation of the research docking program for the CASP2 challenge., Proteins Suppl 1 (1997) 205-209.
[89] B.R. Brooks, R.E. Bruccoleri, B.D. Olafson, D.J. States, S. Swaminathan, M. Karplus, CHARMM—a programm for macromolecular energy, minimization, and dynamics calculations., J Comput Chem 4 (1983) 187-217.
[90] L. Nilsson, M. Karplus, Empirical energy functions for energy minimization and dynamics of nucleic acids., J Comput Chem 7 (1986) 591-616.
[91] W.F. van Gunsteren, H.C. Berendsen, Computer simulations of molecular dynamics: methodology, applications, and perpectives in chemistry., Angew Chem Int Ed 29 (1990) 992-1023.
[92] W.L. Jorgensen, J. Tirado-Rives, The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin., J Am Chem Soc 110 (1988) 1657-1666.
[93] H.J. Böhm, The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure., J Comput Aided Mol Des 8 (1994) 243-256.
[94] M.D. Eldridge, C.W. Murray, T.R. Auton, G.V. Paolini, R.P. Mee, Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes., J Comput Aided Mol Des 11 (1997) 425-445.
[95] D. Rognan, S.L. Lauemoller, A. Holm, S. Buus, V. Tschinke, Predicting binding affinities of protein ligands from three-dimensional models: application to peptide binding to class I major histocompatibility proteins., J Med Chem 42 (1999) 4650-4658.
[96] S. Huo, J. Wang, P. Cieplak, P.A. Kollman, I.D. Kuntz, Molecular dynamics and free energy analyses of cathepsin D-inhibitor interactions: insight into structure-based ligand design., J Med Chem 45 (2002) 1412-1419.
[97] D.F. Sitkoff, K.A. Sharp, B. Honig, Accurate calculation of hydration free energies using macroscopic continuum models., J. Phys. Chem. 98 (1998) 1978–1983.
[98] M. Rarey, B. Kramer, T. Lengauer, G. Klebe, A fast flexible docking method using an incremental construction algorithm., J Mol Biol 261 (1996) 470-489.
[99] H.J. Böhm, LUDI: rule-based automatic design of new substituents for enzyme inhibitor leads., J Comput Aided Mol Des 6 (1992) 593-606.
[100] H.J. Böhm, Prediction of binding constants of protein ligands: a fast method for the prioritization of hits obtained from de novo design or 3D database search programs., J Comput Aided Mol Des 12 (1998) 309-323.
[101] P. Tao, L. Lai, Protein ligand docking based on empirical method for binding affinity estimation., J Comput Aided Mol Des 15 (2001) 429-446.
[102] R.X. Wang, L. Liu, L.H. Lai, Y.Q. Tang, SCORE: A new empirical method for estimating the binding affinity of a protein-ligand complex, J Mol Model 4 (1998) 379-394.
[103] R.X. Wang, L.H. Lai, S.M. Wang, Further development and validation of empirical scoring functions for structure-based binding affinity prediction, J Comput Aided Mol Des 16 (2002) 11-26.

ISBN:978-84-691-1553-4 /DL: T.151-2008

[104] D.K. Gehlhaar, G.M. Verkhivker, P.A. Rejto, C.J. Sherman, D.B. Fogel, L.J. Fogel, S.T. Freer, Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: conformationally flexible docking by evolutionary programming., Chem Biol 2 (1995) 317-324.

[105] M.I. Zavodszky, P.C. Sanschagrin, R.S. Korde, L.A. Kuhn, Distilling the essential features of a protein surface for improving protein-ligand docking, scoring, and virtual screening., J Comput Aided Mol Des 16 (2002) 883-902.

[106] A.N. Jain, Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine., J Med Chem 46 (2003) 499-511.

[107] A.N. Jain, Surflex-Dock 2.1: Robust performance from ligand energetic modeling, ring flexibility, and knowledge-based search., J Comput Aided Mol Des 21 (2007) 281-306.

[108] T.A. Pham, A.N. Jain, Parameter estimation for scoring protein-ligand interactions using negative training data., J Med Chem 49 (2006) 5856-5868.

[109] R.A. Friesner, J.L. Banks, R.B. Murphy, T.A. Halgren, J.J. Klicic, D.T. Mainz, M.P. Repasky, E.H. Knoll, M. Shelley, J.K. Perry, D.E. Shaw, P. Francis, P.S. Shenkin, Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy., J Med Chem 47 (2004) 1739-1749.

[110] T.A. Halgren, R.B. Murphy, R.A. Friesner, H.S. Beard, L.L. Frye, W.T. Pollard, J.L. Banks, Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening., J Med Chem 47 (2004) 1750-1759.

[111] S. Miyazawa, R.L. Jernigan, Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation., Macromolecules 18 (1985) 534-552.

[112] H. Lu, J. Skolnick, Application of statistical potentials to protein structure refinement from low resolution ab initio models., Biopolymers 70 (2003) 575-584.

[113] H. Lu, L. Lu, J. Skolnick, Development of unified statistical potentials describing protein-protein interactions., Biophys J 84 (2003) 1895-1901.

[114] T. Kortemme, A.V. Morozov, D. Baker, An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes., J Mol Biol 326 (2003) 1239-1259.

[115] I. Muegge, PMF scoring revisited., J Med Chem 49 (2006) 5895-5902.

[116] C.Y. Yang, R. Wang, S. Wang, M-score: a knowledge-based potential scoring function accounting for protein atom mobility., J Med Chem 49 (2006) 5903-5911.

[117] M.J. Sippl, Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins., J Mol Biol 213 (1990) 859-883.

[118] S.F. Sousa, P.A. Fernandes, M.J. Ramos, Protein-ligand docking: current status and future challenges., Proteins 65 (2006) 15-26.

[119] H. Alonso, A.A. Bliznyuk, J.E. Gready, Combining docking and molecular dynamic simulations in drug design, Med Res Rev 26 (2006) 531-568.

[120] I. Muegge, Y.C. Martin, A general and fast scoring function for protein-ligand interactions: a simplified potential approach., J Med Chem 42 (1999) 791-804.

[121] I. Muegge, A knowledge-based scoring function for protein-ligand interactions: Probing the reference state, Perspect Drug Discov Des 20 (2000) 99-114.

[122] I. Muegge, Effect of ligand volume correction on PMF scoring, J Comput Chem 22 (2001) 418-425.

[123] J.B.O. Mitchell, R.A. Laskowski, A. Alex, J.M. Thornton, BLEEP – potential of mean force describing protein–ligand interactions: I. Generating potential., J Comput Chem 20 (1999) 1165-1176.

[124] J.B.O. Mitchell, R.A. Laskowski, A. Alex, M.J. Forster, J.M. Thornton, BLEEP – potential of mean force describing protein–ligand interactions: II. calculation of binding energies and comparison with experimental data., J Comput Chem 20 (1999) 1177-1185.

[125] H. Gohlke, M. Hendlich, G. Klebe, Knowledge-based scoring function to predict protein-ligand interactions., J Mol Biol 295 (2000) 337-356.

[126] H.F. Velec, H. Gohlke, G. Klebe, DrugScore(CSD)-knowledge-based scoring function derived from small molecule crystal data with superior recognition rate of near-native ligand poses and better affinity prediction., J Med Chem 48 (2005) 6296-6303.

[127] A.V. Ishchenko, E.I. Shakhnovich, SMall Molecule Growth 2001 (SMoG2001): an improved knowledge-based scoring function for protein-ligand interactions., J Med Chem 45 (2002) 2770-2780.

[128] R.S. DeWhitte, E.I. Shakhnovich, SMoG: de novo design method based on simple, fast, and accurate free energy estimates. 1. Methodology and supporting evidence., J Am Chem Soc 118 (1996) 11733-11744.

[129] A.E. Muryshev, D.N. Tarasov, A.V. Butygin, O.Y. Butygina, A.B. Aleksandrov, S.M. Nikitin, A novel scoring function for molecular docking., J Comput Aided Mol Des 17 (2003) 597-605.

[130] W. Deng, C. Breneman, M.J. Embrechts, Predicting protein-ligand binding affinities using novel geometrical descriptors and machine-learning methods., J Chem Inf Comput Sci 44 (2004) 699-703.

[131] H. Gohlke, G. Klebe, DrugScore meets CoMFA: adaptation of fields for molecular comparison (AFMoC) or how to tailor knowledge-based pair-potentials to a particular protein., J Med Chem 45 (2002) 4153-4170.

[132] S. Radestock, M. Böhm, H. Gohlke, Improving binding mode predictions by docking into protein-specifically adapted potential fields., J Med Chem 48 (2005) 5466-5479.

[133] P.S. Charifson, J.J. Corkery, M.A. Murcko, W.P. Walters, Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins., J Med Chem 42 (1999) 5100-5109.

[134] L. Birch, C.W. Murray, M.J. Hartshorn, I.J. Tickle, M.L. Verdonk, Sensitivity of molecular docking to induced fit effects in influenza virus neuraminidase., J Comput Aided Mol Des 16 (2002) 855-869.

[135] J.A. Erickson, M. Jalaie, D.H. Robertson, R.A. Lewis, M. Vieth, Lessons in molecular recognition: the effects of ligand and protein flexibility on molecular docking accuracy., J Med Chem 47 (2004) 45-55.

[136] C.W. Murray, C.A. Baxter, A.D. Frenkel, The sensitivity of the results of molecular docking to induced fit effects: application to thrombin, thermolysin and neuraminidase., J Comput Aided Mol Des 13 (1999) 547-562.

[137] G.M. Verkhivker, D. Bouzida, D.K. Gehlhaar, P.A. Rejto, S.T. Freer, P.W. Rose, Complexity and simplicity of ligand-macromolecule interactions: the energy landscape perspective., Curr Opin Struct Biol 12 (2002) 197-203.

[138] I. Luque, E. Freire, Structural stability of binding sites: consequences for binding affinity and allosteric effects., Proteins Suppl 4 (2000) 63-71.

[139] H. Alonso, A.A. Bliznyuk, J.E. Gready, Combining docking and molecular dynamic simulations in drug design., Med Res Rev 26 (2006) 531-568.

[140] S.J. Teague, Implications of protein flexibility for drug discovery., Nat Rev Drug Discov 2 (2003) 527-541.

[141] F. Jiang, S.H. Kim, "Soft docking": matching of molecular surface cubes., J Mol Biol 219 (1991) 79-102.

[142] B. Sandak, R. Nussinov, H.J. Wolfson, A method for biomolecular structural recognition and docking allowing conformational flexibility., J Comput Biol 5 (1998) 631-654.

[143] B. Sandak, H.J. Wolfson, R. Nussinov, Flexible docking allowing induced fit in proteins: insights from an open to closed conformational isomers., Proteins 32 (1998) 159-174.

[144] J.H. Lin, A.L. Perryman, J.R. Schames, J.A. McCammon, Computational drug design accommodating receptor flexibility: the relaxed complex scheme., J Am Chem Soc 124 (2002) 5632-5633.

[145] J.H. Lin, A.L. Perryman, J.R. Schames, J.A. McCammon, The relaxed complex method: Accommodating receptor flexibility for drug design with an improved scoring scheme., Biopolymers 68 (2003) 47-62.

[146] L. Schaffer, G.M. Verkhivker, Predicting structural effects in HIV-1 protease mutant complexes with flexible ligand docking and protein side-chain optimization., Proteins 33 (1998) 295-310.

[147] V. Schnecke, C.A. Swanson, E.D. Getzoff, J.A. Tainer, L.A. Kuhn, Screening a peptidyl database for potential ligands to proteins with side-chain flexibility., Proteins 33 (1998) 74-87.

[148] J.Y. Trosset, H.A. Scheraga, PRODOCK: Software package for protein modeling and docking, J Comput Chem 20 (1999) 412-427.

[149] T.M. Frimurer, G.H. Peters, L.F. Iversen, H.S. Andersen, N.P. Møller, O.H. Olsen, Ligand-induced conformational changes: improved predictions of ligand binding conformations and affinities., Biophys J 84 (2003) 2273-2281.

[150] P. Källblad, P.M. Dean, Efficient conformational sampling of local side-chain flexibility., J Mol Biol 326 (2003) 1651-1665.

[151] A.R. Leach, Ligand docking to proteins with discrete side-chain flexibility., J Mol Biol 235 (1994) 345-356.

ISBN:978-84-691-1553-4 /DL: T.151-2008

[152] F. Osterberg, G.M. Morris, M.F. Sanner, A.J. Olson, D.S. Goodsell, Automated docking to multiple target structures: incorporation of protein mobility and structural water heterogeneity in AutoDock., Proteins 46 (2002) 34-40.
[153] S.Y. Huang, X. Zou, Efficient molecular docking of NMR structures: application to HIV-1 protease., Protein Sci 16 (2007) 43-51.
[154] S.Y. Huang, X. Zou, Ensemble docking of multiple protein structures: considering protein structural variations in molecular docking., Proteins 66 (2007) 399-421.
[155] R.M. Knegtel, I.D. Kuntz, C.M. Oshiro, Molecular docking to ensembles of protein structures., J Mol Biol 266 (1997) 424-440.
[156] H. Claussen, C. Buning, M. Rarey, T. Lengauer, FlexE: efficient molecular docking considering protein structure variations., J Mol Biol 308 (2001) 377-395.
[157] H.A. Carlson, Protein flexibility is an important component of structure-based drug discovery., Curr Pharm Des 8 (2002) 1571-1578.
[158] M. Vaqué, A. Arola, C. Aliagas, G. Pujadas, BDT: an easy-to-use front-end application for automation of massive docking tasks and complex docking strategies with AutoDock., Bioinformatics 22 (2006) 1803-1804.
[159] Z. Zsoldos, D. Reid, A. Simon, S.B. Sadjad, A.P. Johnson, eHiTS: A new fast, exhaustive flexible ligand docking system., J Mol Graph Model 26 (2006) 198-212.
[160] R. Taylor, Life-science applications of the Cambridge Structural Database., Acta Crystallogr D Biol Crystallogr 58 (2002) 879-888.
[161] E. Kellenberger, J. Rodrigo, P. Muller, D. Rognan, Comparative evaluation of eight docking tools for docking and virtual screening accuracy., Proteins 57 (2004) 225-242.
[162] S.C. Lovell, J.M. Word, J.S. Richardson, D.C. Richardson, The penultimate rotamer library., Proteins 40 (2000) 389-408.
[163] D.S. Goodsell, A.J. Olson, Automated docking of substrates to proteins by simulated annealing., Proteins 8 (1990) 195-202.
[164] G.M. Morris, D.S. Goodsell, R. Huey, A.J. Olson, Distributed automated docking of flexible ligands to proteins: parallel applications of AutoDock 2.4., J Comput Aided Mol Des 10 (1996) 293-304.
[165] K. Sharp, R. Fine, B. Honig, Computer simulations of the diffusion of a substrate to an active site of an enzyme., Science 236 (1987) 1460-1463.
[166] M.I. Zavodszky, M. Lei, M.F. Thorpe, A.R. Day, L.A. Kuhn, Modeling correlated main-chain motions in proteins for flexible molecular recognition., Proteins 57 (2004) 243-261.
[167] Z. Bikádi, E. Hazai, F. Zsila, S.F. Lockwood, Molecular modeling of non-covalent binding of homochiral (3S,3'S)-astaxanthin to matrix metalloproteinase-13 (MMP-13). Bioorg Med Chem 14 (2006) 5451-5458.
[168] E. Hazai, Z. Bikádi, F. Zsila, S.F. Lockwood, Molecular modeling of the non-covalent binding of the dietary tomato carotenoids lycopene and lycophyll, and selected oxidative metabolites with 5-lipoxygenase., Bioorg Med Chem 14 (2006) 6859-6867.
[169] C. Hetényi, D. van der Spoel, Efficient docking of peptides to proteins without prior knowledge of the binding site., Protein Sci 11 (2002) 1729-1737.
[170] C. Hetényi, D. van der Spoel, Blind docking of drug-sized compounds to proteins with up to a thousand residues., FEBS Lett 580 (2006) 1447-1450.
[171] B. Iorga, D. Herlem, E. Barré, C. Guillou, Acetylcholine nicotinic receptors: finding the putative binding site of allosteric modulators using the "blind docking" approach., J Mol Model 12 (2006) 366-372.
[172] M. Kovács, J. Tóth, C. Hetényi, A. Málnási-Csizmadia, J.R. Sellers, Mechanism of blebbistatin inhibition of myosin II., J Biol Chem 279 (2004) 35557-35563.
[173] J.A.Jr. Yankeelov, D.E.Jr. Koshland, Evidence for conformation changes induced by substrates of phosphoglucomutase., J Biol Chem 240 (1965) 1593-1602.
[174] W. Sherman, T. Day, M.P. Jacobson, R.A. Friesner, R. Farid, Novel procedure for modeling ligand/receptor induced fit effects., J Med Chem 49 (2006) 534-553.
[175] R. Farid, T. Day, R.A. Friesner, R.A. Pearlstein, New insights about HERG blockade obtained from protein modeling, potential energy mapping, and docking studies., Bioorg Med Chem 14 (2006) 3160-3173.
[176] F. Forouhar, M. Hussain, R. Farid, J. Benach, M. Abashidze, W.C. Edstrom, S.M. Vorobiev, R. Xiao, T.B. Acton, Z. Fu, J.J. Kim, H.M. Miziorko, G.T. Montelione, J.F. Hunt, Crystal structures of two bacterial 3-hydroxy-3-methylglutaryl-CoA lyases suggest a common catalytic mechanism among a family of TIM barrel metalloenzymes cleaving carbon-carbon bonds., J Biol Chem 281 (2006) 7533-7545.

[177] T. Jovanovic, R. Farid, R.A. Friesner, A.E. McDermott, Thermal equilibrium of high- and low-spin forms of cytochrome P450 BM-3: repositioning of the substrate?, J Am Chem Soc 127 (2005) 13548-13552.
[178] K. Maeda, D. Das, H. Ogata-Aoki, H. Nakata, T. Miyakawa, Y. Tojo, R. Norman, Y. Takaoka, J. Ding, G.F. Arnold, E. Arnold, H. Mitsuya, Structural and molecular interactions of CCR5 inhibitors with CCR5., J Biol Chem 281 (2006) 12688-12698.
[179] W. Sherman, H.S. Beard, R. Farid, Use of an induced fit receptor structure in virtual screening., Chem Biol Drug Des 67 (2006) 83-84.
[180] M.D. Cummings, R.L. DesJarlais, A.C. Gibbs, V. Mohan, E.P. Jaeger, Comparison of automated docking programs as virtual screening tools., J Med Chem 48 (2005) 962-976.
[181] S. Joy, P.S. Nair, R. Hariharan, M.R. Pillai, Detailed comparison of the protein-ligand docking efficiencies of GOLD, a commercial package and ArgusLab, a licensable freeware., In Silico Biol 6 (2006) 601-605.
[182] M. Kontoyianni, L.M. McClellan, G.S. Sokol, Evaluation of docking performance: comparative data on docking algorithms., J Med Chem 47 (2004) 558-565.
[183] E. Perola, W.P. Walters, P.S. Charifson, A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance., Proteins 56 (2004) 235-249.
[184] G.L. Warren, C.W. Andrews, A.M. Capelli, B. Clarke, J. LaLonde, M.H. Lambert, M. Lindvall, N. Nevins, S.F. Semus, S. Senger, G. Tedesco, I.D. Wall, J.M. Woolven, C.E. Peishoff, M.S. Head, A critical assessment of docking programs and scoring functions., J Med Chem 49 (2006) 5912-5931.
[185] L. David, R. Luo, M.K. Gilson, Ligand-receptor docking with the Mining Minima optimizer., J Comput Aided Mol Des 15 (2001) 157-171.
[186] R.M. Jackson, Q-fit: a probabilistic method for docking molecular fragments by sampling low energy conformational space., J Comput Aided Mol Des 16 (2002) 43-57.
[187] Y.P. Pang, E. Perola, K. Xu, F.G. Prendergast, EUDOC: a computer program for identification of drug interaction sites in macromolecules and drug leads from chemical databases., J Comput Chem 22 (2001) 1750-1771.
[188] R.D. Taylor, P.J. Jewsbury, J.W. Essex, FDS: flexible ligand and receptor docking with a continuum solvent model and soft-core energy function., J Comput Chem 24 (2003) 1637-1656.
[189] C.M. Venkatachalam, X. Jiang, T. Oldfield, M. Waldman, LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites., J Mol Graph Model 21 (2003) 289-307.
[190] D.R. Westhead, D.E. Clark, C.W. Murray, A comparison of heuristic search algorithms for molecular docking., J Comput Aided Mol Des 11 (1997) 209-228.
[191] R.D. Clark, A. Strizhev, J.M. Leonard, J.F. Blake, J.B. Matthew, Consensus scoring for ligand/protein interactions., J Mol Graph Model 20 (2002) 281-295.
[192] H. Chen, P.D. Lyne, F. Giordanetto, T. Lovell, J. Li, On evaluating molecular-docking methods for pose prediction and enrichment factors., J Chem Inf Model 46 (2006) 401-415.
[193] T.N. Doman, S.L. McGovern, B.J. Witherbee, T.P. Kasten, R. Kurumbail, W.C. Stallings, D.T. Connolly, B.K. Shoichet, Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B., J Med Chem 45 (2002) 2213-2221.
[194] O.V. Buzko, A.C. Bishop, K.M. Shokat, Modified AutoDock for accurate docking of protein kinase inhibitors., J Comput Aided Mol Des 16 (2002) 113-127.
[195] H. Gohlke, G. Klebe, Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors., Angew Chem Int Ed Engl 41 (2002) 2644-2676.
[196] J.C. Cole, C.W. Murray, J.W.M. Nissink, R.D. Taylor, R. Taylor, Comparing protein-ligand docking programs is difficult, Proteins 60 (2005) 325-332.
[197] T. Schulz-Gasch, M. Stahl, Binding site characteristics in structure-based virtual screening: evaluation of current docking tools., J Mol Model 9 (2003) 47-57.
[198] M. Cecchini, P. Kolb, N. Majeux, A. Caflisch, Automated docking of highly flexible ligands by genetic algorithms: a critical assessment., J Comput Chem 25 (2004) 412-422.
[199] J.L. Jenkins, R.Y. Kao, R. Shapiro, Virtual screening to enrich hit lists from high-throughput screening: a case study on small-molecule inhibitors of angiogenin., Proteins 50 (2003) 81-93.
[200] B.D. Bursulaya, M. Totrov, R. Abagyan, C.L.3rd. Brooks, Comparative study of several algorithms for flexible ligand docking., J Comput Aided Mol Des 17 (2003) 755-763.
[201] J.W. Nissink, C. Murray, M. Hartshorn, M.L. Verdonk, J.C. Cole, R. Taylor, A new test set for validating predictions of protein-ligand interaction., Proteins 49 (2002) 457-471.

ISBN:978-84-691-1553-4 /DL: T.151-2008

[202] J. Goto, R. Kataoka, N. Hirayama, Ph4Dock: pharmacophore-based protein-ligand docking., J Med Chem 47 (2004) 6804-6811.

[203] R.T. Kroemer, A. Vulpetti, J.J. McDonald, D.C. Rohrer, J.Y. Trosset, F. Giordanetto, S. Cotesta, C. McMartin, M. Kihlén, P.F. Stouten, Assessment of docking poses: interactions-based accuracy classification (IBAC) versus crystal structure deviations., J Chem Inf Comput Sci 44 (2004) 871-881.

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

TABLES AND FIGURES

Table 1. Selected force field-based scoring functions

	Protein-ligand	Internal ligand
SYBYL™ G-Score	$E_{vdW} + E_{H-bond} =$ $\sum_{prot}\sum_{lig}\left[\left(\frac{A_{ij}}{d_{ij}^{8}} - \frac{B_{ij}}{d_{ij}^{4}}\right) + \left(E_{da} + E_{ww}\right) - \left(E_{dw} + E_{aw}\right)\right]$	$E_{vdW} + E_{torsion} =$ $\sum_{lig}\left(\frac{C_{ij}}{d_{ij}^{12}} - \frac{D_{ij}}{d_{ij}^{6}}\right) + \sum_{lig}\frac{1}{2}V\left[1 + \frac{n}{\lvert n\rvert}\cos\left(\lvert n\rvert\omega\right)\right]$
SYBYL™ D-Score	$E_{vdW} + E_{electrostatic} =$ $\sum_{prot}\sum_{lig}\left[\left(\frac{A_{ij}}{d_{ij}^{12}} + \frac{B_{ij}}{d_{ij}^{6}}\right) + 332.0\frac{q_i q_j}{\varepsilon\left(d_{ij}\right)d_{ij}}\right]$	
GoldScore	$E_{vdW} + E_{electrostatic} =$ $\sum_{prot}\sum_{lig}\left[\left(\frac{A_{ij}}{d_{ij}^{a}} + \frac{B_{ij}}{d_{ij}^{b}}\right) + 332.0\frac{q_i q_j}{\varepsilon\left(d_{ij}\right)d_{ij}}\right]$	$E_{vdW} + E_{electrostatic} =$ $\sum_{lig}\left[\left(\frac{A_{ij}}{d_{ij}^{a}} + \frac{B_{ij}}{d_{ij}^{b}}\right) + 332.0\frac{q_i q_j}{\varepsilon\left(d_{ij}\right)d_{ij}}\right]$ $+ optional\ E_{H-bond}$
AutoDock v3.05	$E_{vdW} + E_{H-bond} + E_{electrostatic} =$ $\sum_{prot}\sum_{lig}\left[\left(\frac{A_{ij}}{d_{ij}^{12}} - \frac{B_{ij}}{d_{ij}^{6}}\right) + E(t)\times\left(\frac{C_{ij}}{d_{ij}^{12}} - \frac{D_{ij}}{d_{ij}^{10}}\right) + 332.0\frac{q_i q_j}{\varepsilon\left(d_{ij}\right)d_{ij}}\right]$ where *E(t)*=angular weight factor	$E_{vdW} + E_{H-bond} + E_{electrostatic} =$ $\sum_{lig}\left[\left(\frac{A_{ij}}{d_{ij}^{12}} - \frac{B_{ij}}{d_{ij}^{6}}\right) + E(t)\times\left(\frac{C_{ij}}{d_{ij}^{12}} - \frac{D_{ij}}{d_{ij}^{10}}\right) + 332.0\frac{q_i q_j}{4\left(d_{ij}\right)d_{ij}}\right]$ where *E(t)*=angular weight factor

The total ΔG_{bind} is given by the sum of the terms corresponding to the interaction energy between the receptor and the ligand and the internal energy of the ligand (when available). Depending on the scoring function, these terms can be obtained by adding the following contributions: (a) electrostatic (*i.e.* $E_{electrostatic}$); (b) van der Waals (*i.e.* E_{vdW}); (c) hydrogen bonding (*i.e.* $E_{H\text{-}bond}$); and (d) torsional (*i.e.* $E_{torsion}$). For two atoms **i** and **j**: (a) $\mathbf{A_{ij}}$ and $\mathbf{B_{ij}}$ correspond to van der Waals parameters for the given atom types; (b) $\mathbf{d_{ij}}$ corresponds to their interatomic distance; (c) $\mathbf{q_i}$ and $\mathbf{q_j}$ are the atomic partial charges; and (d) $\boldsymbol{\varepsilon(d_{ij})}$ is a distance-dependent dielectric function. In G-Score, the hydrogen bonding term is a sum of the individual energies (*i.e.* E_{da}, E_{ww}, E_{dw}, E_{aw}) from all the donor-acceptor pairs in the complex.

ISBN:978-84-691-1553-4 /DL: T.151-2008

Table 2. Selected empirical scoring functions

	Functional form
LUDI	$\Delta G_{bind} = \Delta G_{H-bond} \sum_{H-bond} f(\Delta R, \Delta\alpha) + \Delta G_{ionic} \sum_{ionic} f(\Delta R, \Delta\alpha) + \Delta G_{hydrophobic} \sum_{hydrophobic} \lvert A_{hydrophobic} \rvert + \Delta G_{rotor} N_{rotor} + \Delta G_0$
SYBYL™ F-Score	$\Delta G_{bind} = \Delta G_{H-bond} \sum_{H-bond} f(\Delta R, \Delta\alpha) + \Delta G_{ionic} \sum_{ionic} f(\Delta R, \Delta\alpha) + \Delta G_{aromatic} \sum_{aromatic} f(\Delta R, \Delta\alpha) + \Delta G_{contact} \sum_{contact} f(\Delta R) + \Delta G_{rotor} N_{rotor} + \Delta G_0$
SYBYL™ ChemScore	$\Delta G_{bind} = \Delta G_{H-bond} \sum_{H-bond} f(\Delta R, \Delta\alpha) + \Delta G_{metal} \sum_{metal} f(\Delta R, \Delta\alpha) + \Delta G_{lipo} \sum_{lipo} f(\Delta R) + \Delta G_{rotor} \sum_{rotor} f(P_{nl}, P'_{nl}) + \Delta G_0$

The free energy of binding (*i.e.* ΔG_{bind}) is obtained by adding the contribution to the free energy of some terms that correspond to: (a) hydrogen bonding (in LUDI and SYBYL™/F-Score the first two terms account for neutral and ionic hydrogen bonds, respectively); (b) hydrophobic or lipophilic (that accounts for the hydrophobic effect); (c) ligand rotational entropy (a term that counts all the rotatable single bonds in the ligand, which is supposed to be related with the torsional entropy loss of the ligand upon protein-ligand complexation.); (d) contact (that accounts for a general distance-dependent potential for protein-ligand atom contacts); and (e) metal (that accounts for metal ions residing inside the protein binding pocket). Thus, the different scoring functions differ in: (a) the number and the typology of the terms that contributes to ΔG_{bind}; and (b) the mathematical function (*i.e. f*) used to calculate one specific contribution [where this function can depend on an angular ($\Delta\alpha$) and/or a distance (ΔR) parameter/s that penalize/s the deviations from an ideal geometry]. $\Delta G_{H\text{-}bond}$, ΔG_{ionic}, $\Delta G_{hydrophobic}$, ΔG_{rotor}, $\Delta G_{aromatic}$, $\Delta G_{contact}$, ΔG_{metal}, ΔG_{lipo} are regression coefficients for each corresponding free energy term. ΔG_0 is a regression constant. $A_{hydrophobic}$ corresponds to the molecular surface area.

Table 3. Selected knowledge-based scoring functions

	Functional form
PMF	Parametrized pairwise potential PMF score: $PMF = \sum_{prot}\sum_{lig} A_{ij}(d_{ij}) A_{ij}(d_{ij}) = -k_B T \ln\left[f^{j}_{Vol\_corr}(r) \frac{\rho^{ij}_{seg}(r)}{\rho^{ij}_{bulk}} \right]$ where: (a) k_B is the Boltzmann constant; (b) $f^{j}_{Vol\_corr}(r)$ is a ligand volume correction factor; (c) and $\frac{\rho^{ij}_{seg}(r)}{\rho^{ij}_{bulk}}$ indicates a radial distribution function for a protein atom ***i*** and a ligan atom ***j***
DrugScore v1.2	$\Delta W = \gamma \sum_{prot}\sum_{lig} \Delta W_{ij}(r) + (1-\gamma)\left[\sum_{lig} \Delta W_i(SAS, SAS_0) + \sum_{prot} \Delta W_j(SAS, SAS_0) \right]$ where: (a) SAS correspond to the surface accessible area terms; (b) W_{ij} is a distance dependent pairwise potential; and (c) γ is an adjustable weight factor, normally set to 0.5
SMoG	$G = \sum_{ij} -kT \log\left[\frac{p_{ij}}{\overline{p}} \right] \Delta_{ij}$ where: (a) Δ_{ij} is 1 if the distance between atoms i and j is within 5.0 Å (and 0 if it is higher than 5.0 Å); and (b) p_{ij} and $\overline{p}$ are inter atomic and averaged inter atomic interactions

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

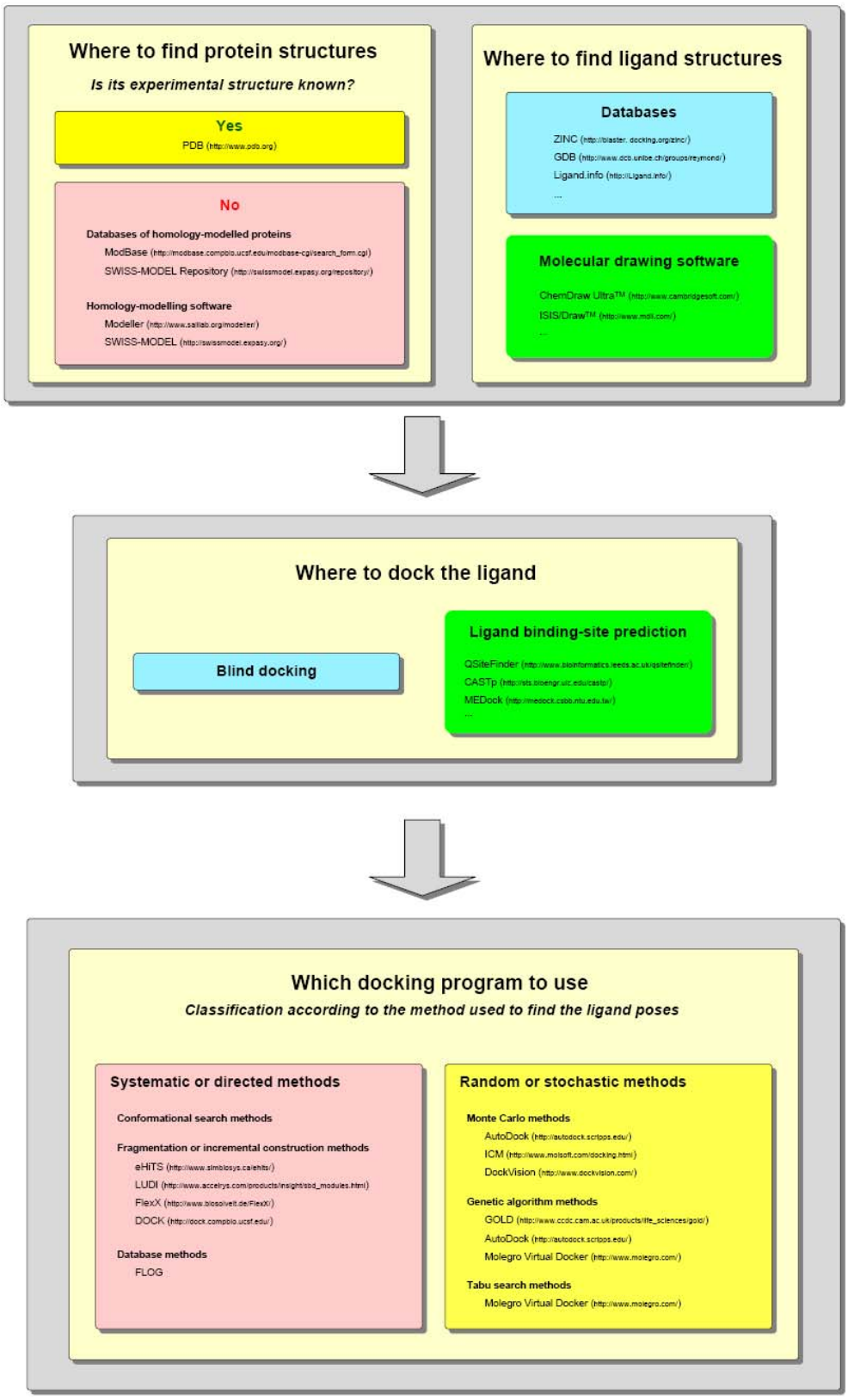
Table 4. Main characteristics of some selected protein-ligand docking programs

Program	Current version	Needs ligand setup?	Needs protein setup?	Input formats accepted	Method for conformational search of the ligand	Scoring functions available by default (and type of scoring function)	Can deal with receptor flexibility?	Allows to dock several ligands with a single set up?	Available OS versions	Graphic Interface	Free for academic research?
eHiTS®	v6.2	No	No	mol sd/sdf pdb mol2 tma tmb	Systematic (F)	eHiTS_Score (hybrid KB-EM) User-trained score	No	Yes	PC-Linux SUN-Solaris SGI-Irix IBM-AIX	CheVi™	Yes
GOLD™	v3.2	Yes	Yes	sd (lig) mol (lig) mol2 (lig/prot) PDB (lig/prot)	Stochastic (GA)	GoldScore (FF) ChemScore (EM) User-defined score	Up to 10 user-defined residues (side and main chain). Only with GoldScore	Yes	Windows 2000/XP PC-Linux SGI-Irix	Gold Front End Silver™ GoldMine™	No
MVD™	v2.2.5	No	No	pdb mol2 mol sd/sdf mdl mvdml	Stochastic (TS) Stochastic (GA)	MolDock Score (EM) MolDock Score [Grid] (EM)	Yes (only with the MolDock Score [GRID] scoring function)	Yes	Windows 2000/XP PC-Linux Mac OS X (Intel) Mac OS X (PowerPC)	Yes	No
AutoDock	v4.0.1	Yes	Yes	pdbqt pdbq mol2 pdb	Stochastic (GA) Stochastic (MC)	AMBER-derived (FF)	Yes	No	PC-Linux Mac OS X (Intel) Mac OS X (PowerPC) Windows (Cywin) SGI-Irix SUN-Solaris (Sparc)	AutoDockTools BDT	Yes
Glide™	v4.5	Yes	Yes	sd pdb maestro (mae) mol2	Systematic pose generation with stochastic optimization	SP 4.5 GlideScore (EM) XP 4.5 GlideScore (EM)	Yes (through Prime™)	Yes	Windows PC-Linux SGI-Irix IBM-AIX	Maestro™	No

Abbreviations for the conformational search method: (a) F (**F**ragmentation method); (b) GA (**G**enetic **A**lgorithm); (c) TS (**T**abu **S**earch); and (d) MC (**M**onte **C**arlo). Abbreviations for the type of scoring function: (a) KB (**K**nowledge-**B**ased); (b) EM (**Em**pirical); and (c) FF (**F**orce **F**ield).

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

Figure 1. Flow-chart of the main steps in a docking experiment



UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

BDT: an easy-to-use front-end Application for automation of massive docking tasks and complex docking strategies with AutoDock

Montserrat Vaque´[1], Anna Arola[1], Carles Aliagas[2] and Gerard Pujadas[1]

[1]Departament de Bioquímica i Biotecnologia, C/ Marcel·lí Domingo s/n and [2]Departament d'Enginyeria Informàtica i Matemàtiques, Av. Països Catalans, 26 Campus de Sant Pere Sescelades, Universitat Rovira i Virgili, Tarragona 43007, Catalonia, Spain

Bioinformatics (2006) 22 (14):1803-4

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

INTRODUCTION

When the specific interaction of small molecules with selected targets is central for research, *in silico* approaches such as protein-ligand docking become important tools. **AutoGrid**/**AutoDock** [1] is one of the most popular software packages for docking. In this package, **AutoGrid** is used to calculate the non-covalent energy of interaction between the receptor and a probe atom that is located in the different grid points of a lattice that divides the receptor's area of interest (*i.e.* the area of the macromolecule where the possibility of ligand binding is studied). **AutoGrid** builds as many files as the number of probe atoms used and these probes are: **(a)** the atoms that will be present in the ligands that will be docked onto the receptor (thus generating the corresponding affinity grid maps); and **(b)** a point charge of +1 (an alternative method can also generate this electrostatic potential grid map using a Poisson-Boltzmann finite difference method as in DELPHI; [2]). The next program in the package (*i.e.* **AutoDock**) uses the full set of grid maps built by **AutoGrid** to guide the docking process of the select ligands through the lattice volume. **AutoGrid**/**AutoDock** also has a graphic interface called **AutoDockTools** (**ADT**) that provides the user with powerful tools for analyzing the docking results and allows an easy set up of: **(a)** the macromolecule and ligand coordinates in the format required by **AutoGrid**/**AutoDock**; **(b)** the file needed to run **AutoGrid** around the receptor's area of interest (*i.e.* the so-called gpf file); and **(c)** the file needed for docking one specific ligand onto this area of the receptor (*i.e.* the so-called dpf file). While **ADT** has strongly decreased the learning-curve needed for using **AutoGrid**/**AutoDock**, it is also true that some docking tasks with this package, although possible, are far from trivial for users without strong computer skills. Examples of such tasks are: **(a)** using receptor flexibility during docking; **(b)** the automatic docking of a large library of ligands onto one or more receptors; and **(c)** docking a ligand library onto one or more receptors without defining one *a priori* ligand-binding site on them (the so-called "blind-docking" analysis) and using a short distance between grid points. To overcome these difficulties, we have developed BDT [3], a graphic front-end application that runs on top of four Fortran programs (*i.e.* make_grids, combine_grids, make_docks and analyze, one under each BDT window tab), which control the conditions of **AutoGrid** and **AutoDock** runs.

BDT is available for free, upon request, for non-commercial research.

ISBN:978-84-691-1553-4 /DL: T.151-2008

GENERAL DESCRIPTION

The BDT window has four *execution* tabs: **(a)** MakeGrids; **(b)** CombineGrids; **(c)** MakeDocks; and **(d)** Analyze (see Figure 1). The tabs are sorted from left to right according to their sequence of use when executing jobs from BDT. This is because the program that runs under each tab needs some input files that result from executing the program under the previous tab. Users are therefore recommend to move to the next tap until BDT informs them by e-mail that the execution of the program under one specific tap is finished.

Four buttons are common to all tabs and have the same function: **(a)** "execute", which starts running the Fortran program under the tab; **(b)** "reset", which replaces custom parameter values with the default ones, and **(c)** "save", which stores the current tab parameter values in a file for later recovery with the "load" button. Each field is filled with default parameter values each time BDT starts.

Users also have to set in some tabs how the underlying program will communicate with them by e-mail by: **(a)** typing their e-mail address in the corresponding window; and **(b)** indicating the amount of information they want to receive from the program.

One more tab (*i.e.* Information) is also available for informative tasks. This provides the user with general information about the main goals of BDT and its compatibility with **AutoGrid**/**AutoDock** versions.

The "?" button allows the user to activate/de-activate the contextual help messages (where default corresponds to contextual help messages activated). The "exit" button allows the user to exit BDT properly (while the Fortran program runs in the background).
Caution messages are also displayed when completing certain important fields.

The MakeGrids tab

Tab description

The MakeGrids tab (see Figure 2) is central to making advanced docking strategies because, from this tab, users can: **(a)** include or exclude the flexibility of the receptors in the calculations; and **(b)** search for the ligand-binding site in all the receptor surface ("Area around the whole receptor surface " option) or in a user-defined portion of it ("Area around one specific point" option) using, in both cases, a grid-point distance as short as they like regardless of the dimensions of the three-dimensional space that is searched. Users can then combine the

flexibility and ligand-binding site location strategies according to their needs (*e.g.* they can do the docking in a specific part of a flexible receptor). For each protein, this tab can automatically deal either with a single PDB file or with a set of PDB files corresponding to different snapshots of its conformations [such as those that can be readily obtained from **FlexWeb** tools [4] (http://flexweb.asu.edu/) or retrieved from the **MODEL** database (http://mmb.pcb.ub.es/MODEL/)]. Here the user does not need to do anything special to take into account receptor flexibility. Receptor flexibility is automatically assumed for those receptors provided to this tab by the user's list that have more than one PDBQS file (where PDBQS corresponds to the input format for the receptors in **AutoGrid**) in the selected PDBQS directory [it is assumed that PDBQS filenames that start with the same four-character code (usually a PDB code) correspond to different conformations of the same receptor].

This tab is used to control where **AutoGrid** has to run (*i.e.* around one specific point or around all the receptor), define the number of grid points per dimension (*i.e.* the dimensions of what we call a "partial box"; see Figure 3), establish the minimum separation between two grid points in the lattice, and set the "security distance" (though this is only used when the "Area around the whole receptor surface" option is activated). The "security distance" is used to increase the length, in the three dimensions, of the smallest box that is able to contain all the receptor's atoms (thus ensuring that the ligand will have enough room to "walk" around all the receptor surface and find the preferred sites for binding; see the algorithm description in the Figure 4A).

If the user selects the "Area around one specific point" option, a new window appears with the list of receptor's PDB codes and for each one the user must indicate: **(a)** the coordinates of the central point of the receptor's area of interest, and **(b)** the number of partial boxes that will be used in each dimension.

The rest of the running conditions are read from a gpf file that is used as a model to build all the gpf files that are needed for the **AutoGrid** runs that follow after the "execute" button is pressed. We provide BDT users with a couple of gpf files that can be used as models [one for proteins without cofactors (*i.e.* the model_gpf_basic file that can be used as a default selection) and one for proteins with Fe as cofactor (*i.e.* the model_gpf_with Fe file)], but users can easily build new ones with **AutoDockTools** (ADT) according to their specific running and receptor needs. This couple of gpf file templates are in the library directory of the original BDT distribution.

If "Quantity of information sent" is set to 5, then *make_grids* (the Fortran program under this tab) can e-mail a PDB file to the users every time *make_grids* stops working with one receptor or with one conformation of the current receptor (if receptor flexibility is used). This file is not

intended for analyzing results but can be used as a record for future reference because it contains: **(a)** the coordinates of the self receptor; **(b)** a box showing the location of the receptor's area of interest; and **(c)** the different "partial boxes" inside this area. This PDB file can be easily handled (*e.g.* coloured, hidden, etc.) by molecular graphic programs such as **RasMol** [5] because coordinates from the receptor, from its area of interest or from the partial boxes contained in this area are labelled as belonging to different subunits ("**R**" for the **r**eceptor, "**T**" for the area of in**t**erest and "**P**" for the **p**artial boxes).

Algorithm description for make_grid

The algorithm consists of the following steps (see Figure 4A):

1) *Make_grids* takes the input information provided by the MakeGrids tab.

2) *Make_grids* takes the first receptor in the list and calculates the dimensions of the smallest box (the so-called "minimum box"; see green box in the Figure3A) that is able to contain all the atoms of its different conformations [where the different conformations for the same receptor are automatically recognized by *make_grids* because their filenames share the first four characters (that usually corresponds to a PDB code)].

3) *Make_grids* adds to each "minimum box" dimension the value that corresponds to the "security distance" and obtains the so-called "total box" (see the blue box in the Figure3A).

4) If the "total box" does not contain an integer number of "partial boxes" (*i.e.* the area that will be used by each **AutoGrid**'s run), the corresponding dimensions of the "total box" are conveniently increased in size until it does.

5) *Make_grids* divides the "total box" into "partial boxes" (see the yellow boxes in the Figure3B) and runs an **AutoGrid** process inside each one. Now, the "total box" corresponds to the receptor's area of interest.

6) **AutoGrid** output files for each "partial box" are conveniently labelled and stored for further use by *combine_grids* (if more than one conformation is available for the receptor) or directly by *make_docks.*

7) After running **AutoGrid** inside all the "partial boxes" of a specific receptor, *make_grids* repeats either step 6 with the next receptor conformation or from step 2 to 6 with the next receptor in the list.

The algorithm (see Figure 4A) follows one slightly different pathway if the "Area around one specific point" option is used because, in that case, no "minimum" and "total" boxes are calculated by *make_grids* and the number of "partial boxes" used by the program correspond to those that have been indicated by the user. The Figure 5 shows the result of running *make_grids* around a selected point (indicated by the red ball).

The CombineGrids tab

Tab description

The CombineGrids tab (see Figure 6) is used to incorporate the receptor's mobility in docking calculations based on the work of Österberg et al. (2002) [6]. Briefly, this method combines all the grid maps from the different receptor conformations and the same probe to obtain a single grid-map file. In this file, the energy of each point is obtained from a weighted average of the energies of the same point in all the original conformational-dependent grid maps (where the corresponding weight is calculated using either a clamped grid or a Boltzmann assumption based on the interaction energy). The Fortran program under this tab is *combine_grids* and the resulting grid maps can be readily used by the MakeDocks tab.

The MakeDocks tab

Tab description

The MakeDocks tab (see Figure 7) enables easy selection of the receptors, ligands and conditions used by **AutoDock** during docking. Depending on the origin of the grid maps used (obtained from either the CombineGrids tab or the MakeGrids tab), the user will or will not consider receptor flexibility during docking. At this point, note that the values the user has to write in the "Dimensions of the grids that **AutoGrid** previously built for the current receptors" and the "Minimum separation between two grid points (Å)" fields of this tab: **(1)** are common for all the receptors included in the list selected in the "Which receptors do you want to use?" field; and **(2)** must be the same as those used by **AutoGrid** when the grid maps for these receptors were built.

AutoDock's running conditions are read from a dpf file that is used as a model to build all the dpf files needed for the **AutoDock** runs that follow after pressing the "execute" button. We provide BDT users with one dpf file (*i.e.* model_GALS.dpf) for using the best **AutoDock** method (the Genetic Algorithm with Local Search; *i.e.* GALS) but they can easily edit it or build a new one with **ADT** to suit their specific needs or preferences.

ISBN:978-84-691-1553-4 /DL: T.151-2008

The MakeDocks user also has to complete the "Dimensions of the grids that will use **AutoDock**" (which can be different from the dimensions of the "Dimensions of the grids that **AutoGrid** previously built for the current receptors" and are usually as large as possible in order to speed up *make_docks*, which is the program running under the present tab) and the "Security length (Å)" fields. Both parameters are only necessary when the receptor's area of interest is formed by more than one "partial box". To understand the meaning of these parameters, it is useful to know how *make_docks* controls where **AutoDock** runs (see the algorithm description in the Figure 4B).

If "Quantity of information sent" is set up to 5, then *make_docks* can e-mail to the users a PDB file every time *make_docks* stops working with one receptor. This file is not for analyzing results but it can be used as a record for future reference because it contains: **(a)** the coordinates of the self receptor; **(b)** a box showing the location of the receptor's area of interest; and **(c)** the different boxes inside the area where **AutoDock** has run (see Figure 8). This PDB file can be easily handled (*e.g.* coloured, hidden, etc.) by molecular graphic programs such as **RasMol** [5] because coordinates from the receptor, from its area of interest or from the **AutoDock** boxes contained in this area are labelled as belonging to different subunits ("**R**" for the **r**eceptor, "**T**" for the area of in**t**erest and "**P**" for the **AutoDock** boxes).

Algorithm description for make_docks

The algorithm consists of the following steps (see Figure 4B):

1) *Make_docks* takes the input information provided by the MakeDocks tab.

2) *Make_docks* reads the coordinates of all the ligands that will be used during the docking and calculates the longest distance between any two atoms from these ligands.

3) *Make_docks* obtains what we call the "maximum ligand length" by adding the "security length" value to the previously calculated distance. The reason for obtaining this "maximum ligand length" is that the coordinates of the ligand in the .OUT.PDBQ files (*i.e.* the ligands format in **AutoDock**) may not correspond to its most extended conformation (though it is strongly recommended that they do).

4) The number of points in the X-, Y- and Z-dimensions of **AutoDock**'s maps (*i.e.* the contents of the "Dimensions of the grids that **AutoGrid** previously built for the current receptors" fields) and the minimum distance between two points in them (*i.e.* the contents of the "Minimum separation between two grid points (Å)" field) are used by *make_docks* to calculate the length of the shortest edge of the parallelepiped (which is not necessarily a cube) where each execution of **AutoDock** will run. From the

difference between this value and the "maximum ligand length", *make_docks* obtains what we call the "stepsize".

5) *Make_docks* selects the first receptor in the list and uses its pre-calculated grid maps to build all those that are needed for the first run of **AutoDock**.

6) After these maps are built, **AutoDock** tries to dock in them all the ligands from the list (one after the other). The results for this first run of **AutoDock** are conveniently labelled and stored for further use by the Analyze tab.

7) *Make_docks* will repeat this process but in a parallelepiped that has been "moved" relative to the previous one by a distance that corresponds to the previously calculated "stepsize". Therefore, two adjacent **AutoDock** parallelepipeds always share enough volume to include (even in their more extended conformations) all the studied ligands. This ensures that ligand-binding sites in the interface between adjacent parallelepipeds are not lost during the blind docking. The Figure 8 shows the sequencial process of construction of boxes where AutoDock runs.

Using this method, *make_docks* analyzes all the area of interest for the current receptor. It is important to note, therefore, that the larger the dimensions of **AutoDock**'s grids, the larger the "stepsize" and, therefore, the fewer iterations needed to cover all the receptor's area of interest (*i.e. make_docks* runs faster).

The Analyze tab

Tab description

The Analyze tab (see Figure 9) can be used to analyze the docking results (*e.g.* by comparing the docking of different ligands onto the same receptor, etc.). From this tab, the user can select or set up: **(1)** the ligand-receptor pairs whose docking results will be analyzed; **(2)** the docking energy threshold below which a docking result for a ligand is rejected; **(3)** the root-mean-square deviation threshold used to cluster docking results for the same ligand (where the one with the highest receptor affinity is chosen as the cluster representative); **(4)** the width of the interval used in the **RasMol** scripts to colour cluster-representatives according to their receptor affinity; and **(5)** how many cluster representatives will be considered for each ligand. Once all this information is set up, the Fortran program under this tab (*i.e.* analyze) can be started.

Algorithm description for analyze

The algorithm consists of the following steps (see Figure 4C):

ISBN:978-84-691-1553-4 /DL: T.151-2008

1) *Analyze* takes the input information provided by the Analyze tab.

2) *Analyze* takes the first receptor from the list and generates a PDB file with its coordinates and with those from the box that corresponds to its studied area of interest.

3) *Analyze* takes all the output files corresponding to the docking of the first ligand into the current receptor and extracts the coordinates for all the predicted conformations of the ligand in its complex with this receptor.

4) All the ligand conformations whose docking energy is more positive than the corresponding user-defined threshold are rejected and the remaining ligand conformations are then clustered. Each cluster is therefore represented by the coordinates that correspond to the one with the highest affinity for the receptor.

5) The various cluster representatives are sorted according to their receptor affinity (*i.e.* the coordinates for the one with the highest affinity are labelled with the number one, and so on) and their coordinates are written into the PDB file that already contains the coordinates from the receptor and box showing its area of interest.

6) Steps from 3 to 5 are carried out with the output files that correspond to the docking of the other selected ligands onto the current receptor.

7) Two **RasMol** scripts are written to make it easier to compare the docking results of the different ligands onto the same receptor. The first one (*i.e.* script_01) colours the ligands according to different intervals of affinity for the receptor irrespective of their identity (*i.e.* from hot to cold colours for decreasing affinity) and is therefore useful for identifying where the most important ligand-binding sites in the receptor structure are located (see Figure 10A). The second one (*i.e.* script_02) colours the ligand according to its identity and is therefore useful for comparing the specificity of the different ligands for each binding site (see Figure 10B).

8) Once all the docking results for one specific receptor have been analyzed, the corresponding PDB file and Rasmol scripts are sent to the user e-mail address for further analysis with **Rasmol [5]**. All this process is repeated with every receptor in the list.

Description of the output PDB file

The output PDB files from analyze have the following parts:

1) Receptor coordinates This is the first part of the output PDB files. It contains the coordinates for all the atoms in the pdbqs file (usually atoms from protein or nucleic acids but sometimes also from cofactor, water molecules, etc). The coordinates are identified with the

label "**R**" in column 22 of the PDB file and can be easily selected with **RasMol [5]** by using the command "**select *R**".

2) Total box coordinates. This is the second part of the output PDB files. It contains the coordinates of the box where the docking of the different ligands onto the receptor has been analyzed (*i.e.* the receptor's area of interest). The coordinates are identified with the label "**T**" in column 22 of the PDB file and can be easily selected with **RasMol [5]** by using the command "**select *T**". The total box coordinates are connected by CONNECT records.

3) Ligand coordinates. This is the third part of the output PDB files. It contains the coordinates of the cluster-representatives obtained from the docking of the various ligands onto the receptor. The coordinates are grouped according to the following criteria:

All conformations belonging to the same ligand are grouped and the ligands are identified in columns 18-20 with their corresponding labels. They can be easily selected with **RasMol [5]** by using the command "**select [LAB]**" (where **LAB** must be replaced by the three-character label of the corresponding ligand).

All the coordinates for one specific conformation of a ligand are labelled with the same ligand label in columns 18-20 and with the same conformation number in columns 23-26. **VERY IMPORTANT:** The docking energy onto the receptor is found in columns 69-76 (whereas the integer part of this energy is located in the same columns in which the "B-factor" values are found in standard PDB files). Therefore, using the **RasMol** command "**select temperature < -11**", it is possible to select all cluster-representatives with energy more negative than -11.00 Kcal/mol. Other comparison operators that can be used with the "**select temperature**" command are >, <= and >=.

The various conformations for one specific ligand are sorted according to decreasing affinity for the receptor. Thus, the conformation with highest receptor affinity is labelled "**1**" in columns 23-26, the conformation with the second highest receptor affinity is labelled "**2**", and so on.

To use this two **RasMol** scripts, please put them in the same folder where **RasMol [5]** looks for the PDB files and load them with the command "**script NAME**" [where **NAME** must be replaced by the complete name of the script (including the file extension, if any)]

The license for BDT has been distributed to 87 universities or research centers all over the world. In the annex there is a table that summarizes the information about the centers that have requested the license.

ISBN:978-84-691-1553-4 /DL: T.151-2008

REFERENCES

[1] G.M. Morris, D.S. Goodsell, R.S. Halliday, R. Huey, W.E. Hart, R.K. Belew, A.J. Olson, Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function, J. Comput. Chem. 19 (1998) 1639-1662.

[2] K. Sharp, R. Fine, B. Honig, Computer simulations of the diffusion of a substrate to an active site of an enzyme., Science 236 (1987) 1460-1463.

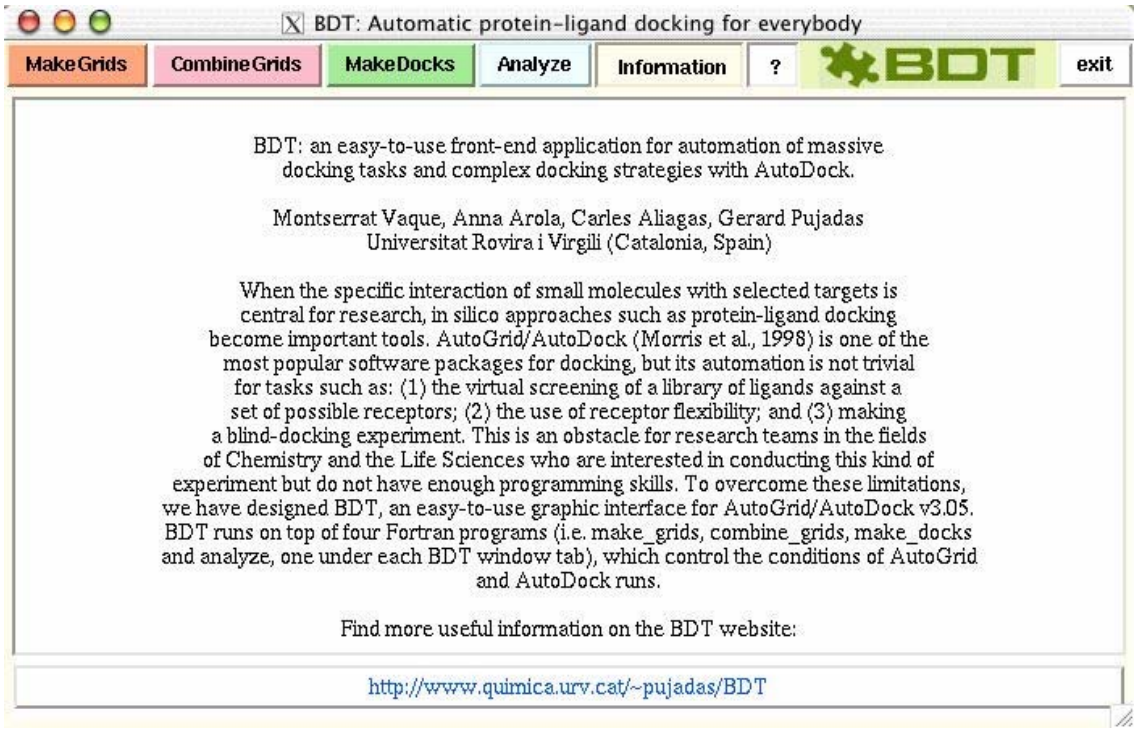
[3] M. Vaque, A. Arola, C. Aliagas, G. Pujadas, BDT: an easy-to-use front-end application for automation of massive docking tasks and complex docking strategies with AutoDock, Bioinformatics 22 (2006) 1803-1804.

[4] M.I. Zavodszky, L. Ming, M.F. Thorpe, A.R. Day, L.A. Kuhn, Modeling correlated main-chain motions in proteins for flexible molecular recognition, Proteins-Structure Function and Bioinformatics 57 (2004) 243-261.

[5] R.A. Sayle, E. J. Milner-White, RASMOL: biomolecular graphics for all., Trends Biochem Sci 20 (1995) 374.

[6] F. Osterberg, G.M. Morris, M.F. Sanner, A.J. Olson, D.S. Goodsell, Automated docking to multiple target structures: Incorporation of protein mobility and structural water heterogeneity in AutoDock, Proteins-Structure Function and Genetics 46 (2002) 34-40.

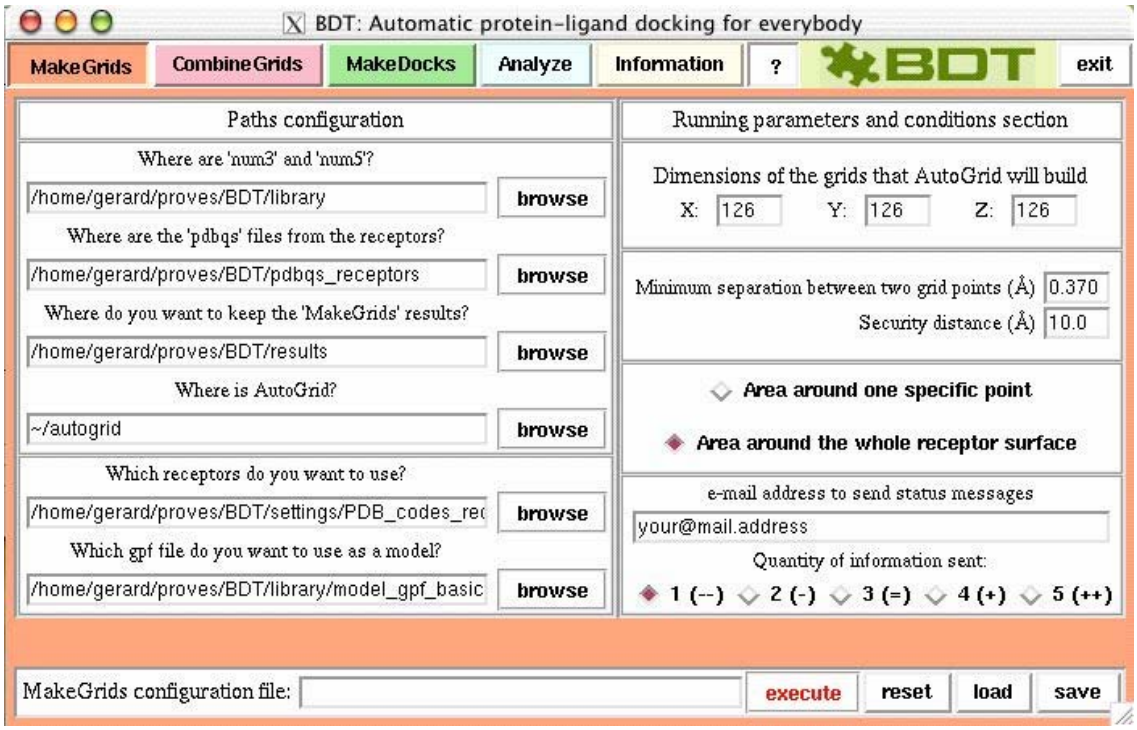
Figure 1. Startup window for BDT



The startup BDT window gives access to four *execution* tabs (*i.e.* **MakeGrids**, **CombineGrids**, **MakeDocks** and **Analyze**), which are sorted from left to right according to the sequence in which they are used when **AutoGrid**/**AutoDock** jobs are executed with BDT. The reason for this sequence of use is that the FORTRAN program that runs under each tab needs the previous tab to provide some input files. The **?** button activates/turns off the contextual help. The **exit** button allows the user to exit BDT properly (*i.e.* while the FORTRAN program runs in the background). The user can return to this startup window from each tab by pushing the **Information** button.

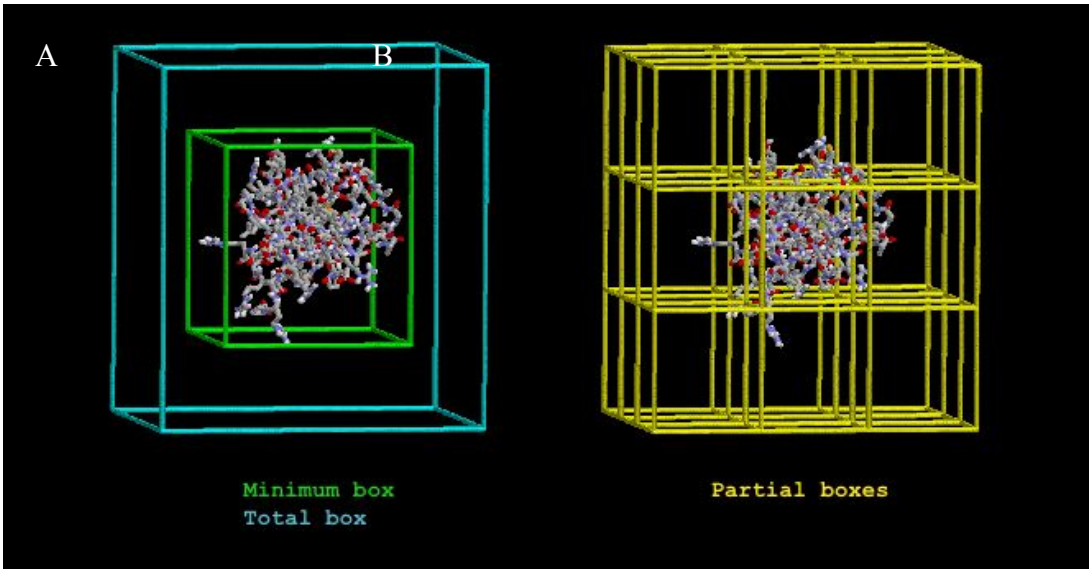
ISBN:978-84-691-1553-4 /DL: T.151-2008

Figure 2. The MakeGrids tab



This tab is used to: (a) control where **AutoGrid** has to run (*i.e.* around one specific point or around the whole the receptor); (b) define the number of grid points per dimension (*i.e.* the dimensions of what we call a *partial box*); (c) establish the minimum separation between two grid points in the lattice; and (d) set up the *Security distance* (though this is only used when the *Area around the whole receptor surface* option is activated). This tab can also be used to determine where **AutoGrid** is installed in the computer and where the various input files are needed for the run. Moreover, users can set up the quantity of information they want to receive by e-mail during **MakeGrids** run by selecting a number from 1 to 5 in the *Quantity of information sent* buttons (where the higher the number, the greater the quantity of information received by e-mail).

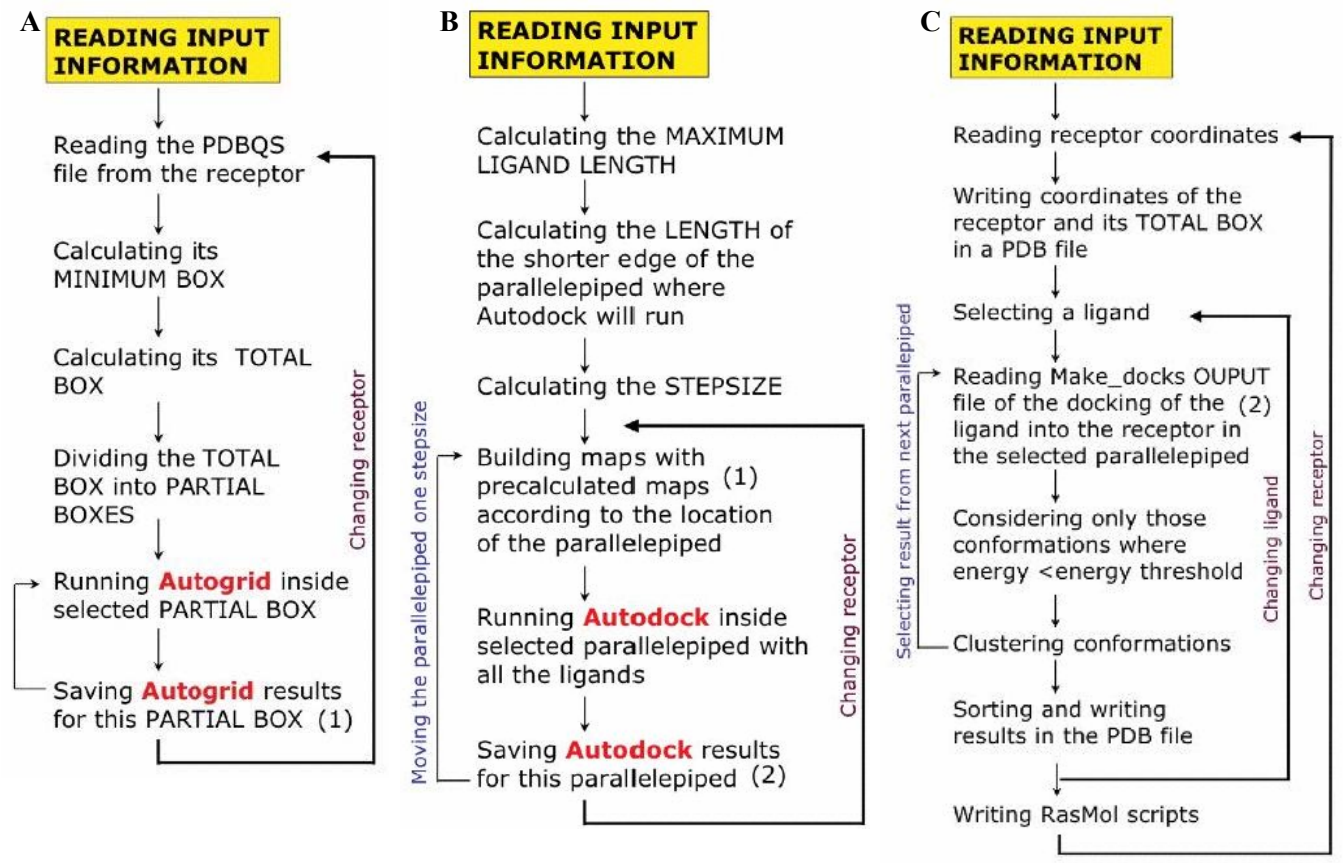
Figure 3. The boxes defined by **MakeGrids**



(A) MakeGrids takes the first receptor in the list and calculates the dimensions of the smallest box (the so-called *minimum box*; box in green) that can contain all the atoms of its different conformations. Then, **MakeGrids** adds the security distance to each *minimum box* dimension to obtain the so-called *total box* (box in blue). **(B) MakeGrids** divides the *total box* into *partial boxes* (boxes in yellow) and runs an **AutoGrid** process inside each one. Once all **AutoGrid** runs for this receptor have finished, the process is repeated with the next receptor in the list (see Figure 4A).

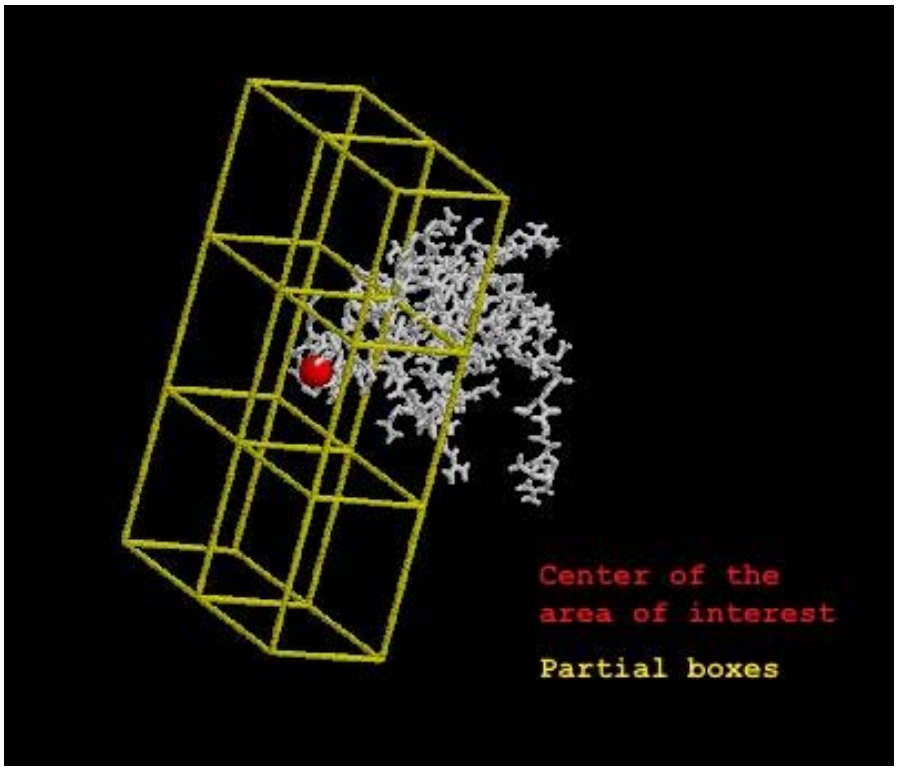
ISBN:978-84-691-1553-4 /DL: T.151-2008

Figure 4. Algorithm description of the programs that run under the **MakeGrids**, **MakeDocks** and **Analyze** tabs



Algorithm description of *make_grids* (panel A), *make_docks* (panel B) and *analyze* (panel C), the FORTRAN programs that run under the **MakeGrids**, **MakeDocks** and **Analyze** tabs, respectively

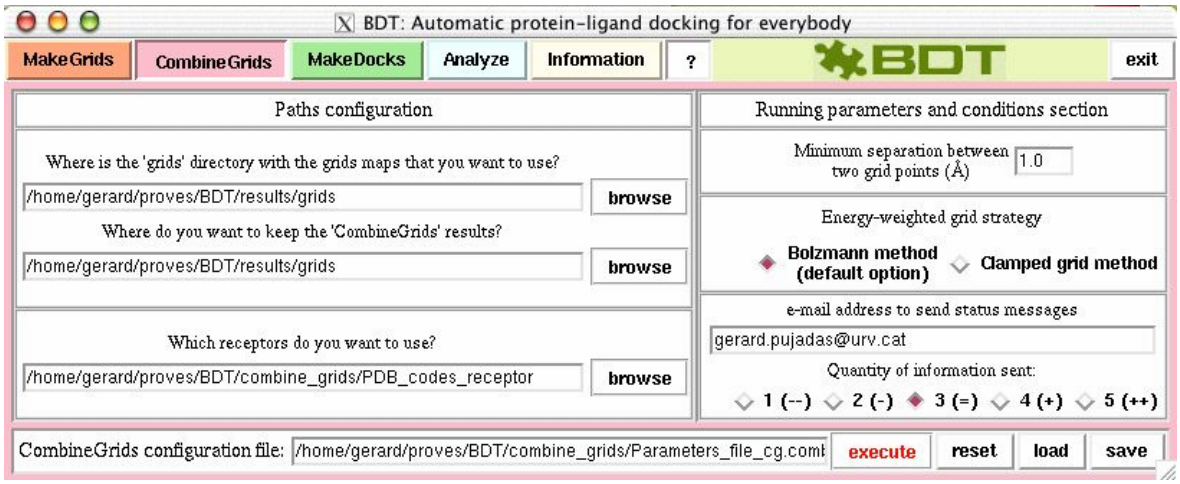
Figure 5. The result of running *make_grids* around a selected point



The user can select the *Area around one specific point* to run *make_grids* by indicating: (a) the coordinates of the central point of the receptor's area of interest (the selected point is indicated in this Figure by a red ball); and (b) the number of partial boxes (in yellow) that will be used in each dimension.

ISBN:978-84-691-1553-4 /DL: T.151-2008

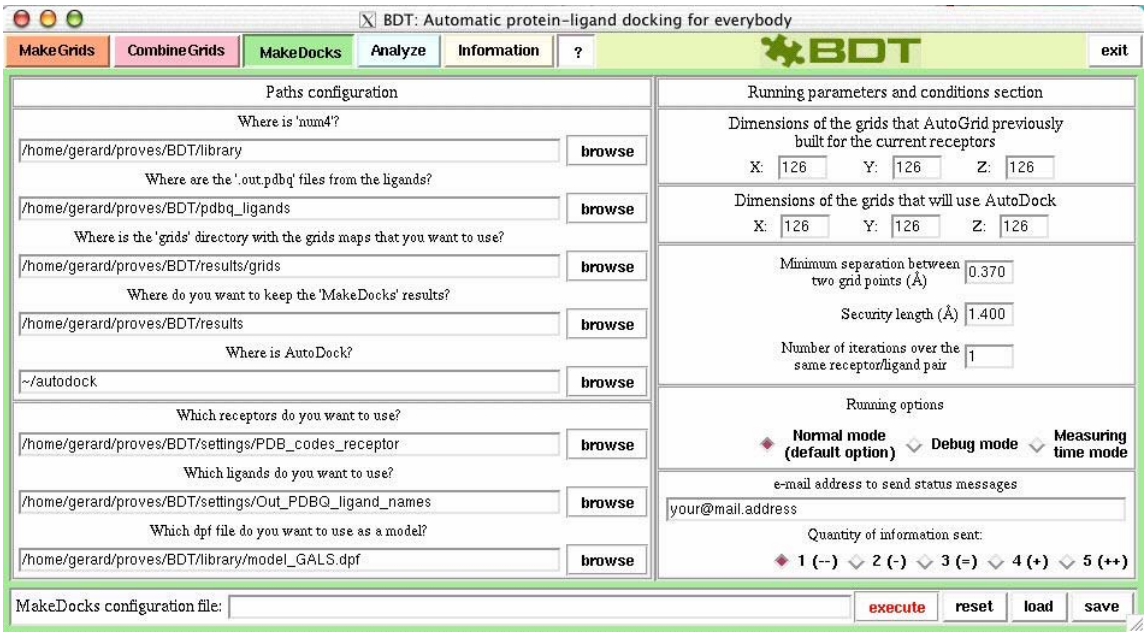
Figure 6. The **CombineGrids** tab



This tab is used to incorporate the receptor's mobility in docking calculations based on the work of Österberg et al. (2002). The resulting grid maps can be readily used by the **MakeDocks** tab.

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

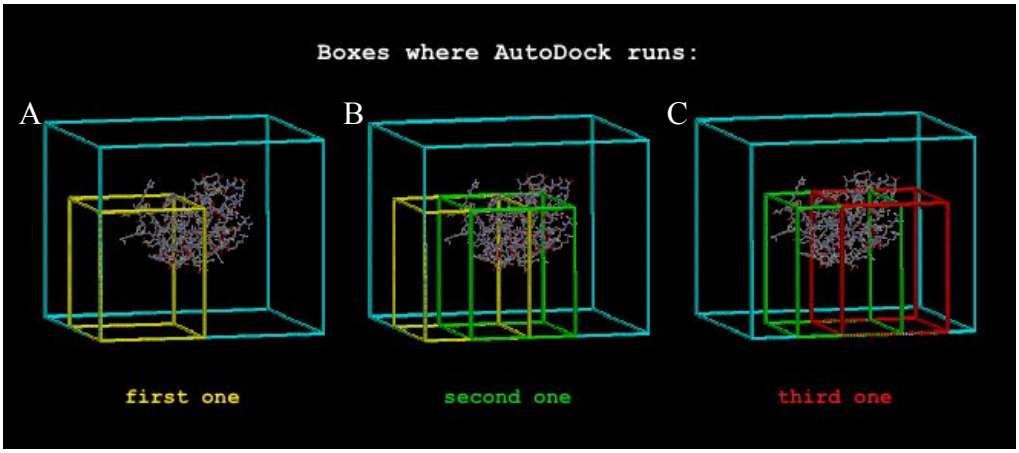
Figure 7. The **MakeDocks** tab



This tab enables the receptors, ligands and conditions used by **AutoDock** during docking to be selected easly. Depending on the origin of the grid maps used (obtained from either the **CombineGrids** or the **MakeGrids** tabs), the user will or will not consider receptor flexibility during docking. In this tab the user also has to fill in the next fields: (a) *Dimensions of the grids that will use AutoDock* (which may be different from the dimensions of the *Dimensions of the grids that AutoGrid previously built for the current receptors* and are usually as large as possible in order to speed up *make_docks* calculations); and (b) the *Security length (Å)* (a quantity that is added to the largest distance between any two atoms in the ligands studied). Both parameters are only necessary when the receptor's area of interest is formed by more than one *partial box*.

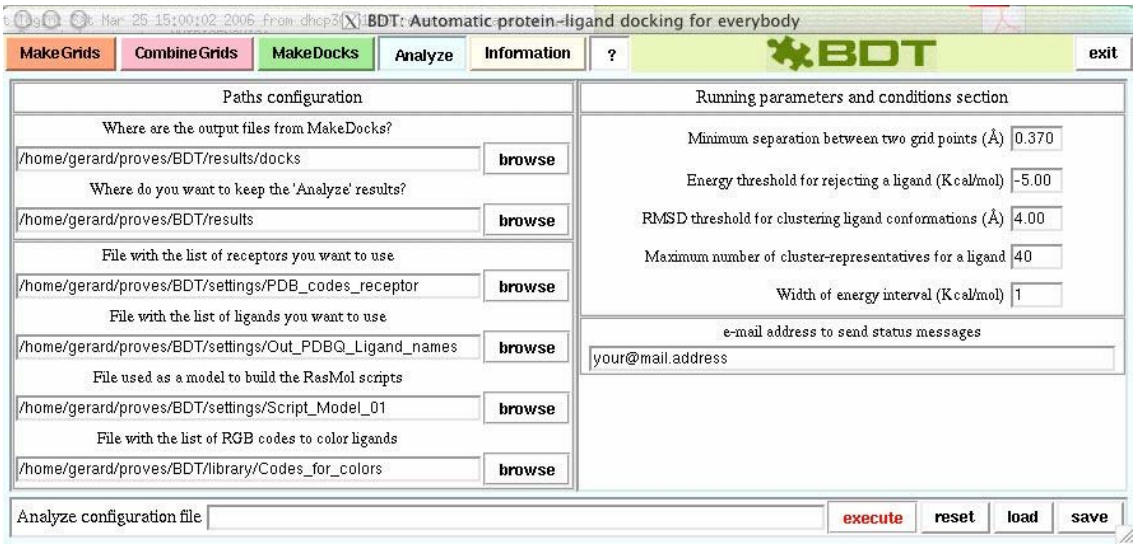
ISBN:978-84-691-1553-4 /DL: T.151-2008

Figure 8. The sequencial process of construction of boxes during **MakeDocks** runs.



At first, *make_docks* calculates the length of the shortest edge of the parallelepiped (which is not necessarily a cube) where each execution of **AutoDock** will run. From the difference between this value and the *maximum ligand length*, *make_docks* obtains what we call the *stepsize*. Then, *make_docks* selects the first receptor in the list and uses its pre-calculated grid maps to build all those that are needed to run **AutoDock** in the first parallelepiped (see panel **A**). After these maps are built, **AutoDock** tries to dock in them all the ligands from the list (one after the other). Then, *make_docks* will repeat this process but in a parallelepiped that has been *moved* away from the previous one by a distance that corresponds to the previously calculated *stepsize* (see panels **B** and **C**). Therefore, two adjacent **AutoDock** parallelepipeds always share enough volume to contain (even in their most extended conformations) all the ligands studied. This ensures that ligand-binding sites at the interface between adjacent parallelepipeds are not lost during blind docking. Using this method, *make_docks* analyzes the whole area of interest for the current receptor.

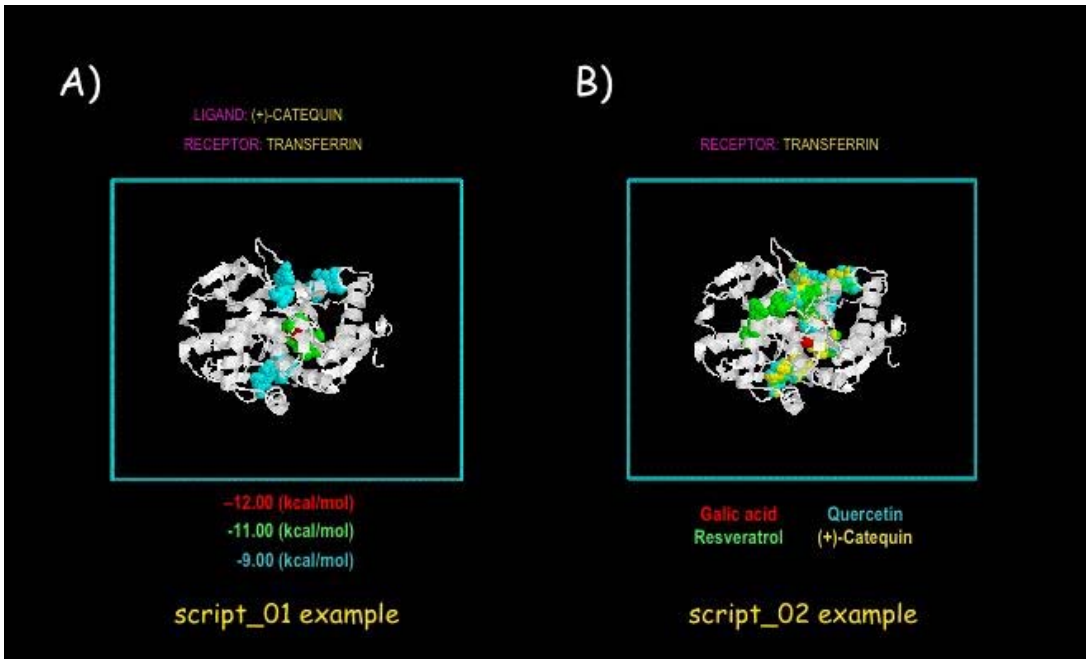
Figure 9. The Analyze tab



This tab can be used to analyze the docking results (*e.g.* by comparing the docking of different ligands onto the same receptor, etc.). From this tab, the user can select or set up: **(1)** the ligand-receptor pairs whose docking results will be analyzed; **(2)** the docking energy threshold below which a docking result for a ligand is rejected; **(3)** the root-mean-square deviation threshold used to cluster docking results for the same ligand (where the one with the highest receptor affinity is chosen as the cluster representative); **(4)** the width of the interval used in the **RasMol** scripts (http://www.openrasmol.org/) to colour cluster-representatives according to their receptor affinity; and **(5)** how many cluster representatives will be considered for each ligand. Once all this information is set up, the program under this tab (*i.e.* analyze) can be started.

ISBN:978-84-691-1553-4 /DL: T.151-2008

Figure 10. Visualization of the docking results with RasMol



Two **RasMol** scripts (http://www.openrasmol.org/) enable the docking results to be visualized more easly. The first script colours the ligands according to different intervals of affinity for the receptor irrespective of their identity (*i.e.* from hot to cold colours for decreasing affinity) and is therefore useful for identifying where the most important ligand-binding sites in the receptor structure are located (see panel A)**.** The second one colours the ligand according to its identity and is therefore useful for comparing the specificity of the different ligands for each binding site (see panel B).

BIOINFORMATICS APPLICATIONS NOTE Vol. 22 no. 14 2006, pages 1803–1804
doi:10.1093/bioinformatics/btl197

Structural bioinformatics

BDT: an easy-to-use front-end application for automation of massive docking tasks and complex docking strategies with AutoDock

Montserrat Vaqué[1], Anna Arola[1], Carles Aliagas[2] and Gerard Pujadas[1,*]

[1]Departament de Bioquímica i Biotecnologia, C/ Marcel·lí Domingo s/n and [2]Departament d'Enginyeria Informàtica i Matemàtiques, Av. Països Catalans, 26 Campus de Sant Pere Sescelades, Universitat Rovira i Virgili, Tarragona 43007, Catalonia, Spain

Received on March 27, 2006; revised on May 11, 2006; accepted on May 17, 2006
Advance Access publication May 23, 2006
Associate Editor: Martin Bishop

ABSTRACT

Motivation: AutoGrid/AutoDock is one of the most popular software packages for docking, but its automation is not trivial for tasks such as (1) the virtual screening of a library of ligands against a set of possible receptors; (2) the use of receptor flexibility and (3) making a blind-docking experiment with the whole receptor surface. This is an obstacle for research teams in the fields of Chemistry and the Life Sciences who are interested in conducting this kind of experiment but do not have enough programming skills. To overcome these limitations, we have designed BDT, an easy-to-use graphic interface for AutoGrid/AutoDock.

Availability: BDT is available for free, upon request, for non-commercial research.

Supplementary information: http://www.quimica.urv.cat/~pujadas/BDT/

Contact: gerard.pujadas@urv.cat

1 INTRODUCTION

Bioinformatic tools such as sequence similarity searches, sequence analysis and the homology modeling of protein sequences with unknown experimental 3D structure are currently used as standard research routines in biochemical and biomedical laboratories without extensive user-training. Crucial to extending the use of such tools beyond the borders of bioinformatic groups has been the development of graphic or web-based interfaces that are on top of such sophisticated algorithms. These interfaces are usually set up with default values for the parameters that control the algorithm that are valid in most situations and that allow novice or occasional users to use the tools easily. Moreover, the graphic interfaces make it easy to customize the values of these parameters for more specific or expert uses. A nice example of these graphic interfaces is Swiss-PdbViewer/Deep View (Guex *et al*., 1999), which has dramatically increased the use of homology modeling beyond bioinformatic groups.

It is important to understand intermolecular interactions between small ligands and their macromolecular receptors in order to explain crucial life processes such as gene expression, the regulation of metabolic pathways and enzyme catalysis. Using docking algorithms to predict the 3D structure of macromolecule/ligand complexes is therefore of interest for a wide range of biochemical and biomedical investigations. In this context, one of the most reliable, robust and popular energy-based docking packages is AutoGrid/AutoDock (Morris *et al*., 1998) because it allows a very efficient docking of flexible ligands (e.g. substrates, drug candidates, inhibitors, peptides, etc.) onto receptors (e.g. enzymes, antibodies, nucleic acids, etc.) even when the receptors are also flexible (Österberg *et al*., 2002).

While ADT (the graphic interface for AutoGrid/AutoDock) has strongly decreased the learning-curve needed for using this package, it is also true that some docking tasks with AutoGrid/AutoDock, although possible, are far from trivial for users without strong computer skills. Examples of such tasks are (1) using receptor flexibility during docking; (2) the automatic docking of a large library of ligands onto one or more receptors [because it is necessary to set up (a) one specific AutoGrid's command file for each receptor used and (b) one specific AutoDock's command file for each ligand–receptor pair that is assayed] and (3) docking a ligand library onto one or more receptors without defining one *a priori* ligand-binding site on them (therefore, using the whole receptor surface) and using a distance between grid points as short as the user needs. To overcome these difficulties, we have developed BDT, a Tcl/Tk graphic front-end application that runs on top of four Fortran programs (i.e. *make_grids*, *combine_grids*, *make_docks* and *analyze*, one under each BDT window tab), which control the conditions of AutoGrid and AutoDock runs.

2 DESCRIPTION

The BDT window has four tabs: (1) MakeGrids, (2) CombineGrids, (3) MakeDocks and (4) Analyze (see the BDT website at http://www.quimica.urv.cat/~pujadas/BDT/ for a detailed description of the tabs and the algorithms of the programs under them). The tabs are sorted from left to right according to their sequence of use [from (1) to (4) in the above list]. This is because the program that runs under each tab needs some input files that result from executing the program under the previous tab. Users are therefore recommended to move to the next tap until BDT informs them by e-mail that the execution of the program under one specific tap is finished.

*To whom correspondence should be addressed.

Four buttons are common to all tabs and have the same function: (1) 'execute', which starts running the Fortran program under the tab; (2) 'reset', which replaces custom values with the default ones and (3) 'save', which stores the current tab parameter values in a file for later recovery with the 'load' button. Users also have to set in all tabs how the underlying program will communicate with them by e-mail by (1) typing their e-mail address in the corresponding windows and (2) indicating the amount of information they want to receive from the program (this option is not available for the Analyze tab). BDT also has a contextual help (activated by default) for guiding users when filling the different fields in the tabs. Default parameter values are also provided for the different fields that are useful in most common situations.

From the MakeGrids tab, users can (1) include or exclude the flexibility of the receptors in the calculations and (2) search for the ligand-binding site in all the receptor surface ('Area around the whole receptor surface' option) or in a user-defined portion of it ('Area around one specific point' option) using, in both cases, a grid-point distance as short as they like regardless of the dimensions of the 3D space that is searched. Users can then combine the flexibility and ligand-binding site location strategies according to their needs (e.g. they can do the docking in a specific part of a flexible receptor). Thus, for each protein, this tab can automatically deal with either a single PDB file or with a set of PDB files corresponding to different snapshots of its conformations [such as those that (1) can be found when a specific protein has been crystallized in a set of different conditions and the resulting structures deposited in the PDB (http://www.pdb.org, Berman *et al*., 2000); (2) can be readily obtained from FlexWeb tools (http://flexweb.asu.edu/, Zavodszky *et al*., 2004) or (3) can be retrieved from the MODEL database (http://mmb.pcb.ub.es/MODEL)]. Here the user does not need to do anything special to take into account receptor flexibility. Receptor flexibility is automatically assumed for those receptors provided to this tab by the user's list that has more than one PDBQS file (where PDBQS corresponds to the input format for the receptors in AutoGrid) in the selected PDBQS directory [it is assumed that PDBQS filenames that start with the same four-character code (usually a PDB code) correspond to different conformations of the same receptor]. Therefore, this tab is used to control where AutoGrid has to run and its output files can be used either with the CombineGrids or the MakeDocks tabs (depending on whether the receptor's flexibility is considered).

The CombineGrids tab is used to incorporate the receptor's mobility in docking calculations based on the work of Österberg *et al*. (2002). Briefly, this method combines all the grid maps from the different receptor conformations and the same probe to obtain a single grid-map file. In this file, the energy of each point is obtained from a weighted average of the energies of the same point in all the original conformational-dependent grid maps (where the corresponding weight is calculated using either a clamped grid or a Boltzmann assumption based on the interaction energy). The resulting grid maps can be readily used by the MakeDocks tab.

The MakeDocks tab enables easy selection of the receptors, ligands and conditions used by AutoDock during docking. Depending on the origin of the grid maps used (obtained from either the CombineGrids tab or the MakeGrids tab), the user will or will not consider receptor flexibility during docking.

The Analyze tab can be used to analyze the docking results (e.g. by comparing the docking of different ligands onto the same receptor, etc.). For each studied receptor, Analyze e-mails the user a PDB file and two RasMol (Sayle and Milner-White, 1995) scripts to use with it. The PDB file contains the coordinates for (1) the self receptor; (2) the box where the possibility of ligand binding is studied and (3) the cluster representatives for docking solutions of the assayed ligands. The RasMol scripts make it easier to compare the docking results of the different ligands on the same receptor. The first script (i.e. script_01) colors the ligands according to different intervals of affinity for the receptor irrespective of their identity (i.e. from hot to cold colors for decreasing affinity) and is therefore useful for identifying where the most important ligand-binding sites in the receptor structure are located. The second one (i.e. script_02) colors the ligand according to its identity and is therefore useful for comparing the specificity of the different ligands for each binding site.

3 CONCLUSIONS

With just a few clicks of the mouse, BDT enables sophisticated docking strategies to be carried out, not only for research but also for teaching. Therefore, BDT contributes significantly to the progress of bioinformatics and biomedical research.

ACKNOWLEDGEMENTS

We thank the authors of AutoGrid/Autodock for providing us with version 3.0.5 of their software and, especially, Dr Garrett Morris and Dr Ruth Huey for their help. We also thank Kevin Costello of our University's Language Service for correcting the manuscript. This study was supported by grant number CO3/O8 from the Fondo de Investigación Sanitaria (FIS) and AGL2005-04889 from the Comisión Interministerial de Ciencia y Tecnología (CICYT) of the Spanish Government. Montserrat Vaqué is the recipient of a fellowship from grant number CO3/O8.

Conflict of Interest: none declared.

REFERENCES

Berman,H.M. *et al*. (2000) The Protein Data Bank. *Nucleic Acids Res*., **28**, 235–242.
Guex,N. *et al*. (1999) Protein modelling for all. *Trends Biochem. Sci*., **24**, 364–367.
Morris,G.M. *et al*. (1998) Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem*., **19**, 1639–1662.
Österberg,F. *et al*. (2002) Automated docking to multiple target structures: incorporation of protein mobility and structural water heterogeneity in AutoDock. *Proteins*, **46**, 34–40.
Sayle,R. and Milner-White,E.J. (1995) RasMol: biomolecular graphics for all. *Trends Biochem. Sci*., **20**, 333–379.
Zavodszky,M.I. *et al*. (2004) Modeling correlated main-chain motions in proteins for flexible molecular recognition. *Proteins*, **57**, 243–261.

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

Phosphoinositide 3-kinase α (PI3K α)

III

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

Introduction

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

Brief introduction

Phosphoinositide 3-kinases (also known as phosphatidylinositol 3-OH kinases, PI3-kinases or PI3Ks) are a ubiquitous family of enzymes that catalyze the transfer of the γ-phosphate group of ATP to the D3-hydroxyl position in the inositol head group of membrane phosphoinositides (PtdIns; for the atomic head-group nomenclature, see Hinchliffe & Irvine [1]). They are found in a wide range of species (including eukaryotes such as mammals, yeast, flies, slime mold, plants and algae) [2, 3]. PI3Ks and their second messenger lipid products are implicated in a plethora of cellular processes such as cell growth and survival, vesicle trafficking, cytoskeletal reorganization and chemotaxis, cell adhesion, B and T-cell development, superoxide production and insulin signaling [2-6]. The downstream effects of PI3Ks are spatially and temporally transmitted by their lipid products through specific interactions with effector proteins. Indeed, the majority of known PI3K-dependent events can be explained, at least in part, by identifying and characterizing the proteins which bind specific lipid products of PI3K largely by means of specific domains which are specialized for phosphoinositide binding [3].

Eight mammalian PI3Ks have been identified and classified into three major classes (I, II and III) on the basis of their sequence homology, binding of adaptor proteins (*i.e.* regulation) and substrate specificity [7-11]. Agonist-sensitive class I PI3Ks are the most extensively studied and, *in vitro,* they can 3'-phosphorylate different PtdIns (*i.e.* PtdIns, PtdIns(4)P and PtdIns(4,5)P_2; although *in vivo* they preferentially phosphorylate PtdIns(4,5)P_2 to produce PtdIns(3,4,5)P_3 [12, 13]). PtdIns(3,4)P_2 and PtdIns(3,4,5)P_3 are nominally absent in resting cells, while their levels rise rapidly in response to a wide range of stimuli [3]. Class I PI3Ks (the best characterized type) are heterodimeric proteins formed by a 110 kDa catalytic subunit (known as p110 subunits) and a smaller regulatory/adaptor subunit. According to the structure and mode of activation, class I PI3Ks are divided into subclasses IA and IB. At present, p110γ (the class IB catalytic subunit isoform) has been crystallized free or complexed with a variety of ligands or with Ras (see Table 1 and Figure 1; [14-18]) but no known three-dimensional structure has yet been obtained for the complete sequence of any class IA catalytic subunit isoforms (*i.e.* p110α, p110β, and p110δ). Nevertheless, recently, a structure of a short sequence-segment that corresponds to the adaptor-binding domain of p110α (first 108 residues in the sequence of p110α) in a complex with a segment of the p85α regulatory subunit that corresponds to its inter–Src homology 2 (inter-SH2) domain has been resolved [19]. However, previous data have shown that there are additional interactions between p85 and p110α [20]. Therefore, the complete description of structural details leading to the interaction of p110 subunits with their regulatory/adaptor subunit is still missing.

There is clear evidence to suggest that PI3Ks are involved in the pathogenesis of important human diseases [21]. For example, many tumour suppressor genes and oncogenes associated with tumorigenesis are components of cellular signaling networks that utilize PI3Ks. It should be noted that mutations of PTEN [an enzyme which dephosphorylates PtdIns(3,4,5)P3 to antagonize PI3K] are found in a significant number of cancers [22, 23]. Recent data suggest that most of the frequent mutations in cancer constitutively activate p110α [24, 25]. However, the increase in PI3K activity in cancer cells can also be produced by the loss of PTEN activity [26-28].

Although the catalytic domain of p110α, p110β, p110δ and p110γ shows a high sequence identity (only 4-5 out of 26 of the residues located at 5.0 Å of the ligand binding site are not identical; see Figure 2), isoform-specific inhibitors [21, 29-36] with promising therapeutic benefits in cancer, and immunological, inflammation and proliferative-based deseases have been developed and are expected to show few side effects [37-45]. Moreover, these inhibitors have overcome the limitations of interpreting the results obtained with knockout mice [46-49] and have shown that each isoform plays a different role in normal cell physiology and that they are subject to differential regulation.

Classification of class I PI3K subunits and isoforms

Class I PI3Ks are heterodimeric proteins consisting of a 110 kDa catalytic subunit that is encoded by at least four mammalian genes (*i.e.* PIK3CA, PIK3CB, PIK3CD, PIK3CG; usually known as PI3Kα, β, δ, and γ) and a constitutively associated smaller regulatory/adaptor subunit. On the basis of their structure and mode of activation, class I PI3Ks can be further subdivided into two groups (IA and IB).

Isoforms p110α, p110β, and p110δ are the catalytic subunits of class IA whereas the p110γ isoform is the catalytic subunit of class IB. Table 2 shows the sequence homology between class I catalytic subunits. From a structural point of view, the catalytic subunits of these PI3Ks are composed of several modular domains (see Figure 1). Most of these are common to all class I catalytic subunits (*i.e.* the catalytic lipid-kinase domain, the PI kinase ―PIK or helical― domain, the C2 phospholipid binding domain and the Ras-binding domain) whereas one is specifically found in α, β, and δ isoforms (*i.e.* the N-terminal domain which interacts with the regulatory subunit) [3].

Nine regulatory/adaptor subunits derived from three separate genes (PIK3R1, PIK3R2 and PIK3R3) have been identified in mammalians for class IA PI3Ks [50-54]. The adaptor subunits

of class IA PI3Ks have no intrinsic catalytic activity but they have a number of modular domains that enable protein-protein interactions (*i.e.* an SH3 domain, a Bcr homology domain flanked by two proline-rich regions and two C-terminal SH2 domains separated by an interSH2 domain; see Figure 3) such as the ones with the catalytic subunit (through its interSH2 domain [19, 55-59]). Three out of nine adaptor subunits are obtained by translating the corresponding genes (p85α, p85β and p55γ are obtained from PIK3R1, PIK3R2 and PIK3R3, respectively; see Figure 3) whereas the rest are derived either: (a) from an alternative splicing of the p85α mRNA that produces 53-55 kDa adaptor subunits in which the SH3 and BH domains are replaced by 6 or 34 residues to become p50α [60, 61] or p55α (also called p85/AS53 [62, 63]), respectively; (b) from a 24-nucleotide insertion that replaces the Asp605 in p85α by 9 other residues and which affects both p85α and p55α to produce two additional regulatory subunit variants ($p85\alpha_{+8a.a}$ and $p55\alpha_{+8a.a}$, respectively [62]); or (c) from an alternative splicing of the p55γ mRNA that produces 43-47 kDa adaptor subunits in which a 59 residue-long segment in the interSH2 domain (affecting both isoforms) or a 36 residue-long segment in the region between the BH and the first SH2 domains (affecting only the third isoform according to the information in the Q92569 entry of Swiss-Prot) are absent. Two regulatory subunits have been identified for class IB, the p101 [64] and p84/87 [65]. They have no homology with their equivalents in class IA and no protein-protein interaction motif in their sequence.

All mammalian cell types investigated to date express at least one of the three class IA PI3Ks. In this respect, the catalytic subunit isoforms α and β are ubiquitously expressed but δ is more restricted (it is primarily found in leukocytes). On the other hand, the regulatory subunit also shows differential tissue distribution (p85 α and β are ubiquitously expressed, whereas p50α and p55α are expressed in fat, muscle, liver and brain, and p55γ is mainly expressed in the brain) [2].

Regulation and activation of class I PI3Ks

Class IA PI3Ks are primarily regulated downstream of receptors with intrinsic tyrosine kinase activity or associated with non-receptor tyrosine kinases. In this respect, activation of PI3Ks by extracellular agonists involves the binding of SH2 domains from p85 to specific phosphotyrosine residues within the pYXXM motives in its receptors or in other signaling proteins. Moreover, it has been established that p110 with p85 subunits are constitutively bound and inactive in the cytosol [66]. Thus, it has been proposed that the p85 inhibition of p110α occurs through a charge-charge interaction between the p110 helical domain and one of the C-terminal SH2 domains in the p85 subunit [19]. Therefore, when p85 interacts with the tyrosine-phosphorylated motifs, the p85-mediated constraint on the p110 activity is liberated and the

ISBN:978-84-691-1553-4 /DL: T.151-2008

cytosolic complex is translocated to the plasma membrane in such a way that it accesses its physiological lipid substrate (*i.e.* PtdIns(4,5)P_2) [67, 68]. Synergistic with this activation, class IA PI3Ks are directly stimulated by the GTP-bound form of Ras [5, 6]. Class IB PI3K is activated directly by βγ subunits of heterotrimeric G-proteins (Gβγ) downstream of G-protein-coupled receptor activation (GPCR). The sensitivity to Gβγ is enhanced by the presence of p101, which facilitates the tethering of the PI3Kγ isoform to the plasma membrane. p110γ alone or p101/110γ are also activated directly by GTP-Ras, in synergy with Gβγ. On the other hand, class IA PI3Kβ can be promiscuously activated by both tyrosine kinase receptors and GPCR, although PI3Kγ is the only one that can be coupled to this receptor [69]. In the regulatory process of PI3K, the roles of the regulatory/adaptor subunits are numerous: for example, the down-regulation of the basal activity, stabilization of the catalytic subunit, activation downstream of receptor tyrosine kinases, and sequential activation by tyrosine kinases and Ras [70]).

The lipid products of class I PI3K have numerous biological functions because they have docking sites for proteins which control various metabolic process. These proteins contain specific lipid-binding domains that can bind to PtdIns such as the pleckstrin homology (PH) domain [which is found in Akt and in phospholipase C (PLC)] [71], the phox homology (PX) domain [72] and FYVE (which is found in Fab1, YOTB, Vac1 and EEA1) among others [73]. The activation of serine/threonine kinase of the protein kinase A, G and C family (such as PDK1 and the three isoforms of PKB/Akt) by PI3K are the signalling events that have been most studied. Thus, for instance, the activated PKB/AKT is able to phosphorylate a vast number of proteins which are crucial for promoting cell survival, controlling glucose homeostasis by promoting glucose transport through GLUT4 and controlling protein synthesis [66]. On the other hand, PI3K lipid products are the substrate of a series of phosphatases that specifically remove phosphate groups from their inositol ring. For example, the phosphatase and tensin homolog deleted on chromosome ten (PTEN) hydrolyses the phosphate in the D3 position of the inositol ring of PtdIns(3,4,5)P_3 which ends the PI3K-mediated signaling [66].

Role of class IA PI3K isoforms in the insulin signalling pathway and the glucose metabolism

PI3Ks mediate in a series of signal transduction pathways that are initiated by a wide variety of peptide or steroid hormones. However, the PI3K-dependent biochemical mechanism, in which PI3K isoforms are involved and from which PI3K-mediated specific biological responses are obtained, is only well characterized for a few peptides (for instance, insulin). It is well known that PI3K plays an essential role in insulin-stimulated glucose transport [15, 74-77]. In this respect, the binding of the hormone insulin to its receptor (IR) results in a receptor autophosphorylation on tyrosine residues. Then, activated IR triggers PI3K activity mainly by binding and phosphorylating adaptor proteins of the insulin receptor substrate (IRS) family such as IRS1, IRS2, IRS3 and IRS4 because phosphorylated IRS are docking sites for the SH2 domains of class IA p85 regulatory/adaptor subunits. Thus, the p85α, p85β and p55γ subunits modulate IRS by associating to PI3Ks differently [53].

Several reports using mutant mice as well as novel selective inhibitors suggest that different p110 subunits might play different roles. In this respect, recent experiments suggest that p110α is the most important form present in insulin signalling complexes and is required for signalling to downstream cell events [33]. Although several studies have indicated that p110β plays an important role in insulin signalling [76], biochemical analysis shows that only p110α is selectively recruited and activated by IRS proteins [78]. Furthermore, the use of p110 inhibitors with significant isoform selectivity shows that p110α is the major PI3K effector downstream from the IR and that, in contrast, p110β is dispensable but sets a phenotypic threshold for p110α activity [33]. Recently, a mechanism for activating p110α has been elucidated [*i.e.* the adaptor-binding domain of p110α in a complex with the p85α inter–Scr homology 2 (interSH2) has been crystallized]. Therefore, this model for the p110α/p85 heterodimer gives new insights into the architecture and mechanism of PI3Ks [19, 68].

The interaction between IRS and the PI3K regulatory subunit activates the enzyme and triggers the recruitment of PI3K to the membrane where, in turn, the PtdIns(3,4,5)P_3 production activates downstream effectors that control such metabolic processes as glucose uptake, lipolysis inhibition, triglyceride formation and glycogen synthesis. In adipocytes and muscle cells, after the phosphorylation of the lipid substrate, the signalling events start with the binding of the Ser/Thr kinase PDK1 to PtdIns(3,4,5)P_3 [which activates atypical forms of PKC (*i.e.* aPKC)] and the protein kinase PKB/Akt [79]. Of these downstream effectors, it is PKBβ/Akt2 that mainly stimulates glucose uptake from the blood, promotes synthesis of glycogen, and inhibits gluconeogenesis. Therefore, the final events of this signal transduction cascade are

ISBN:978-84-691-1553-4 /DL: T.151-2008

responsible for processes such as the translocation of the glucose tranporter GLUT4 from an intracellular compartment to the plasma membrane, and for other metabolic effects of insulin [77, 80] (it is well known that insulin increases glucose uptake mainly by enriching the concentration of GLUT4 proteins at the plasma membrane, rather than by increasing the intrinsic activity of the transporter). Furthermore, PI3K activation is attenuated by PtdIns(3,4,5)P_3 dephosphorylation via 3'-phosphatases such as PTEN or 5'-phosphatases such as SHIP2 [81]. For more information about insulin signalling pathways, see the schema proposed by Saltiel and Kahn (see Figure 4) [77].

References

[1] K. Hinchliffe, R. Irvine, Inositol lipid pathways turn turtle., Nature 390 (1997) 123-124.
[2] M.P. Wymann, L. Pirola, Structure and function of phosphoinositide 3-kinases, Biochim. Biophys. Acta 1436 (1998) 127-150.
[3] K.E. Anderson, S.P. Jackson, Class I phosphoinositide 3-kinases, Int. J. Biochem. Cell Biol. 35 (2003) 1028-1033.
[4] R. Katso, K. Okkenhaug, K. Ahmadi, S. White, J. Timms, M.D. Waterfield, Cellular function of phosphoinositide 3-kinases: Implications for development, immunity, homeostasis, and cancer, Annu. Rev. Cell Dev. Biol. 17 (2001) 615-675.
[5] L. Stephens, A. McGregor, P. Hawkins, Phosphoinositide 3-kinases: regulation by cell-surface receptors and function of 3-phosphorylated lipids, Biology of Phosphoinositides, vol. 27, 2000, pp. 32-108.
[6] B. Vanhaesebroeck, S.J. Leevers, K. Ahmadi, J. Timms, R. Katso, P.C. Driscoll, R. Woscholski, P.J. Parker, M.D. Waterfield, Synthesis and function of 3-phosphorylated inositol lipids, Annu. Rev. Biochem. 70 (2001) 535-602.
[7] M.J. Zvelebil, L. MacDougall, S. Leevers, S. Volinia, B. Vanhaesebroeck, I. Gout, G. Panayotou, J. Domin, R. Stein, F. Pages, Structural and functional diversity of phosphoinositide 3-kinases., Philos Trans R Soc Lond B Biol Sci 351 (1996) 217-223.
[8] J. Domin, M.D. Waterfield, Using structure to define the function of phosphoinositide 3-kinase family members., FEBS Lett 410 (1997) 91-95.
[9] B. Vanhaesebroeck, S.J. Leevers, G. Panayotou, M.D. Waterfield, Phosphoinositide 3-kinases: a conserved family of signal transducers., Trends Biochem. Sci 22 (1997) 267-272.
[10] A. Toker, L.C. Cantley, Signalling through the lipid products of phosphoinositide-3-OH kinase., Nature 387 (1997) 673-676.
[11] B. Vanhaesebroeck, M.D. Waterfield, Signaling by distinct classes of phosphoinositide 3-kinases., Exp Cell Res 253 (1999) 239-254.
[12] L.R. Stephens, K.T. Hughes, R.F. Irvine, Pathway of phosphatidylinositol(3,4,5)-trisphosphate synthesis in activated neutrophils., Nature 351 (1991) 33-39.
[13] P.T. Hawkins, T.R. Jackson, L.R. Stephens, Platelet-derived growth factor stimulates synthesis of PtdIns(3,4,5)P3 by activating a PtdIns(4,5)P2 3-OH kinase., Nature 358 (1992) 157-159.
[14] M. Camps, T. Rückle, H. Ji, V. Ardissone, F. Rintelen, J. Shaw, C. Ferrandi, C. Chabert, C. Gillieron, B. Françon, T. Martin, D. Gretener, D. Perrin, D. Leroy, P.A. Vitte, E. Hirsch, M.P. Wymann, R. Cirillo, M.K. Schwarz, C. Rommel, Blockade of PI3Kgamma suppresses joint inflammation and damage in mouse models of rheumatoid arthritis., Nat. Med. 11 (2005) 936-943.
[15] Z.A. Knight, B. Gonzalez, M.E. Feldman, E.R. Zunder, D.D. Goldenberg, O. Williams, R. Loewith, D. Stokoe, A. Balla, B. Toth, T. Balla, W.A. Weiss, R.L. Williams, K.M. Shokat, A pharmacological map of the PI3-K family defines a role for p110 alpha in insulin signaling, Cell 125 (2006) 733-747.
[16] M.E. Pacold, S. Suire, O. Perisic, S. Lara-Gonzalez, C.T. Davis, E.H. Walker, P.T. Hawkins, L. Stephens, J.F. Eccleston, R.L. Williams, Crystal structure and functional analysis of Ras binding to its effector phosphoinositide 3-kinase gamma, Cell 103 (2000) 931-943.
[17] E.H. Walker, M.E. Pacold, O. Perisic, L. Stephens, P.T. Hawkins, M.P. Wymann, R.L. Williams, Structural determinants of phosphoinositide 3-kinase inhibition by wortmannin, LY294002, quercetin, myricetin, and staurosporine, Mol. Cell 6 (2000) 909-919.
[18] E.H. Walker, O. Perisic, C. Ried, L. Stephens, R.L. Williams, Structural insights into phosphoinositide 3-kinase catalysis and signalling, Nature 402 (1999) 313-320.
[19] N. Miled, Y. Yan, W.C. Hon, O. Perisic, M. Zvelebil, Y. Inbar, D. Schneidman-Duhovny, H.J. Wolfson, J.M. Backer, R.L. Williams, Mechanism of two classes of

cancer mutations in the phosphoinositide 3-kinase catalytic subunit., Science 317 (2007) 239-242.

[20] J. Yu, C. Wjasow, J.M. Backer, Regulation of the p85/p110alpha phosphatidylinositol 3'-kinase. Distinct roles for the n-terminal and c-terminal SH2 domains., J Biol Chem 273 (1998) 30199-30203.

[21] M. Camps, T. Ruckle, H. Ji, V. Ardissone, F. Rintelen, J. Shaw, C. Ferrandi, C. Chabert, C. Gillieron, B. Francon, T. Martin, D. Gretener, D. Perrin, D. Leroy, P.A. Vitte, E. Hirsch, M.P. Wymann, R. Cirillo, M.K. Schwarz, C. Rommel, Blockade of PI3K gamma suppresses joint inflammation and damage in mouse models of rheumatoid arthritis, Nat. Med. 11 (2005) 936-943.

[22] D. Black, F. Bogomolniy, M.E. Robson, K. Offit, R.R. Barakat, J. Boyd, Evaluation of germline PTEN mutations in endometrial cancer patients., Gynecol Oncol. 96 (2005) 21-24.

[23] N.T. Nassif, G.P. Lobo, X. Wu, C.J. Henderson, C.D. Morrison, C. Eng, B. Jalaludin, E. Segelov, PTEN mutations are common in sporadic microsatellite stable colorectal cancer., Oncogene 23 (2004) 617-628.

[24] J.D. Carson, G. Van Aller, R. Lehr, R.H. Sinnamon, R.B. Kirpatrick, K.R. Auger, D. Dhanak, R.A. Copeland, R. Gontarek, P.J. Tummino, L. Luo, Effects of oncogenic p110alpha subunit mutations on the lipid kinase activity of phosphatidylinositol 3-kinase., Biochem. J (2007).

[25] Y. Samuels, K. Ericson, Oncogenic PI3K and its role in cancer., Curr Opin Oncol 18 (2006) 77-82.

[26] J.R. McMullen, P.Y. Jay, PI3K(p110alpha) inhibitors as anti-cancer agents: minding the heart., Cell Cycle 6 (2007) 910-913.

[27] P.K. Vogt, S. Kang, M.A. Eisliger, M. Gymnopoulos, Cancer-specific mutations in phosphatidylinositol 3-kinase, Trends Biochem. Sci 32 (2007) 342-349.

[28] L.H. Saal, K. Holm, M. Maurer, L. Memeo, T. Su, X. Wang, J.S. Yu, P.O. Malmström, M. Mansukhani, J. Enoksson, H. Hibshoosh, A. Borg, R. Parsons, PIK3CA mutations correlate with hormone receptors, node metastasis, and ERBB2, and are mutually exclusive with PTEN loss in human breast carcinoma., Cancer Res 65 (2005) 2554-2559.

[29] M. Hayakawa, H. Kaizawa, K. Kawaguchi, N. Ishikawa, T. Koizumi, T. Ohishi, M. Yamano, M. Okada, M. Ohta, S. Tsukamoto, F.I. Raynaud, M.D. Waterfield, P. Parker, P. Workman, Synthesis and biological evaluation of imidazo[1,2-a]pyridine derivatives as novel PI3 kinase p110 alpha inhibitors, Bioorg. Med. Chem. 15 (2007) 403-412.

[30] M. Hayakawa, H. Kaizawa, H. Moritomo, T. Koizumi, T. Ohishi, M. Okada, M. Ohta, S. Tsukamoto, P. Parker, P. Workman, M. Waterfield, Synthesis and biological evaluation of 4-morpholino-2-phenylquinazolines and related derivatives as novel PI3 kinase p110 alpha inhibitors, Bioorg. Med. Chem. 14 (2006) 6847-6858.

[31] T.B. Lanni, K.L. Greene, C.N. Kolz, K.S. Para, M. Visnick, J.L. Mobley, D.T. Dudley, T.J. Baginski, M.B. Liimatta, Design and synthesis of phenethyl benzo[1,4]oxazine-3-ones as potent inhibitors of PI3Kinase gamma, Bioorg. Med. Chem. Lett. 17 (2007) 756-760.

[32] Q.W. Fan, W.A. Weiss, Isoform specific inhibitors of PI3 kinase in glioma, Cell Cycle 5 (2006) 2301-2305.

[33] Z. Knight, B. Gonzalez, M.E. Feldman, E.R. Zunder, D.D. Goldenberg, O. Williams, R. Loewith, D. Stokoe, A. Balla, B. Toth, T. Balla, W.A. Weiss, R.L. Williams, K.M. Shokat, A pharmacological map of the PI3-K family defines a role for p110alpha in insulin signaling., Cell 125 (2006) 733-747.

[34] Z.A. Knight, G.G. Chiang, P.J. Alaimo, D.M. Kenski, C.B. Ho, K. Coan, R.T. Abraham, K.M. Shokat, Isoform-specific phosphoinositide 3-kinase inhibitors from an arylmorpholine scaffold, Bioorg. Med. Chem. 12 (2004) 4749-4759.

[35] V. Pomel, J. Klicic, D. Covini, D.D. Church, J.P. Shaw, K. Roulin, F. Burgat-Charvillon, D. Valognes, M. Camps, C. Chabert, C. Gillieron, B. Francon, D. Perrin, D. Leroy, D. Gretener, A. Nichols, P.A. Vitte, S. Carboni, C. Rommel, M.K. Schwarz, T.

Ruckle, Furan-2-ylmethylene thiazolidinediones as novel, potent, and selective inhibitors of phosphoinositide 3-kinase gamma, J. Med. Chem. 49 (2006) 3857-3871.

[36] M. Hayakawa, H. Kaizawa, H. Moritomo, T. Koizumi, T. Ohishi, M. Yamano, M. Okada, M. Ohta, S. Tsukamoto, F.I. Raynaud, P. Workman, M.D. Waterfield, P. Parker, Synthesis and biological evaluation of pyrido[3',2':4,5]furo[3,2-d]pyrimidine derivatives as novel PI3 kinase p110alpha inhibitors., Bioorg. Med. Chem. Lett. 17 (2007) 2438-2442.

[37] J. Doukas, W. Wrasidlo, G. Noronha, E. Dneprovskaia, R. Fine, S. Weis, J. Hood, A. DeMaria, R. Soll, D. Cheresh, Phosphoinositide 3-kinase gamma/delta inhibition limits infarct size after myocardial ischemia/reperfusion injury, Proc. Nat. Acad. Sci. U.S.A. 103 (2006) 19866-19871.

[38] C. Perrino, H.A. Rockman, M. Chiariello, Targeted inhibition of phosphoinositide 3-kinase activity as a novel strategy to normalize beta-adrenergic receptor function in heart failure, Vascul Pharmacol. 45 (2006) 77-85.

[39] T. Ruckle, M.K. Schwarz, C. Rommel, PI3K gamma inhibition: towards an 'aspirin of the 21st century'?, Nat. Rev. Drug Discovery 5 (2006) 903-918.

[40] P.M. Finan, S.G. Ward, PI3-kinase inhibition - A target for therapeutic intervention, Protein Tyrosine Kinases: from Inhibitors to Useful Drugs, 2005, pp. 53-69.

[41] S. Ward, Y. Sotsios, J. Dowden, I. Bruce, P. Finan, Therapeutic potential of review phosphoinositide 3-kinase inhibitors, Chem. Biol. 10 (2003) 207-213.

[42] R.C. Stein, M.D. Waterfield, PI3-kinase inhibition: a target for drug development?, Mol. Med. Today 6 (2000) 347-357.

[43] B.E. Drees, G.B. Mills, C. Rommel, G.D. Prestwich, Therapeutic potential of phosphoinositide 3-kinase inhibitors, Expert Opinion on Therapeutic Patents 14 (2004) 703-732.

[44] K.S. Lee, H.K. Lee, J.S. Hayflick, Y.C. Lee, K.D. Puri, Inhibition of phosphoinositide 3-kinase delta attenuates allergic airway inflammation and hyperresponsiveness in murine asthma model, FASEB J. 20 (2006) 455-465.

[45] K.S. Lee, S.J. Park, S.R. Kim, K.H. Min, S.M. Jin, K.D. Puri, Y.C. Lee, Phosphoinositide 3-kinase-delta inhibitor reduces vascular permeability in a murine model of asthma, Journal of Allergy and Clinical Immunology 118 (2006) 403-409.

[46] S.M. Brachmann, K. Ueki, J.A. Engelman, R.C. Kahn, L.C. Cantley, Phosphoinositide 3-kinase catalytic subunit deletion and regulatory subunit deletion have opposite effects on insulin sensitivity in mice, MCB 25 (2005) 1596-1607.

[47] E. Patrucco, A. Notte, L. Barberis, G. Selvetella, A. Maffei, M. Brancaccio, S. Marengo, G. Russo, O. Azzolino, S.D. Rybalkin, L. Silengo, F. Altruda, R. Wetzker, M.P. Wymann, G. Lembo, E. Hirsch, PI3K gamma modulates the cardiac response to chronic pressure overload by distinct kinase-dependent and -independent effects, Cell 118 (2004) 375-387.

[48] Z.A. Knight, K.M. Shokat, Features of selective kinase inhibitors, Chem. Biol. 12 (2005) 621-637.

[49] B. Vanhaesebroeck, K. Ali, A. Bilancio, B. Geering, L.C. Foukas, Signalling by P13K isoforms: insights from gene-targeted mice, Trends Biochem. Sci 30 (2005) 194-204.

[50] B.R. Dey, R.W. Furlanetto, S.P. Nissley, Cloning of human p55 gamma, a regulatory subunit of phosphatidylinositol 3-kinase, by a yeast two-hybrid library screen with the insulin-like growth factor-I receptor., Gene 209 (1998) 175-183.

[51] J.A. Escobedo, S. Navankasattusas, W.M. Kavanaugh, D. Milfay, V.A. Fried, L.T. Williams, cDNA cloning of a novel 85 kd protein that has SH2 domains and regulates binding of PI3-kinase to the PDGF beta-receptor., Cell 65 (1991) 75-82.

[52] M. Otsu, I. Hiles, I. Gout, M.J. Fry, F. Ruiz-Larrea, G. Panayotou, A. Thompson, R. Dhand, J. Hsuan, N. Totty, Characterization of two 85 kd proteins that associate with receptor tyrosine kinases, middle-T/pp60c-src complexes, and PI3-kinase., Cell 65 (1991) 91-104.

[53] S. Pons, T. Asano, E. Glasheen, M. Miralpeix, Y.T. Zhang, T.L. Fisher, M.G. Myers, X.J. Sun, M.F. White, The structure and function of p55(PIK) reveal a new regulatory subunit for phosphatidylinositol 3-kinase, MCB 15 (1995) 4453-4465.

[54] E.Y. Skolnik, B. Margolis, M. Mohammadi, E. Lowenstein, R. Fischer, A. Drepps, A. Ullrich, J. Schlessinger, Cloning of PI3 kinase-associated p85 utilizing a novel method for expression/cloning of target proteins for receptor tyrosine kinases., Cell 65 (1991) 83-90.

[55] K.H. Holt, L. Olson, W.S. Moye-Rowley, J.E. Pessin, Phosphatidylinositol 3-kinase activation is mediated by high-affinity interactions between distinct domains within the p110 and p85 subunits., Mol Cell Biol 14 (1994) 42-49.

[56] P. Hu, A. Mondino, E.Y. Skolnik, J. Schlessinger, Cloning of a novel, ubiquitously expressed human phosphatidylinositol 3-kinase and identification of its binding site on p85., Mol Cell Biol 13 (1993) 7677-7688.

[57] A. Klippel, J.A. Escobedo, Q. Hu, L.T. Williams, A region of the 85-kilodalton (kDa) subunit of phosphatidylinositol 3-kinase binds the 110-kDa catalytic subunit in vivo., Mol Cell Biol 13 (1993) 5560-5566.

[58] R. Dhand, K. Hara, I. Hiles, B. Bax, I. Gout, G. Panayotou, M.J. Fry, K. Yonezawa, M. Kasuga, M.D. Waterfield, PI 3-kinase: structural and functional analysis of intersubunit interactions., EMBO J 13 (1994) 511-521.

[59] A. Klippel, J.A. Escobedo, M. Hirano, L.T. Williams, The interaction of small domains between the subunits of phosphatidylinositol 3-kinase determines enzyme activity., Mol Cell Biol 14 (1994) 2675-2685.

[60] D.A. Fruman, L.C. Cantley, C.L. Carpenter, Structural organization and alternative splicing of the murine phosphoinositide 3-kinase p85 alpha gene., Genomics 37 (1996) 113-121.

[61] K. Inukai, M. Funaki, T. Ogihara, H. Katagiri, A. Kanda, M. Anai, Y. Fukushima, T. Hosaka, M. Suzuki, B.C. Shin, K. Takata, Y. Yazaki, M. Kikuchi, Y. Oka, T. Asano, p85alpha gene generates three isoforms of regulatory subunit for phosphatidylinositol 3-kinase (PI 3-Kinase), p50alpha, p55alpha, and p85alpha, with different PI 3-kinase activity elevating responses to insulin., J Biol Chem 272 (1997) 7873-7882.

[62] D.A. Antonetti, P. Algenstaedt, C.R. Kahn, Insulin receptor substrate 1 binds two novel splice variants of the regulatory subunit of phosphatidylinositol 3-kinase in muscle and brain., Mol Cell Biol 16 (1996) 2195-2203.

[63] K. Inukai, M. Anai, E. Van Breda, T. Hosaka, H. Katagiri, M. Funaki, Y. Fukushima, T. Ogihara, Y. Yazaki, A novel 55-kDa regulatory subunit for phosphatidylinositol 3-kinase structurally similar to p55PIK Is generated by alternative splicing of the p85alpha gene., J Biol Chem 271 (1996) 5317-5320.

[64] L.R. Stephens, A. Eguinoa, H. Erdjument-Bromage, M. Lui, F. Cooke, J. Coadwell, A.S. Smrcka, M. Thelen, K. Cadwallader, P. Tempst, P.T. Hawkins, The G beta gamma sensitivity of a PI3K is dependent upon a tightly associated adaptor, p101., Cell 89 (1997) 105-114.

[65] S. Suire, J. Coadwell, G.J. Ferguson, K. Davidson, P. Hawkins, L. Stephens, p84, a new Gbetagamma-activated regulatory subunit of the type IB phosphoinositide 3-kinase p110gamma., Curr Biol 15 (2005) 566-570.

[66] E. Hirsch, C. Costa, E. Ciraolo, Phosphoinositide 3-kinases as a common platform for multi-hormone signaling., J Endocrinol 194 (2007) 243-256.

[67] Z. Fu, E. Aronoff-Spencer, J.M. Backer, G.J. Gerfen, The structure of the inter-SH2 domain of class IA phosphoinositide 3-kinase determined by site-directed spin labeling EPR and homology modeling., Proc Natl Acad Sci U S A 100 (2003) 3275-3280.

[68] J.Y. Lee, J.A. Engelman, L.C. Cantley, Biochemistry. PI3K charges ahead., Science 317 (2007) 206-207.

[69] C. Murga, S. Fukuhara, J.S. Gutkind, A novel role for phosphatidylinositol 3-kinase beta in signaling from G protein-coupled receptors to Akt., J Biol Chem 275 (2000) 12069-12073.

[70] J. Yu, Y. Zhang, J. McIlroy, T. Rordorf-Nikolic, G.A. Orr, J.M. Backer, Regulation of the p85/p110 phosphatidylinositol 3'-kinase: stabilization and inhibition of the p110alpha catalytic subunit by the p85 regulatory subunit., Mol Cell Biol 18 (1998) 1379-1387.

[71] M.A. Lemmon, K.M. Ferguson, R. O'Brien, P.B. Sigler, J. Schlessinger, Specific and high-affinity binding of inositol phosphates to an isolated pleckstrin homology domain., Proc Natl Acad Sci U S A 92 (1995) 10472-10476.

[72] Y. Xu, L.F. Seet, B. Hanson, W. Hong, The Phox homology (PX) domain, a new player in phosphoinositide signalling., Biochem J 360 (2001) 513-530.

[73] T.G. Kutateladze, Phosphatidylinositol 3-phosphate recognition and membrane docking by the FYVE domain., Biochim Biophys Acta 1761 (2006) 868-877.

[74] P. Bevan, Insulin signalling., J Cell Sci 114 (2001) 1429-1430.

[75] C. Chaussade, G.W. Rewcastle, J.D. Kendall, W.A. Denny, K. Cho, L.M. Grønning, M.L. Chong, S.H. Anagnostou, S.P. Jackson, N. Daniele, P.R. Shepherd, Evidence for functional redundancy of class IA PI3K isoforms in insulin signalling., Biochem J 404 (2007) 449-458.

[76] R. Hooshmand-Rad, L. Hájková, P. Klint, R. Karlsson, B. Vanhaesebroeck, L. Claesson-Welsh, C.H. Heldin, The PI 3-kinase isoforms p110(alpha) and p110(beta) have differential roles in PDGF- and insulin-mediated signaling., J Cell Sci 113 Pt 2 (2000) 207-214.

[77] A.R. Saltiel, C.R. Kahn, Insulin signalling and the regulation of glucose and lipid metabolism., Nature 414 (2001) 799-806.

[78] L.C. Foukas, M. Claret, W. Pearce, K. Okkenhaug, S. Meek, E. Peskett, S. Sancho, A.J.H. Smith, D.J. Withers, B. Vanhaesebroeck, Critical role for the p110 alpha phosphoinositide-3-OH kinase in growth and metabolic regulation, Nature 441 (2006) 366-370.

[79] B. Vanhaesebroeck, D.R. Alessi, The PI3K-PDK1 connection: more than just a road to PKB., Biochem J 346 Pt 3 (2000) 561-576.

[80] P.H. Ducluzeau, L.M. Fletcher, H. Vidal, M. Laville, J.M. Tavaré, Molecular mechanisms of insulin-stimulated glucose uptake in adipocytes., Diabetes Metab 28 (2002) 85-92.

[81] T. Maehama, J.E. Dixon, PTEN: a tumour suppressor that functions as a phospholipid phosphatase., Trends Cell Biol 9 (1999) 125-128.

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF
PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

Table 1. Three-dimensional structures available for class I PI3Ks in the Protein Data Bank (http://www.pdb.org; [62]) with their most important features

PDB code	Resolution	R-Value	Source	Complex with	Bibliographic reference
1E7U	2.00	0.254	Sus scrofa	Wortmannin	Walker et al. (2000)
1E7V	2.40	0.273	Sus scrofa	LY294002	Walker et al. (2000)
1E8W	2.50	0.265	Sus scrofa	Quercetin	Walker et al. (2000)
1E8X	2.20	0.255	Sus scrofa	ATP	Walker et al. (2000)
1E90	2.70	0.272	Sus scrofa	Myricetin	Walker et al. (2000)
1E8Y	2.00	0.245	Homo sapiens		Walker et al. (2000)
1E8Z	2.40	0.232	Homo sapiens	Staurosporine	Walker et al. (2000)
1HE8	3.00	0.212	Homo sapiens	Ras	Pacold et al. (2000)
2A4Z	2.90	0.264	Homo sapiens	AS-604850	Camps et al. (2005)
2A5U	2.70	0.285	Homo sapiens	AS-605240	Camps et al. (2005)
2CHW	2.60	0.240	Homo sapiens	PIK-39	Knight et al. (2006)
2CHX	2.50	0.236	Homo sapiens	PIK-90	Knight et al. (2006)
2CHZ	2.60	0.246	Homo sapiens	PIK-93	Knight et al. (2006)

Table 2. Percent identity between the catalytic subunits from class I PI3Ks when the complete sequences (values above the diagonal) or the sequences from the catalytic lipid-kinase domains (values below the diagonal) are compared. The sequences were compared with the Clustal V method implemented in the MegAlign program from the Lasergene v7.1 package (http://www.dnastar.com). Default parameters were used for the comparisons.

p110α	**p110β**	**p110δ**	**p110γ**	
	33.7	33.9	27.5	**p110α** (P42336)
46.8		55.6	26.6	**p110β** (P42338)
45.6	72.9		25.2	**p110δ** (O00329)
41.0	41.8	39.5		**p110γ** (P48736)

Figure 1. Three-dimensional structure from the complex between pig p110γ and ATP corresponding to the PDB entry 1E8X [17]

Three-dimensional structure from the complex between pig p110γ (cartoon model) and ATP (ball & stick representation in red). The different domains are colored as follows: (a) Ras-binding domain (orange); (b) C2 domain (blue); (c) helical domain (magenta); and (d) catalytic domain (green). The figure has been done with the PDB entry 1E8X [17] and RasMol v2.7.3 (http://www.openrasmol.org/).

Figure 2. Multialignment of the sequences of the different PI3K-human isoforms and the p110γ isoform from pig. The figure also shows: (a) the Swiss-Prot code from each sequence; (b) the location of the α-helices and the β-strands according to annotations in 1E8Y (free human p110γ [17]); (c) the relative location in the sequence of the residues from which no coordinates are found in 1E8Y (this is shown as an absence of information in the line that contains the secondary structure assignment); (d) the secondary structure elements from residues with known coordinates in 1E8Y colored according to the domain to which they belong (*i.e.* brown, blue, magenta and green for Ras-binding, C2, helical [or PIK] and catalytic domain, respectively; [17]); (e) the relative location of important loops [*i.e.* the CBR1, the CBR2 and the CBR3 loops (the loops through C2 domains are often involved in Ca^{2+}-dependent or Ca^{2+}-independent phospholipid membrane binding and are located at one end of the domain); the P-loop; the NC-loop (the one that links the two catalytic-domain lobes); the catalytic loop; and the activation loop]; (f) the residues in p110γ with at least one atom within a radius of 5.0 Å from the ligands in any of the crystallized p110γ complexes (red dots at the bottom of a position in the alignment); (g) residues where mutations would abolish kinase activity (yellow background; [60]); (h) residues that interact putatively with the phosphate groups from PtdIns(4,5)P_2 (red characters; [17]); and (i) the residue that is thought to act as a general base to deprotonate the 3-hydroxyl of the PtdIns(4,5)P_2 substrate and generate a nucleophile that attacks the γ-phosphate of ATP (blue triangle at the bottom of the alignment). It is worth mentioning that differences in the secondary structure assignment between 1QMM (Walker et al., 1999) and 1E8Y [17] mean that the secondary-structure labels in the present figure are not equivalent to the ones in Figure 3 from Walker et al. (1999) (even though the criteria used to do the labelling was the same). The multialignment was done with the Clustal V method implemented in the MegAlign program from the Lasergene v7.1 package (http://www.dnastar.com) and default parameters were used for the comparisons.

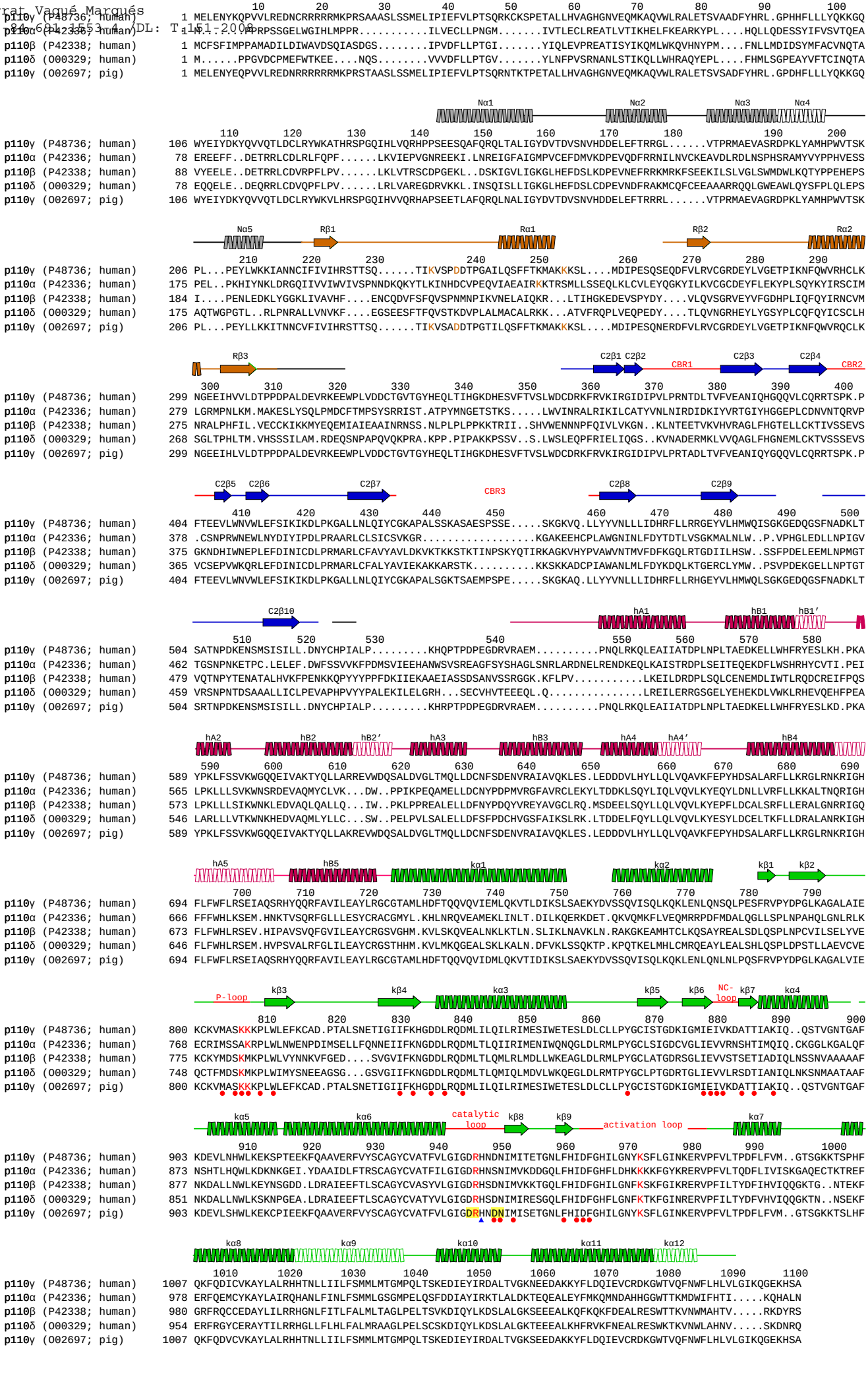
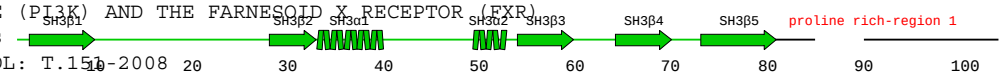


Figure 3. Multialigment between the three adaptor subunits from human class IA PI3Ks obtained by translation of PIK3R1 (p85α), PIK3R2 (p85β) and PIK3R3 (p55γ) genes. The corresponding codes in Swiss-Prot of the resulting proteins are also indicated beside the subunit name. The location of secondary structures relative to the p85α sequence has been taken from several PDB files because no single structure has been crystallized that corresponds to the complete sequence (the sequence interval that covers each PDB file is indicated in brackets): (a) 1PHT (3-85); (b) 1AZG (91-104); (c) 1PBW (115-298); (d) 2IUG (324-433); and (e) 1H9O (616-720). The absence of secondary structure elements in a sequence segment indicates that no structure has been solved for that segment of p85α. Secondary structure elements are colored according to the domain to which they belong (the sequence interval that corresponds to each domain is indicated in brackets): (a) the Src homology or SH3 domain in green (3-79); (b) the breakpoint-cluster-region homology or BH domain in blue (113-301); (c) the N-terminal SH2 or NSH2 domain in magenta (333-428); and (d) the C-terminal SH2 or CSH2 domain in yellow (624-718). Different p85α isoforms are obtained either by the alternate splicing of the mRNA (p50α and p55α are obtained by removing the 1-306 or 1-304 segments from p85α and replacing them by a 6 or 34 residue-long segment, respectively) or by a 24-nucleotide insertion that replaces the Asp605 (shown in a blue box) in p85α by 9 other residues. This insertion affects both p85α and p55α to produce two additional regulatory subunit variants ($p85\alpha_{+8a.a}$ and $p55\alpha_{+8a.a}$). Two p55γ isoforms have also been identified in which the 256-314 segment is absent and the 36-71 segment is only in one of them (shown in a green box). The residues involved in the interactions with the catalytic subunits in p85α (478-492) within the interSH2 domain are highlighted by a red box. Tyrosine or serine residues reported to be phosphorylated have a yellow background. Residues in which variants reduce insulin-stimulated activity have a red background (*i.e.* $R_{409}\rightarrow Q$ in p85α) while those that do not reduce this activity have a green background (where the variants are $M_{326}\rightarrow I$, $E_{451}\rightarrow K$ and $D_{330}\rightarrow N$ in p85α; and $R_{234}\rightarrow S$ and $P_{313}\rightarrow S$ in p85β). The multialignment was done with the Clustal V method implemented in the MegAlign program from the Lasergene v7.1 package (http://www.dnastar.com) and default parameters were used for the comparisons.

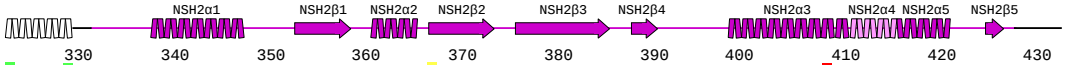
UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008




```
                10        20        30        40        50        60        70        80        90       100       110
p85α (P27986)    1 MSA.EGYQYRALYDYKKEREEDIDLHLGDILTVNKGSLVALGFSDGQEARPEEIGWLNGYNETTGERGDFPGTYVEYIGRKKISPPTPKPRPPRPLPVAPGSSKTEADVEQQA
p85β (O00459)    1 MAGPEGFQYRALYPFRRERPEDLELLPGDVLVVSRAALQALGVAEGGERCPQSVGWMPGLNERTRQRGDFPGTYVEFLGPVALARPGPRPRGPRPLPARPRDGAPEPG.....
p55γ (Q92569)      ..........................................................................................................
```


```
                120       130       140       150       160       170       180       190       200       210       220
p85α (P27986)  113 LTLPDLAEQFAPPDIAPPLLIKLVEAIEKKGLECSTLYRTQSSSNLAELRQLLDCDTPSVDLEMIDVHVLADAFKRYLLDLPNPVIPAAVYSEMISLAPEVQSSEEYIQLLKK
p85β (O00459)  109 LTLPDLPEQFSPPDVAPPLLVKLVEAIERTGLDSESHYRPELPAPRTDWSL........SDVDQWDTAALADGIKSFLLALPAPLVTPEASAEARRALREAAGP......VGP
p55γ (Q92569)      ..........................................................................................................
```


```
                230       240       250       260       270       280       290       300       310       320
p85α (P27986)  226 LIRSPSIPHQYWLTLQYLLKHFFKLSQTSSKNLLNARVLSEIFSPMLFRFSAASSDNTENLIKVIEILISTEW.............NERQPAPALPPKPPKPTTVANNGMNNN
p85β (O00459)  208 ALEPPTLPLHRALTLRFLLQHLGRVARRAPALGPAVRALGATFGPLLLRAPPPPSSPPPGGAPDGSEPSPDFPALLVEKLLQEHLEEQEVAPPALPPKPPKAKPAPTVLANGG
p55γ (Q92569)    1 ...........................................................MYNTVWSMDRDDADWREVMMPYSTELIFYIEMDPPALPPKPPKPMTSAVPNGMKD
```




```
                330       340       350       360       370       380       390       400       410       420       430
p85α (P27986)  326 M..SLQDAEWYWGDISREEVNEKLRDTADGTFLVRDASTKMHGDYTLTLRKGGNNKLIKIFHRDGKYGFSDPLTFSSVVELINHYRNESLAQYNPKLDVKLLYPVSKYQQDQV
p85β (O00459)  321 SPPSLQDAEWYWGDISREEVNEKLRDTPDGTFLVRDASSKIQGEYTLTLRKGGNNKLIKVFHRDGHYGFSEPLTFCSVVDLINHYRHESLAQYNAKLDTRLLYPVSKYQQDQI
p55γ (Q92569)   56 SSVSLQDAEWYWGDISREEVNDKLRDMPDGTFLVRDASTKMQGDYTLTLRKGGNNKLIKIYHRDGKYGFSDPLTFNSVVELINHYHHESLAQYNPKLDVKLMYPVSRYQQDQL
```

```
                440       450       460       470       480       490       500       510       520       530       540
p85α (P27986)  437 VKEDNIEAVGKKLHEYNTQFQEKSREYDRLYEEYTRTSQEIQMKRTAIEAFNETIKIFEEQCQTQERYSKEYIEKFKREGNEKEIQRIMHNYDKLKSRISEIIDSRRRLEEDL
p85β (O00459)  434 VKEDSVEAVGAQLKVYHQQYQDKSREYDQLYEEYTRTSQELQMKRTAIEAFNETIKIFEEQGQTQEKCSKEYLERFRREGNEKEMQRILLNSERLKSRIAEIHESRTKLEQQL
p55γ (Q92569)  169 VKEDNIDAVGKKLQEYHSQYQEKSKEYDRLYEEYTRTSQEIQMKRTAIEAFNETIKIFEEQCHTQEQHSKEYIERFRREGNEKEIERIMMNYDKLKSRLGEIHDSKMRLEQDL
```

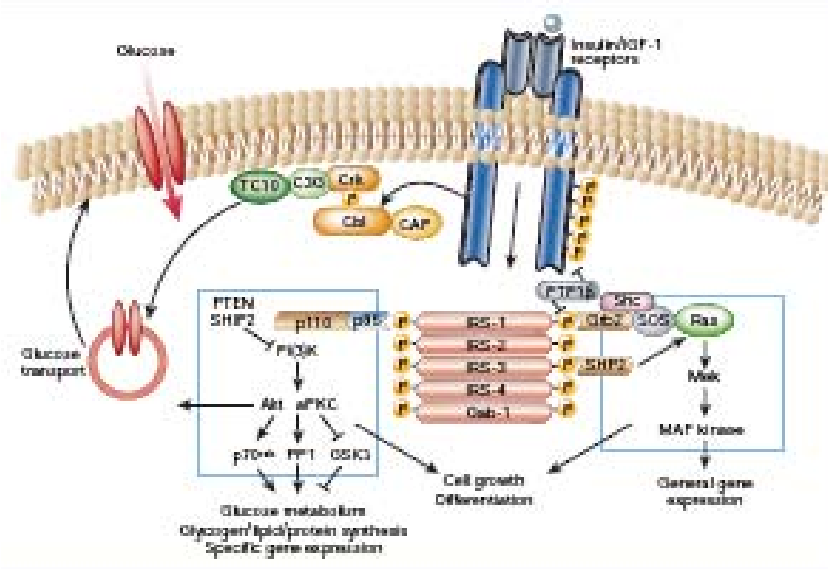

```
                550       560       570       580       590       600       610       620       630       640       650       660
p85α (P27986)  550 KKQAAEYREIDKRMNSIKPDLIQLRKTRDQYLMWLTQKGVRQKKLNEWLG..NENTEDQYSLVEDDEDLPHHDEKTWNVGSSNRNKAENLLRGKRDGTFLVRESSKQGCYACS
p85β (O00459)  547 RAQASDNREIDKRMNSLKPDLMQLRKIRDQYLVWLTQKGARQKKINEWLGIKNE.TEDQYALMEDEDDLPHHEERTWYVGKINRTQAEEMLSGKRDGTFLIRESSQRGCYACS
p55γ (Q92569)  282 KKQALDNREIDKKMNSIKPDLIQLRKIRDQHLVWLNHKGVRQKRLNVWLGIKNEDADENYFINEEDENLPHYDEKTWFVEDINRVQAEDLLYGKPDGAFLIRESSKKGCYACS
```


```
                670       680       690       700       710       720
p85α (P27986)  661 VVVDGEVKHCVINKTATGYGFAEPYNLYSSLKELVLHYQHTSLVQHNDSLNVTLAYPVYAQQRR......
p85β (O00459)  659 VVVDGDTKHCVIYRTATGFGFAEPYNLYGSLKELVLHYQHASLVQHNDALTVTLAHPVRAPGPGPPPAAR
p55γ (Q92569)  395 VVADGEVKHCVIYSTARGYGFAEPYNLYSSLKELVLHYQQTSLVQHNDSLNVRLAYPVHAQMPSLCR...
```

ISBN:978-84-691-1553-4 /DL: T.151-2008

Phosphoinositide 3-kinase α

Figure 4. Insulin signalling pathways according to Saltiel and Kahn [77]



UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

In silico prediction of the inhibitory activity of naturally occurring and bioactive forms of phenolic compounds on p110α

Montserrat Vaqué, Esther Sala, Gemma Montagut, Anna Ardévol, Cinta Bladé, M. Josepa Salvadó, Mayte Blay, Juan Fernández-Larrea, Lluís Arola, Gerard Pujadas*

Departament de Bioquímica i Biotecnologia. Universitat Rovira i Virgili,
C/ Marcel·lí Domingo s/n, Campus de Sescelades.
Tarragona 43007, Catalonia (Spain)

A short version of this manuscript has been submitted to
Chemical Information and Modeling

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

ABSTRACT

Recently it has been shown that the specific inhibition of the catalytic subunit of the class I PI3Kα isoform (*i.e.* p110α) blocks insulin-stimulated glucose uptake in 3T3-L1 adipocytes, L6 myotubes and mice that were treated with the same p110α inhibitors that have been used in both culture cells. In contrast, the specific inhibition of the catalytic subunits of the rest of class I PI3K isoforms does not affect insulin activity. It has also been reported that some naturally-occurring phenolic compounds, which are present in plant extracts that can be added during food production to obtain functional foods, inhibit p110α (*i.e.* resveratrol and naringenin) whereas others do not (*i.e.* genistein). Therefore, it seems that the p110α inhibitory activity of the phenolic compounds that are frequently found in these extracts (and their resulting metabolites) needs to be predicted in order to evaluate the possible side effects associated to their consumption. In order to do this, we have made a 3D quantitative structure-activity relationship study with the goal of predicting the concentration of each polyphenol that reduces the p110α activity to 50%. Our results show that stilbenes, flavonols (except myricetin), flavones, flavanones, anthocyanidins (except delphinidin and delphinidin 3-glucoside), most flavanol monomers [except (+)-epicatechin, (+/-)-epigallocatechin, (+/-)-gallocatechin, (-)-catechin and 4'-*O*-methyl-(-)-epigallocatechin], all procyanidin dimers and some isoflavones (*i.e.* daidzin, genistin, glycitin and glycitein) are able to map most of the features of a pharmacophore built with data from synthetic p110α inhibitors and, therefore, they are potential inhibitors of this enzyme.

ABBREVIATIONS

EGCG: Epigallocatechin gallate; PDB: Protein Data Bank; PI3K: Phosphoinositide 3-kinase; PKB: Protein Kinase B; QSAR: Quantitative Structure Relationship; RMS: Root-Mean Square; RMSD: Root-Mean Square Deviation

KEYWORDS

Catalyst; protein-ligand docking; eHiTS, procyanidin extract, SAR study, 3D-QSAR study, IC50 prediction

SUPPLEMENTARY INFORMATION

Supplementary materials and information for this paper are available at the following URL: http://www.quimica.urv.cat/~pujadas/PI3K_01

ISBN:978-84-691-1553-4 /DL: T.151-2008

INTRODUCTION

One of the major biological responses induced by insulin via PI3K in muscles and adipose tissue is the translocation of the glucose transporter GLUT4 at the plasma membrane, which leads to an increase in glucose uptake [1, 2]. Recent studies with isoform-specific inhibitors have shown that p110α (the catalytic subunit of PI3Kα) is crucial for this insulin-stimulated glucose-uptake in such cell lines as 3T3-L1 adipocytes, L6 myotubes and CHO-IR [3, 4]. Furthermore, when the remaining class IA PI3K catalytic subunit isoforms (*i.e.* p110β, p110δ) are specifically inhibited, no effect is observed on the glucose uptake of the same cell lines [3, 4]. Similar effects to those described *in vitro* for ATP-competitive inhibitors are also observed *in vivo* [4]. Thus, insulin-tolerance tests demonstrate that the same p110α specific-inhibitors can prevent blood glucose levels from decreasing in fasted mice that have been intravenously injected with insulin [4] whereas p110β/p110δ specific-inhibitors have no effect on blood glucose levels in response to insulin [4]. Moreover, all these results have been corroborated by an approach that uses transgenic gene knock-in mice in which the p110α gene carries a mutation (D933A) that suppresses its lipid-kinase activity [5]. In this experiment, the mice that are homozygous for this mutation show embryonic lethality whereas those that are heterozygous are viable but have reduced p110α activity [5]. The main results obtained by this knock-in approach were that: (a) p110α is the most important PI3K catalytic isoform present in insulin signalling complexes; (b) heterozygous mice become insulin resistant with age; and (c) insulin signalling to protein kinase B (*i.e.* PKB) is relatively normal in liver but is severely impaired in muscle and fat (this is consistent with recent findings that show that p110α is required but is not sufficient to mediate insulin signalling in HepG2 hepatoma cells [3]). Therefore, it is concluded that p110α is required for the signalling of insulin-downstream events in two primary insulin-responsive tissues like muscle and fat [3-5] and that its dysfunction may have deleterious effects on glucose homeostasis [6].

Phenolic compounds are commonly found in fruits and vegetables [7-9] and overall have beneficial effects on human health: for example, they are cardioprotective, antioxidant, antigenotoxic, anti-inflammatory and anticarcinogenic [10-13]. Extracts that contain these molecules can be added to food to convert it into functional food (*i.e.* food that has health benefits beyond the traditional nutrients it contains). Nevertheless, it is also well known that whereas some of these phenolic compounds (*i.e.* catechin and gallic acid) do not affect glucose uptake, others (*i.e.* quercetin, myricetin, catechin-gallate, resveratrol and naringenin) reversibly inhibit it in isolated rat adipocytes, 3T3-L1 adipocytes, rat L6 myotubes and human muscle-derived cell lines [14-16]. In this respect, it has been shown that resveratrol acts as a reversible and competitive inhibitor of ATP in p110α [14] whereas naringenin has no effect on the kinase

activity of the insulin receptor or on the tyrosine phosphorylation of its immediate substrates. Rather, it inhibits the phosphorylation of the downstream signaling molecule PKB [15]. In contrast, it has been suggested that quercetin, myricetin and catechin-gallate act directly by competing with glucose for its binding site in GLUT4 [16]. Nevertheless, it is plausible that quercetin, myricetin and catechin-gallate also partially inhibit an insulin-dependent end-point response such as insulin-induced glucose uptake in isolated rat adipocytes by direct binding to p110α. This is supported by two facts: (a) complexes between another class I PI3K isoform (*i.e.* p110γ) and quercetin or myricetin have been obtained and both ligands bind to the ATP-binding site [Protein Data Bank (PDB; http://www.pdb.org) entries have 1e8w and 1e90 codes for quercetin and myricetin complexes, respectively] [17]; and (b) no experiment have discounted that these phenolic compounds can bind to p110α [16]. Therefore, it is reasonable to think that the inhibition on the glucose uptake produced by quercetin, myricetin, catechin-gallate, resveratrol and naringenin [14-16] is, at least in part, a consequence of their competition with ATP for the ATP-binding site in p110α. In fact, some molecules that non-specifically inhibit all PI3K catalytic subunit isoforms by competing with ATP have also been synthesized on the basis of naturally-occurring phenolic compounds (*e.g.* the chromome LY294002 was synthesized using quercetin as a model [18]).

Therefore, it seems that the potential of the different phenolic compounds for inhibiting p110α must be predicted to prevent the possible side effects associated with their consumption because, for instance, it has been suggested that the regular consumption of naringenin-rich grapefruit may exacerbate insulin resistance in susceptible individuals via impaired glucose uptake in the adipose tissue [15]. It has also been recently suggested that p110α is essential for maintaining cardiac function in response to a pathological cardiac insult and that p110α inhibitors result in more fibrosis, which adversely increases the stiffness of the myocardium [19]. Therefore, there are concerns about the use of molecules that inhibit p110α in cancer patients with cardiovascular risk factors [20]. Thus, we have made a 3D quantitative structure-activity relationship study (*i.e.* 3D-QSAR study) with the goal of predicting the IC50 (*i.e.* the polyphenol concentration that reduces p110α activity to 50%) of: (a) the most frequent phenolic compounds found in plant extracts [21]; and (b) the bioactive structures of the phenolic compounds detected in plasma or urine [22-25]. The latter group of molecules have been metabolized by mechanisms of intestinal and/or hepatic conjugation in which they are glucuronidated, methylated, glycosilated or sulfated, and/or hydrolized by the intestinal microflora and they have been added to the study because they may contribute to the biological effects of the phenolic compounds in humans. Therefore, in order to carry out the 3D-QSAR study, we obtained a pharmacophore (*i.e.* a 3D arrangement of the ligand's functional groups that are important for its activity when it binds to a defined target) by using available data on the

ISBN:978-84-691-1553-4 /DL: T.151-2008

IC50 and the molecular structure of a relatively large set of synthetic p110α inhibitors [26, 27]. Then, the resulting pharmacophore was used to map all the conformations that are possible for each of the phenolic compounds analyzed and predict their IC50 value by assuming that the IC50 of a molecule is the lowest one that can be obtained by any of its conformations. Our predictions show excellent agreement with the reported experimental findings [14-16]. Therefore, this supports the reliability of the IC50 values obtained for the rest of the phenolic compounds with the same pharmacophore.

MATERIALS AND METHODS

Pharmacophore development

Set up of the training set

We are working with a set of 36 synthetic compounds which, according to the literature, act as p110α inhibitors [26, 27]. Their IC50 values span 5 orders of magnitude (ranging from 0.0018 to 60 μM) and each order of magnitude is represented by three compounds or more. Furthermore, all the activity values have been obtained in the same experimental conditions by the same scientific team.

The most important aspect of using HypoGen™ [a module from Catalyst™ v4.11 (Accelrys Inc., San Diego, CA, USA; http://www.accelrys.com/products/catalyst/)], is the selection of the molecules that will be used in the hypotheses or pharmacophore generation (where a hypothesis or pharmacophore is a combination of chemical functions in the three-dimensional space that correlates the ligand structure with its effect on the activity of the receptor to which it binds). For this reason, this set (the so called *training set*) must contain molecules that are diverse enough in structure and activity to ensure the statistic relevance of the calculated pharmacophore models (special care must also be taken into account to avoid redundant information in the training set). Since the 36 compounds from which the training set is to be obtained have a common receptor binding mode, the resulting pharmacophores should be able to capture the essential features of the ligand structure effect on p110α activity. Furthermore, the training set must contains a mixture of active and less active compounds (sometimes called inactive compounds), which will depend on the rules applied by HypoGen™ during the *constructive* and *subtractive* phases (see below in the *HypoGen™ phases* section of **Hypothesis generation**). This will allow HypoGen™ to build pharmacophore candidates that correctly distinguish between active and inactive compounds when it is used to predict the IC50 values of other compounds that bind to the p110α target. Thus, during the HypoGen™ *constructive* phase,

active compounds are identified as those in which the result of the expression **MA*Unc(MA) – A/Unc(A)** is higher than 0.0 [where: (a) **MA** is the activity of the most active compounds; (b) **A** is the activity of the analyzed compound; and (c) **Unc(MA)** and **Unc(A)** are the uncertainties associated with **MA** and **A**, respectively (where these account for the intrinsic variability associated with the measure of biological data; 3 by default)]. During the second phase (the so called *subtractive* phase), inactive compounds are identified as those in which the activity (*i.e.* IC50) is 3.5 orders of magnitude greater than the corresponding value for the most active compound (*i.e.* **MA**). The value of 3.5 is used by default and it is very important to make sure that there are enough inactive compounds in the training set. If there are not enough, more can be obtained by setting a lower value for the *GenerateHypo.inactive.spread* parameter in the **.Catalyst** file generated by Catalyst™ [where this file is located both in the run and the home directories and it is read every time a Catalyst™ job (either interactively or in background) is started]. At this point, it is worth to pointing out that the higher the IC50 of a molecule is, the lower is its inhibitory power.

Taking into account all the above considerations, we selected 21 out of the 36 molecules to build the training set (see Figure 1).

Ligand structure building and conformation generation

Once the molecules that constitute the training set had been identified, their 3D structure was built and further minimized so that they could be used within Catalyst™. This was done with ChemDraw Ultra™ v10.0 (CambridgeSoft Corporation, Cambridge, MA, USA; http://www.cambridgesoft.com/software/details/?ds=2&dsv=9) and the same procedure was used with the other p110α ligands needed in this study (where the resulting molecular structures can be downloaded from the paper's website). Once built and minimized, all the molecules were imported into Catalyst™ where they were submitted to conformational analysis. Catalyst™ provides two types of conformational search, BEST and FAST. As recommended when these conformations are used in a HypoGen™ process, we used the BEST conformational search [Catalyst™ v4.11 Tutorials, Accelrys Inc., San Diego, CA, USA, 2005]. The BEST option uses the Poling algorithm [28], which: (a) optimizes the conformers in both torsional and Cartesian space; (b) is more precise; and (c) provides more complete coverage of the conformational space. Moreover, we set 255 to be the maximum number of conformers [in order to ensure the maximum coverage of the conformational space (Catalyst™ v4.11 Tutorials, Accelrys Inc., San Diego, CA, USA, 2005)] whereas the energy constraint (*i.e.* a value that controls the maximum difference in energy between the lowest and the highest-energy conformers generated) was set to 15 Kcal/mol. At this point, it is worth mentioning that setting a very narrow energy range for this constraint is dangerous because it restricts the number of conformers that Catalyst™

generates (because they must all have an energy that is within this range) and, therefore, negatively affects subsequent hypothesis generation and activity prediction.

Hypothesis generation

We used the HypoGen™ module from Catalyst™ to generate pharmacophore candidates by correlating the IC50 experimental data on p110α of the molecules in the training set with their previously generated conformations. The uncertainty value used in all HypoGen™ runs was 3 (*i.e.* the default value). Prior to any HypoGen™ run, an initial analysis of the training set was made and it was detected that these molecules have four main chemical functions (*i.e.* functional groups, features or functions in the Catalyst™ argot): (a) the H-bond acceptor (HA); (b) the H-bond donor (HD); (c) the hydrophobic group (HY); and (d) the aromatic ring (RA). Therefore, we decided to select only these four functions in an attempt to minimize the execution time of HypoGen™ and achieve a good combination of the functions that describe the relation between the structure and activities in our training set. These chemical functions also describe the most important kind of interactions between receptor and ligand.

HypoGen™ phases

In order to build a pharmacophore that correlates the ligand structures with their biological activity on a specific receptor (here p110α), HypoGen™ followed three consecutive phases: (1) constructive; (2) subtractive; and (3) optimization. In the constructive phase, HypoGen™ used all the conformations of the two most active compounds in the training set (*i.e.* 2g and 15e; see Figure 1) to build all the possible pharmacophore candidates by combining all the considered chemical functions (*i.e.* HA, HD, HY and RA). Therefore, the number of candidates generated in this way depends on the structural and chemical diversity of these two compounds. Then, HypoGen™ considered that a minimum subset of features of the remaining most active compounds (see Figure 1) must also be also matched by the pharmacophore candidates and, therefore, only those hypotheses which fit a minimum subset of features of these molecules were retained for further evaluation. Finally, this first phase ended when a database which contained all the remaining pharmacophore candidates was built (the number of candidates in it depends on the chemical diversity of the most active compounds in the training set). In the subtractive HypoGen™ phase, the pharmacophore candidates in the database were analyzed and, when one of them was matched by more than half of the least active compounds (*i.e.* the ones with an IC50 for p110α that is 3.5 orders of magnitude greater than the value for the most active compound; see Figure 1), it was removed. Finally, in the optimization phase, HypoGen™ tried to improve the remaining hypothesis candidates either by modifying, adding or deleting features and/or by rotating vectors via a simulated annealing approach. Then, each modification was evaluated using the error cost value (see section below for a definition) and finally, the best

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

ten resulting hypotheses were reported (scored according to their overall cost; see next section for overall cost calculation details).

During the hypothesis generation process, we reviewed the parameters of the HypoGen™ run [for instance, the configuration cost (see next section for its meaning) which is calculated early in the process] in order to abort incorrect generation of pharmacophores. In our first HypoGen™ run, the configuration cost was too high and had to be aborted. The HypoGen™ module can reduce the value of the configuration cost in a number of different ways (for instance by imposing restrictions on the number and types of features that are included in the hypothesis). We had already restricted the types of feature that our compounds could contain to HA, HD, HY and RA. Therefore, we further restricted the minimum number of each of the functions that the hypothesis allows. We specified 1 for the individual minimum value for HA, HY and RA (this means that the program was forced to include one HA, one HY and one RA function in the pharmacophore) but we did not modify the maximum number of features allowed by default during pharmacophore generation (*i.e.* 5). All these restrictions resulted in a new HypoGen™ run with better configuration costs and the top ten resulting hypotheses are shown in Table 1.

Evaluation of the cost of each hypothesis

During a HypoGen™ run a large number of candidate hypotheses were generated, some of which were considered and others discarded. Catalyst™ uses the *cost* of each hypothesis to decide whether to discard or to accept it (where the cost of a hypothesis is a measure of the number of bits required to describe it completely and the lower the cost, the better the hypothesis). The overall cost of a hypothesis is calculated by summing three cost factors [*i.e.* the **w**eight cost (***W***), the **e**rror cost (***E***) and the **c**onfiguration cost (***C***)] using the expression $Cost = eE + wW + cC$ (where *e*, *w* and *c* are the coefficients associated with each factor, the default value of which is 1.0). The weight cost is a value that increases in a Gaussian form as the feature weight in a model deviates from an idealized value of 2.0 and it is the main contribution to establishing the final hypothesis cost. This cost factor is designed to favour hypotheses in which the feature weights are close to 2.0. On the other hand, the error cost is a value that represents the RMS (*i.e.* root-mean square) differences between the estimated and measured activities for the training set. This cost factor is designed to favour those models with a high correlation between estimated and measured activities (*i.e.* with low RMS values). Finally, the configuration cost is a fixed value (*i.e.* constant among all the pharmacophore candidates built during the same HypoGen™ run; see Table 1) that is calculated early in the run and which can be found in the file with the extension .full. This file is located in the directory created by HypoGen™ inside the run directory. The configuration cost corresponds to the exponent of a base 2 potency that is equal to the number of hypothesis candidates that

ISBN:978-84-691-1553-4 /DL: T.151-2008

HypoGen™ will attempt to optimize during the run. If this number is less than 18, all the remaining pharmacophore models will be thoroughly analyzed. In contrast, if this value is equal to or higher than 18, not all of the data will be considered when the hypotheses are optimized so some candidates that have *survived* the subtractive phase will be left out of the optimization phase. In these cases, there are more degrees of freedom in the training set than HypoGen™ can properly deal with, so the hypotheses resulting from this last phase may be only a part of the some of those that would have been obtained. Configuration costs are equal to or higher than 18 when the training set is too complex (*e.g.* due to the molecules in this set being too flexible) and, in these situations, it is better to abort the HypoGen™ run, redefine the composition of the training set used for generating the hypotheses and start the HypoGen™ execution again. Therefore, the configuration cost depends on the complexity of the hypothesis space being optimized for a given training set and, therefore, the higher it is, the more computational resources are required for the calculation. It is also worth mentioning that the configuration cost is often referred to as the *entropy* of the hypothesis space.

As well as the total cost of each hypothesis, at the beginning of an automated hypothesis generation Catalyst™ calculates the cost of two theoretical hypotheses. The first is the ideal (or fixed) hypothesis in which the error cost component is minimal because all the compounds fall along a line of slope 1. The second is the the null hypothesis in which the error cost is high because all the compounds fall along a line of slope 0. These pharmacophores are considered to be the upper and lower limits for a pharmacophore candidate derived from the training set. Therefore, the cost values for these two theoretical hypotheses are useful guides for an early estimation of the probability of a successful experiment because they are available around 15 minutes from the start of the run and therefore help to decide if it is interesting for the current run to proceed or not. The fixed hypothesis cost (*i.e.* the fixed cost) tends to have a number of bits in the 70-100 interval and this is reported in the file with the .full extension (near the value for the configuration cost). This value corresponds to the lowest possible cost of the simplest hypothetical model that fits all the data perfectly and it is calculated by adding the values from: (a) the minimum achievable error; (b) the weight cost; and (c) the constant configuration cost. On the other hand, the null hypothesis cost (*i.e.* the null cost) is reported in the file with the .log extension (located in the same place as the .full file) and it is usually higher than the fixed cost. It represents the maximum cost of a pharmacophore with no features and with an activity that corresponds to the average of the activities from training set molecules. The difference between the fixed and the null hyphotesis costs is very important (the greater this difference is, the higher the probability of finding useful hypotheses). Nevertheless, in terms of hypothesis significance, what really matters is the magnitude of the difference between the cost of the null hypothesis and the cost of a returned hypothesis. In general, if this difference is greater than 60 bits, there is

an excellent chance that the pharmacophore model represents a true correlation. Since most returned hypotheses will be higher in cost than the fixed cost model, the difference between fixed and null costs must be at least 70 if the 60 bit difference is to be achieved. If a returned hypothesis has a cost that differs from the null hypothesis by 40-60 bits, it has a 75-90% chance of representing a true correlation in the data. When the difference is less than 40 bits, the likelihood of the hypothesis representing a true correlation in the data rapidly drops below 50%. Under these conditions, it may be difficult to find a pharmacophore model that can be shown to be predictive. In the extreme situation in which the differences between fixed and null cost are small (<20), there is little chance of succeeding and it is advisable to consider a new composition of the training set before proceeding.

In our study, the fixed cost is 88.116 (which is a good fit between the data of training set) whereas the null hypothesis cost (which indicates the maximum occurring error cost) has a value of 156.603. The difference between the cost of the null hypothesis and the total cost of the best hypothesis is between the 40-60 interval which indicates a 75-90% chance of representing a true correlation in the data. Table 1 summarizes the information about the top ten hypotheses obtained.

Generation of new hypotheses with HypoRefine™

HypoGen™ assumes that ligand activity increases with the number of hypothesis features present in the molecule. However, it must also be considered that although an inactive ligand can contain all the hypothesis features, a steric bulk may interfere in the intermolecular interaction with the target. In this situation, HypoGen™ would over predict the activity of the inactive ligand. HypoRefine™ is another Catalyst™ module that also generates pharmacophore models but, unlike HypoGen™, it also considers the possibility of steric clashes between the ligand and the target. Therefore, additional pharmacophore candidates were also generated with HypoRefine™ and submitted to the same validation process as the ones directly obtained from HypoGen™ (see below in the **Pharmacophore validation** part).

With HypoRefine™, exclusion volumes (*i.e.* excluded volume spheres) can be automatically placed at strategic locations around the previously generated HypoGen™ hypotheses (therefore, it is very important that the HypoRefine™ run is done in the same conditions as the previous HypoGen™ run so that the resulting hypotheses are comparable). The location of these volumes is inferred from the location of atoms of well-pharmacophore fitting but inactive compounds, which describe "forbidden zones" where the molecules cannot fit. In this respect, once the training set compounds have been classified as active or inactive (either by the user or by the default HypoRefine™ criteria), excluded volume spheres are automatically identified in four

ISBN:978-84-691-1553-4 /DL: T.151-2008

consecutive steps: (1) alignment of active molecules; (2) alignment of inactive molecules; (3) identification of atoms in the aligned inactive compounds that are far away from those in the aligned active compounds; and (4) random selection from the points where these atoms are located and assignment as excluded volume spheres. The constructive phase of HypoRefine™ is identical to the phase in HypoGen™ whereas the subtractive phase is not performed. So, excluded volume spheres are included in the simulated annealing optimization process. Finally, pharmacophore models that also fit inactive molecules are automatically penalized (*i.e.* their total cost is increased). If the hypothesis models reported by HypoRefine™ have lower costs than the ones obtained with a standard HypoGen™ run, it is concluded that the differences in activities between the training set molecules are better explained if excluded volumes are considered as well as feature mapping.

Our HypoRefine™ pharmacophore models were obtained in the same experimental conditions as the ones described above for HypoGen™. The training set molecules that were chosen as active or inactive in order to define the location of the excluded volumes. The information about the top ten hypotheses obtained is summarized in Table 1.

Pharmacophore mapping

Active compounds contain common hypothesis features that are not found in inactive compounds and HypoGen™ can identify them within conformationally allowable regions. These functional groups are represented with spheres that indicate where they are located. The centres of these spheres indicate the preferred location of the functional group (where the centre of a base, an acid, a hydrogen bond donor or acceptor is usually defined as the position of an atom but, for a hydrophobic region or aromatic ring, the centre is defined as the centroid of the group). Their radii indicate some kind of tolerance in their location relative to the ideal place. Moreover, for a more accurate description of some functions (*e.g.* hydrogen bond donors and acceptors), a vector shows the direction of the interaction. In this respect, a vector representation is more accurate than a point representation since it imposes an additional constraint of bond directionality between complementary features between the ligand feature and the receptor.

Pharmacophore validation

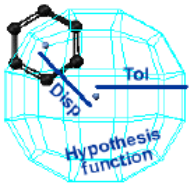
When the various hypothesis generation processes had finished, the resulting pharmacophore candidates (see Table 1) were analyzed in order to select the one with the best predictive power for further use in IC50 prediction. There are several selection methods and we used them all to validate the hypothesis candidates reported by HypoGen™ and HypoRefine™. One strategy was to statistically evaluate each hyphotesis by: (a) predicting the training set activities; (b) performing a randomization test with a Catalyst™ utility called catScramble™; and (c)

analyzing the cost parameters (as has already been described). Another strategy would be to use a set of compounds with known experimental IC50 values and then test the predictive power of each hyphotesis. These compounds must be different from the ones in the training set and, ideally, their IC50 values have to be measured under the same experimental conditions and by the same scientific team as the ones used in the training set. In the current experiment, the 15 molecules of the initial 36 that were not included in the training set were used in this way (*i.e.* as the test set).

Training set activity prediction

The activity of the compounds in the training set was estimated using the regression parameters. The use of regression parameters is a hypothesis validation method that evaluates the capacity of the self hypothesis for correctly predicting the activity (IC50 values in this case) of each training set compound. In order to calculate these parameters for one specific pharmacophore, a term called *geometric fit* has to be calculated for each molecule in the set. To do so, Catalyst™ evaluates whether each functional group in the analyzed hyphotesis is present or not in the corresponding molecule. When this analysis has finished, Catalyst™ calculates its geometric fit with the next expression (where the more external summatory only accounts for those pharmacophore functions that are also present in the molecule and the internal one applies only when the same function is found more than once in the pharmacophore and in the analyzed molecule):

$$Fit = \sum_{\substack{mapped\\ hypo\\ functions}} w\left(1 - \sum_{spheres}\left(\frac{Disp}{Tol}\right)^2\right)$$



Where: (a) **Disp** is the distance term that measures the separation between the functional group in the molecule and its centroid in the pharmacophore; (b) **Tol** is the radius of the corresponding function in the hypothesis (*i.e.* a tolerance measure of the location of the pharmacophore function); and (c) **w** is a factor that weights the contribution of the function to the global activity.

The resulting geometric fit value for each training set molecule is plotted against the negative logarithm of its experimental activity and then the resulting points are submited to a linear regression analysis that provides a line that is used to predict the activity for each training compound (where the resulting r value for the linear regression corresponds to the correlation value of the corresponding hypothesis). Finally, the quality of the prediction for the training set is reported with the RMS value that represents the deviation of the logarithm of estimated

ISBN:978-84-691-1553-4 /DL: T.151-2008

activities (y_i) from the logarithm of the measured activities (x_i) normalized by the logarithm of uncertainties (Unc_i). This RMS is calculated by the following expression:

$$RMS = \left(\frac{1}{n} \sum_{i=1}^{n} \frac{(X_i - Y_i)^2}{\log(Unc_i)} \right)^{1/2}$$

Table 1 shows the RMS and the correlation values for each hypothesis.

For each molecule in the training set, Catalyst™ also computes the difference between its experimental activity and hypothesis-estimated activity. This is known as the *error value*. The error value is computed as the ratio of the experimental IC50 relative to the estimated IC50 (or the inverse if the latter is greater than the former). An error with a negative sign indicates that the experimental IC50 is higher than the corresponding estimation. Table 2 shows the experimental IC50, the estimated IC50 and the error values for all the molecules in the training set in accordance with the best hypotheses described in Table 1 [*i.e* the first hypothesis from the HypoGen™ run (hereafter Hypo1) and the first three hypotheses from the HypoRefine™ run (hereafter Hypo2, Hypo3 and Hypo4)]. The graph of estimated against experimental activities for the training set is shown in Figure 2.

Randomization test

The aim of this test is to evaluate the significance of the hypothesis analyzed in order to check whether there is a strong correlation between the chemical structures of the training set and its biological activity. In this test, Catalyst™ statistically validates a pharmacophore by performing a randomization with a utility called catScramble™, which is based on Fisher's randomization test. To do so, catScramble™ randomly reassigns the activity values to the training set compounds. Each random reassignment generates a new spreadsheet in which the active molecules can become inactive or have an intermediate level of activity, and the originally inactive molecules can become active. The number of these new spreadsheets will depend on the level of statistical significance that is desired. For instance, to achieve a 95% confidence level, 19 random spreadsheets (or 19 HypoGen™ runs) have to be generated [whereas for a 98 or 99% confidence level, 49 or 99 random spreadsheets (or HypoGen™ runs) have to be generated, respectively]. Then, a HypoGen™ process is performed with each randomized spreadsheet in experimental conditions (*i.e.* features, parameters) that are identical to those used in the original HypoGen™ run. After the new hypotheses had been generated, the statistical significance of the original HypoGen™ run was calculated with the expression:

$$Significance = (1-(1+n)/N)*100$$

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

where: (a) **n** is the total number of new hypotheses with a total cost lower than that of the hypotheses obtained by the original HypoGen™ run; and (b) **N** is the total number of HypoGen™ runs (initial + random runs). Therefore, if the randomized data set generates a large number of pharmacophores with similar or better cost values, RMS and correlation, then the hypotheses generated by the original HypoGen™ run are considered to be generated by chance and, therefore, not reliable (*i.e.* they have low significance).

We decided to do a randomization test with a confidence level of 95% which required the generation of 19 random spreadsheets. Figure 3 and Table 3 show the differences in costs between the HypoGen™ and HypoRefine™ runs and the scrambled runs. As can be seen, none of the random-generated hypotheses have a total cost that is lower than the cost of the original HypoGen™ and HypoRefine™ runs. Therefore, we concluded that there is, at least, a 95% probability that the best ten hypotheses generated by the original HypoGen™ and HypoRefine™ runs report a true correlation between the structural and biological data for the training set.

Test set activity prediction

The predictive capacity of a hypothesis must be determined by estimating the activities of some molecules not included in the training set (*i.e.* the so called *test set*). Fifteen molecules (see Figure 1) were not used to generate the pharmacophore and their corresponding IC50 data were obtained under the same experimental conditions as the ones in the training set. Therefore, they were used as a test set to help us to select the best hyphotesis of all the ones that had been identified as correct by the above statistical analyses. Thus, Catalyst™ uses each of the statistically correct pharmacophore candidates (see Table 2) to predict the activity values for the test set molecules. Then their predicted IC50 values were correlated with the experimental ones (see Table 4). Finally, the pharmacophore with the best predictive power was Hypo3 (see Figure 4) because only two molecules in the test set had an error value above ten and it shows a higher correlation with real and predicted IC50 values than Hypo1 (see Table 4). Therefore, it was used to fit and estimate the activity of compounds that were not in the training or test sets but which shared the p110α binding site with them (*e.g.* the naturally-ocurring phenolic compounds in vegetable extracts and their bioactive forms; see Figure 5 and Tables 5 and 6). For some of them, the experimental IC50 value has been reported in the bibliography (although experimental conditions other than those of the training and test sets) but for others it is unknown.

ISBN:978-84-691-1553-4 /DL: T.151-2008

p110α structure used in the study

No structure is available in the PDB for the catalytic subunit of the p110α isoform. So a homology model, built from its human sequence (P42336 accession number in Swiss-Prot), was downloaded from ModBase [29] (http://modbase.compbio.ucsf.edu/). According to this database information, this model was built by using a *Sus scrofa* p110γ structure as template. It was deposited in the PDB with the code 1E7V [17] and shows a 37% sequence identity with P42336. Moreover, this p110α homology model has coordinates for the sequence segment that corresponds to residues from 106 to 1062 and when it is superposed on the template structure used to model it, the resulting RMSD is 1.11 Å. Before it was used in docking experiments, the p110α model was subjected to energy minimization with Swiss-PdbViewer/DeepView (http://www.expasy.org/spdbv/) [30]. The coordinates for the resulting minimized p110α structure are available from the paper's website.

Docking studies

Docking studies were done with the eHiTS® v6.1 software package (SimBioSys Inc., Toronto, Canada; http://www.simbiosys.ca/ehits/). The p110α region around which possible ligand binding has been studied is defined by the smallest box that can contain all the ligand atoms from the 1E7V structure (*i.e.* LY294002) plus 15 Å in each dimension (*-margin 15* option in eHiTS®). Before any docking experiments, then, the p110α homology model downloaded from ModBase and further minimized with Swiss-PdbViewer/DeepView was structurally superposed on 1E7V in order to obtain the LY294002 coordinates in the same coordinate system as the p110α receptor. The superposition was done with Swiss-PdbViewer/DeepView. The resulting LY294002 coordinates were then used by eHiTS® to focus the docking on the selected p110α area (*-clip* option in eHiTS®) and can be found on the paper's website. For the rest of the eHiTS® parameters and options, default values were used.

When an eHiTS® run is finishes, the program reports the top 32 conformations (energy optimized) of the docked ligand that have the lowest (*i.e.* most negative) interaction energy with the p110α area under study. All eHiTS® results can be downloaded from the supplementary materials website.

RESULTS AND DISCUSSION

Pharmacophores generated by HypoGen™ and HypoRefine™

Ten hypotheses were initially obtained by using the training molecules in Figure 1 and the HypoGen™ module from Catalyst™. During the HypoGen™ run, the cost parameters (*i.e.* the costs of configuration, fixed hypothesis, null hypothesis and self hypotheses), the rms deviation and the correlation coefficient were calculated and further used to evaluate the quality of the pharmacophore candidates. The values of these parameters are listed, together with the pharmacophore features mapped by each hypothesis, in Table 1.

Table 1 shows that seven out of ten pharmacophore candidates (*i.e.* all except hypotheses 6, 7 and 9) have four chemical features (one HA, two HY and one RA functions) whereas the remaining three had only three functions (one HA, one HY and one RA function). The configuration cost (also known as entropy) of a good pharmacophore must be below 17 bits and Table 1 shows that it was 16.355. Hence, this entropy value indicates that the training set, the restrictions in terms of the feature types (*i.e.* HA, HD, HY and RA) and the minimum number of features required (one for HA, HY and RA) to derive the hypotheses were suitable. HypoGen™ also performs: (1) a null hypothesis calculation (related to a null cost) which presumes that there is no relationship in the dataset and that the experimental activities are normally distributed around their average value; and (2) a fixed hypothesis calculation (related to a fixed cost) which represents the simplest model that fits all data perfectly. Therefore, a meaningful pharmacophore hypothesis may result when the difference between these null and fixed cost values is large (*i.e.* the greater the difference, the higher the probability of finding useful pharmacophores). In our HypoGen™ run, the null cost value of the top 10 hypotheses was 156.603 and the fixed cost value was 88.116 (with a difference between null and fixed cost of 68.487 bits). On the other hand, the magnitude of the difference between the cost of the null hypothesis and the cost of any returned hypothesis is also very important in terms of hypothesis significance. Thus, the total cost of a good hypothesis is expected to be close to the fixed cost of the fixed hyphotesis. Table 1 reports the differences between null and total cost and all the hypotheses differ from the null hypothesis by 40-60 bits. It should be pointed out that the first one (so called Hypo1) shows a difference close to 60 bits which means there is a 90% chance of it representing a true correlation in the data. Table 1 also reports the quality of the linear regression derived from the geometric fit index [*i.e.* correlation coefficient (r)] and the quality of the correlation between the estimated and the experimental activity data (*i.e.* the RMS). The former parameter value for Hypo1 was higher than 0.9 (*i.e.* 0.950), which shows a good correlation (because 1.0 would be a perfect correlation). The graph of estimated against experimental activities for the training set is shown in Figure 2. The RMS value of 0.897 for

Hypo1 also showed a good quality of prediction for the training set (it is the lowest RMS divergence achieved in the HypoGen™ run; see Table 1). To complete this information, Table 2 shows the validation done by using Hypo1 to predict the activity for each compound of the training set. As well as showing the estimated IC50 values, we classified the IC50 values on an activity scale with the following criteria: (a) "++++" for highly active molecules (*i.e.* IC50 < 0.1 μM); (b) "+++" for active compounds (*i.e.* 0.1 μM ≤ IC50 < 1 μM); (c) "++" for moderately active molecules (*i.e.* 1 μM ≤ IC50 < 10 μM); (d) and "+" for inactive compounds (*i.e.* IC50 >10 μM). With this activity scale, we realized that Hypo1 was able to predict most of the compounds. In this respect, it is worth mentioning that only 5 out of 21 molecules in the training set were incorrecly classified on this activity scale (but in all 5 cases, the predicted IC50 values were in a group on the activity scale immediately after or before the one in which experimental IC50 are classified; see Table 2). For instance, one inactive compound (*i.e.* 5d) was predicted as moderately active. One of the active compounds (*i.e.* 6k) was predicted as moderately active and another one (*i.e.* 8) was predicted as highly active. Moreover, it should also be said that the hypothesis predicted no highly active or active compounds as inactive compounds, and *vice versa*. The analysis of the error value is also reported in Table 2 for Hypo1 (*i.e.* Error column) and showed that all 21 compounds in the training set had errors below 10, which means that the predicted activity of these compounds is between 10 times greater than and 1/10 of the experimental activity (where a negative ratio means that the experimental IC50 value is higher than the estimated one). All this, then, confirmed that Hypo1 was a reliable pharmacophore candidate for describing the structure activity relationship (SAR) in the training set. Thus, we concluded that Hypo1 was the most statistically significant hypothesis generated by HypoGen™ and the remaining nine hypotheses were discarded for further use.

After the HypoGen™ run, we used the HypoRefine™ module to evaluate whether the steric clashes in the intermolecular interaction between ligand and receptor can affect the biological activity. Using the same training set of 21 compounds (see Figure 1) and the same conditions as in the previously described HypoGen™ run, another set of 10 hypothesis was generated. The HypoRefine™ module also calculated the cost parameters, the rms deviation and the correlation coefficient in order to analyze the statistical significance of these 10 hypotheses. The value for these parameters and the corresponding pharmacophore features are listed in Table 1. Thus, Table 1 shows that nine out of the ten HypoRefine™ hypotheses consisted of four features (*i.e.* one HA, two HY and one RA feature for hypotheses 1 to 6; and two HA, one HY and one RA feature for hypotheses 7 to 9). Only the last hypothesis (*i.e.* hypothesis number 10) shows three chemical functions (one HA, one HY and one RA). Besides the chemical functions, one excluded volume was included in some hypotheses by the HypoRefine™ run (Table 1 shows which hypotheses contain one excluded volume). Relative to the parameter values for the

HypoRefine™ run, the configuration cost was correct (*i.e.* 16.388) and the difference between the null and total cost of Hypo2, Hypo3 and Hypo4 (the best three hypotheses) was 60.296, 60.118 and 59.068 bits, respectively. Therefore, these three hypotheses showed a 90% chance of representing a true correlation in the data. In terms of correlation coefficient (*i.e.* r) and RMS, the quality of Hypo2, Hypo3 and Hypo4 (see Table 1) was confirmed. For these three hypotheses, the r values were higher than 0.9 and their RMS were the lowest observed in the ten HypoRefine™ derived hypotheses. To complete this information, Table 2 shows the validation done by using Hypo2, Hypo3 and Hypo4 to predict the activity for each compound of the training set. Thus, the scale activity column shows that, in general, these hypotheses were able to correctly predict the most active and inactive compounds. In only a few situations: (a) inactive compounds were predicted as moderately active (*i.e.* 5b in Hypo4) or *vice versa* (*i.e.* 10f in Hypo2 and Hypo4); (b) some active molecules were predicted as moderately active (*i.e.* 6k in Hypo2, Hypo3 and Hypo4 and ly294002 in Hypo4) and *vice versa* (*i.e.* 10g in Hypo3); and (c) some active compounds were predicted as highly active (*i.e.* 8 in Hypo2, Hypo3 and Hypo4 and 6h in Hypo2 and Hypo4) or *vice versa* (*i.e.* 11 in Hypo3). It should be pointed out that none of the hypotheses predicted any highly active or active compounds to be inactive compounds or *vice versa*. The error values (see Table 2) for all 21 compounds in the training set were less than 10 when they were computed for Hypo2, Hypo3 and Hypo4, which means that the activity prediction of these compounds is between 10 times greater than and 1/10 of the actual activity. Therefore, the IC50 activity values of all compounds are considered to have been correctly predict. Taking into account all this, we concluded that Hypo2, Hypo3 and Hypo4 were the most statistically significant hypotheses generated by HypoRefine™ and the remaining seven hypotheses were not considered any further in the prediction evaluation.

Therefore, the results confirmed that Hypo1, Hypo2, Hypo3 and Hypo4 were reliable pharmacophore candidates for describing the structure activity relationship (SAR) in the training set. However, they also have to be able to explain the activity values of our test set.

Validation with the randomization test

Another approach for validating the quality of HypoGen™ and HypoRefine™ hypotheses was to use the catScramble™ program to apply a cross validation. In this validation test, we selected a 95% confidence level, so 19 random assignments were made of the activity values among the training set compounds for further use with HypoGen™ and HypoRefine™ (see Table 3). The data of this cross validation clearly indicated that all values generated after randomization produced hypotheses with no predictive values (see Table 3). Thus, the results of the HypoGen™ runs done with the randomized data showed that: (a) of the 19 runs, no hypothesis had a correlation higher than 0.8; (b) the RMS values of the best hypothesis in each run were

higher than the ones in Hypo1; and (c) the total cost of the best hypothesis in each run was closer to the null cost (*i.e.* 156.603), which is not desirable for a good hypothesis. Results were similar obtained when the randomized data was used by HypoRefine™. Thus, of the 19 runs, only two had a correlation higher than 0.8 (*i.e.* runs 3 and 5 with correlations of 0.820 and 0.837, respectively). Relative to the RMS values, these were higher than the ones in Hypo2, Hypo3 and Hypo4. Finally, the total cost of the best hypothesis from each run was close to the null cost (*i.e.* 156.603), which is not desirable for a good hypothesis. Figure 3A shows the differences in costs between the ten best hypotheses generated by the original HypoGen™ run and the ten best hypotheses obtained by the 19 HypoGen™ runs done with the randomized data. Figure 3B is equivalent to Figure 3A but compares HypoRefine™ not HypoGen™ runs.

Pharmacophore validation with the test set

Two validation procedures (*i.e.* the analysis of parameters and randomization test) provided strong statistical confidence on the Hypo1, Hypo2, Hypo3 and Hypo4 pharmacophore candidates. Nevertheless, this validation process does not reveal anything about the predictive capacity of these hypotheses. Therefore, we evaluated their predictive capacity by estimating the activities of some compounds outside the training set (*i.e.* the so called *test set*; see Figure 1), which are structurally distinct molecules (*i.e.* structurally diverse) from those used to generate the pharmacophore candidates but whose activity values had been measured in the same experimental conditions. The IC50 activity values of the test set compounds range from 0.0028 to 30 μM (see Table 4) and were classified following the same criteria as the training set [where "++++" indicates highly active (IC50 < 0.1 μM), "+++" indicates active (0.1 μM ≤ IC50 < 1 μM), "++" indicates moderately active (1 μM ≤ IC50 < 10 μM); and "+" indicates inactive (IC50 >10 μM)]. Analyzing the activities predicted by Hypo1, Hypo2, Hypo3 and Hypo4 (see Table 4), we can observe that no highly active or active compounds were predicted as inactive or *vice versa*. When the predictive power of the four hypotheses is compared, Hypo3 is the hypothesis that gave the best prediction and had the fewest differences in the activity scale between the predicted and the experimental IC50 data (see Table 4). These differences were not found in compounds at the extremes of the activity range (the most and the least active compounds). Moreover, the correlation coefficients performed by Hypo1, Hypo2, Hypo3 and Hypo4 were 0.839, 0.790, 0.851 and 0.748, respectively, which show a good correlation between the experimental and estimated activities. The calculated Error value for each test set compound is also reported in Table 4. It shows that when Hypo1 or Hypo3 are used, 13 out of 15 compounds in the test set have Error values that are lower than 10, which means that the activity prediction of these compounds is between 10 times greater than and 1/10 of the experimental activity. Moreover, the two molecules that both Hypo1 and Hypo3 fail to predict are the same (*i.e.* 2d and 6b). In contrast, Hypo2 and Hypo4 predict the IC50 activity values of

four and five compounds, respectively, with an Error value higher than 10. Therefore, it can be concluded that the best predictive hypothesis candidates are Hypo1 and Hypo3. In this respect, the estimated inhibitory activity in both molecules is higher than the experimental one, which means that the inhibitory activity is over-predicted by Hypo1 and Hypo3. The 6b compound has an additional methyl group relative to 6a (methoxy for 6b and hydroxyl for 6a in R2 position; see Figure 1), which could conflict slightly with the p110α receptor and prevent the interaction with the receptor. In the same way, the 2d compound has one NO2 substituent in the phenyl group (see Figure 1), which might also conflict with the receptor. Therefore, after the validation with the test set, we confirm that only Hypo1 and Hypo3 are statistically right and predictive.

Pharmacophore candidate selection

Comparing Hypo1 (obtained with standard HypoGen™ run) with Hypo3 (obtained by running HypoRefine™), we realized that both hypotheses have similar total cost values (*i.e.* 97.357 and 96.485) and neither RMS (*i.e.* 0.897 and 0.856) nor correlation values (*i.e.* 0.950 and 0.955) show significant differences (see Table 1). Furthermore, when we compared the pharmacophore features, we realized that the two hypotheses explained the SAR with one HA, two HY and one RA and similar locations of these functions in the 3D space (results not shown). Despite the similarity between both pharmacophore candidates, the best one is Hypo3 because it has: (a) the highest cost difference with the null cost (*i.e.* 60.118 vs 59.246; see Table 1); (b) the best correlation coefficient (*i.e.* 0.955 vs 0.950; see Table 1); (c) the lowest RMS divergence (*i.e.* 0.856 vs 0.897; see Table 1); and (d) the values of the Error were lower than for Hypo1 (see Table 2). Although Hypo3 was generated with HypoRefine™, this hypothesis does not include any excluded volumes (see Table 1). Therefore, we can conclude that Catalyst™ is able to explain the differences in activities between the molecules of our dataset simply by using the feature mapping and, in the present study, the excluded volumes do not help us to discriminate between those molecules that inhibit p110α and those that do not. Therefore, the validation study suggests that Hypo3 can map a structurally diverse group of compounds quite effectively and makes us confident that it can be used to identify new p110α inhibitors.

Analysis of the pharmacophore for p110α (*i.e.* Hypo3)

Hypo3 is formed by: (a) one hydrogen bond acceptor; (b) two hydrophobic groups; and (c) one aromatic ring [see Figure 4 where the features are represented with spheres and color-coded with green for hydrogen-bond acceptor (HA), light-blue for hydrophobic functions (HY) and orange for the aromatic ring function (RA)]. The hydrogen bond acceptor and aromatic ring include a vector that shows the direction of the interaction. The figure also shows the interval of distances between the different pharmacophore functions where **a**, **b**, **c**, **d**, **e** and **f** correspond to [2.993-4.993], [5.621-7.621], [2.355-4.355], [7.058-9.058], [6.978-8.978] and [7.291-9.291] Å,

ISBN:978-84-691-1553-4 /DL: T.151-2008

respectively. The most active compounds map all the features (see Table 2) and the fit between the pharmacophore and the compounds is higher than for the active or moderately active compounds. Most of the inactive compounds map three out of four functions in the pharmacophore (see Table 2). Nevertheless, it is specially significant that three out of four of these molecules map only one hydrophobic function whereas the most active compounds map the two that are present in the pharmacophore (thus, suggesting that the intermolecular interaction of the hydrophobic function has a very important role). Therefore, the ability to fit both or just one of the hydrophobic functions seems to correlate well to predict whether a compound is an active p110α inhibitor or not.

IC50 prediction of phenolic compounds

We used eHiTS® to dock the most frequent phenolic compounds in plant extracts and their bioactive forms (see Figure 5) in a homology model from p110α that was downloaded from ModBase (where the ligand binding site used during docking was around the area where the phenolic compounds quercetin and myricetin are found to be in their crystallized complexes with p110γ). Therefore, it is important to point out that the ligand conformations obtained from the docking experiment were generated inside the protein active site (*i.e.* in allowed regions within the receptor). The phenolic compounds studied belong to two different classes: flavonoids and non-flavonoids [31, 32]. The non-flavonoids tested can be further divided into two different groups: phenolic acids and stilbenes (see Figure 5). On the other hand, flavonoid compounds are grouped into six structurally diverse families [*i.e.* flavonols, flavones, flavanones, anthocyanidins, flavanols (monomers and oligomers or procyanidins) and isoflavones; see Figure 5]. The docking results show that some phenolic compounds (*i.e.* tetramers of the procyanidins) are not able to dock into the binding site of p110 α because their molecular structures are too big for the ligand binding site (see Figure 5 and Table 5).

Taking into account that the chemical functions of Hypo3 describe the most important kind of interactions between the receptor (*i.e.* p110α) and the ligands (*i.e.* p110α inhibitors), we decided to use it to estimate the IC50 activities for the phenolic compounds in Figure 5, assuming that they also interact with p110α and inhibit it by using the same mechanism as the molecules in the training and test sets. Estimating the corresponding IC50 values required the different conformations for all the phenolic compounds in Figure 5 to be generated. Therefore, we used either the different poses obtained by the docking process described above or the conformations generated with the BEST conformational search algorithm from Catalyst™ [28]. Finally, the IC50 values were estimated by mapping the resulting conformations (either from the docking results or from the BEST conformational search) on Hypo3 by means of the "Fast fit" Catalyst™ option. Then, the lowest IC50 value (*i.e.* the highest inhibitory activity) obtained in

the two mappings was reported (see Tables 5 and 6). The resulting IC50 values span 6 orders of magnitude (from 0.0043 to 210 μM) (see Tables 5 and 6) and they have been classified on the same activity scale as the molecules in the training and the test sets. In general, there is a high level of agreement between the predicted IC50 values obtained by the two conformational search methods. In this respect, significant discrepancies are only restricted to (+)-epicatechin, (-)-catechin, procyanidin trimers and tetramers and tannic acid (see Tables 5 and 6). Thus, when these discrepancies occur, we consider that the IC50 results derived from conformations obtained by docking are more reliable than the ones that have been obtained without considering any steric hindrance caused by the p110α receptor (such as when the conformations obtained from the BEST algorithm are used). For instance, procyanidin trimers and tannic acid are big molecules with a high degree of flexibility (see Figure 5) that are forced to adopt poses in the ATP-ligand binding site of p110α that are very different from the ones obtained when the binding-site structure is not considered. In the case of procyanidin tetramers (molecules that are larger than procyanidin trimers or tannic acid; see Figure 5), the docking algorithm cannot find a ligand pose that is compatible with the binding-site geometry (the ligand is too big for the size of the binding-site cavity). In contrast, the BEST algorithm suggests different conformations for the procyanidin tetramers that the docking shows not to be compatible with the ligand-binding site dimensions. This, then, seems to explain the large differences between the IC50 values that are predicted by the two conformational search methods for tannic acid and procyanidin trimers and tetramers.

Before describing our predicted IC50 values on p110α, we should point out that Gamet-Payrastre and colleagues studied how the activity of PI3K isolated from human blood platelets and immunoprecipitated with anti-p85α antibody is inhibited by fourteen naturally-occurring flavonoids of different chemical classes [33, 34]. Nevertheless, isoform-selective PI3K inhibitors were not available when they carried out their study so they could not assess the relative amount of the total activity that was attributable to a particular class IA isoform (in contrast to similar experiments that are done at present [3]). Therefore, the IC50 values reported by Gamet-Payrastre and colleagues must be seen as the effect of the assayed flavonoids on a *pool* that contains the different class IA PI3K isoforms and not as result of their inhibitory power on p110α. In consequence, their results cannot be compared with the values reported in the present study because it is well known that the same inhibitor can provide very different IC50 values when assayed with different class IA PI3K isoforms [4, 26, 27, 35-38].

Prediction of non-flavonoid activity

The IC50 values predicted for stilbenes are lower than for phenolic acids (see Table 5 and Table 6). Thus, phenolic acids, which are the smallest phenolic compounds (see Figure 5), are

predicted as inactive p110α inhibitors with very high IC50 activity values (*i.e.* between 11 and 210 μM; see Tables 5 and 6). In agreement with these results, a previous study showed that gallic acid does not inhibit methyl glucose uptake [16]. In contrast, although stilbenes have a small structure (see Figure 5), they are predicted to be active inhibitors of p110α. In this respect, the presence of two aromatic rings in the stilbene structure allows it to fit onto the pharmacophore better than the phenolic acids. Therefore, the inhibitory activity of stilbenes is higher. Thus, just as we found with resveratrol (IC50 = 0.46 μM; see Table 5), a recent study has also found that this stilbene is a competitive inhibitor that targets the ATP binding pocket of p110α and p110β [14].

Prediction of flavonol activity

The results in Table 5 and Table 6 show that most flavonols were predicted as active inhibitors of p110α. At this point, it should be pointed out that previous experimental studies have shown that quercetin inhibits glucose uptake [16] and *in silico* experiments have suggested that this inhibition occurs as a consequence of the formation of a complex between quercetin and GLUT4 [16]. In contrast, our results suggest that the inhibition of the glucose uptake is probably a consequence of the quercetin binding to the ATP binding site of p110α (this is strongly supported by the recent finding that confirms that inhibitors targeting p110α isoform block the acute effect of insulin *in vivo* [4] and by the fact that quercetin can bind to the ATP-binding site of the p110α-related p110γ isoform [17]). Isorhamnetin (*i.e.* 3-*O*-methylquercetin), a bioactive molecule, is the only flavonol that is predicted to be a highly active p110α inhibitor (see Table 6). Its structure (see Figure 5) has a methoxy group whose methyl moiety can fit in the second hydrophobic function of the pharmacophore (results not shown) and so, map all its functions (which therefore produces better interaction with the receptor). On the other hand, the flavonol myricetin is predicted to be an inactive (*i.e.* low activity) inhibitor within both docking and Catalyst^{TM}-derived conformations (*i.e.* 110 and 160 μM, respectively; see Table 5). Experimental studies on how myricetin influences glucose uptake in adipocytes are contradictory. On the one hand, one study has reported that myricetin inhibits glucose uptake in isolated rat adipocytes [16]. On the other, another study with the same type of cells states that it increases the glucose transport of the same cells and mimics insulin stimulating lipogenesis without affecting the insulin receptor function or GLUT4 translocation [39]. The molecular structure of the myricetin is very similar (see Figure 5) to the structure of the quercetin but its estimated activities are very different (*i.e.* they differ by three orders of magnitude). In order to explain this difference, we analyzed how both molecules fit in the pharmacophore and we realized that myricetin cannot able to fit the two hydrophobic functions defined in the pharmacophore (see Figure 4). Thus, when myricetin is analyzed, it can be observed that it has three hydroxyl groups joined to phenyl ring B and the one in position R1 is not present in the

quercetin structure (see Figure 5). Therefore, this third hydroxyl group in the B aromatic ring of myricetin may be responsible for its low activity because it might hinder accessibility to the ring aromatic function of the pharmacophore. In fact, it has been shown that when myricetin forms a complex with the p110α-related p110γ isoform, its relative orientation in the binding site is flipped end-for-end relative to the orientation of quercetin in its complex with p110γ [17]. Therefore, the situation may also be similar when the target is p110α and this could explain the large differences in IC50 values that are predicted for these two highly similar ligands.

Prediction of flavone and flavanone activity

Tables 5 and 6 also show that flavones and flavanones (which have molecular structures very similar to flavonols; see Figure 5) are also predicted to be active inhibitors of p110α with IC50 values between 0.29 and 1 μM. They all fit in the pharmacophore by mapping the same functions (one hydrophobic function, one aromatic ring and one hydrogen acceptor; results not shown). To our knowledge, and with the exception of naringenin, there are no previous experimental studies on how these phenolic compounds inhibit the activity of p110α. In this respect, naringenin has been shown to inhibit insulin-stimulated glucose uptake in 3T3-L1 adipocytes in a dose-dependent manner through the inhibition of p110α [15], and our results confirm this experimental evidence (see Table 5).

Prediction of anthocyanidin activity

There is no experimental data about the inhibitory p110α activity of anthocyanidins and our results predict that they can be highly active, active or inactive depending on their molecular structure (see Tables 5 and 6). Interestingly, the bioactive structures of most anthocyanidins show that the structure of the naturally-occurring form from which they derive is determinant for their activity values, (*i.e.* the conjugation does not substantially modify the situation in the activity scale of the bioactive forms relative to the naturally-occurring form from which they derive; see Table 6). Thus, delphinidin and its conjugated forms were predicted to be inactive p110α inhibitors (*i.e.* IC50 ranges from 19 to 160 μM). This low inhibitory activity of delphinidin is the result of its molecular structure (see Figure 5), which does not allow a simultaneous fit of the two hydrophobic functions defined in the pharmacophore because delphinidin has an extra hydroxyl group joined to the phenyl ring B. This is not the case in other anthocyanidins such as cyanidin (see Figure 5 and Table 5). Therefore, this third hydroxyl group in the phenyl ring is responsible for the decrease in activity because it influences the hydrophobic surface of the ligand and, as has been mentioned above, fitting to the two hydrophobic functions of the pharmacophore seems to be an essential feature for distinguishing inactive p110α ligands from the rest (see Table 2). In contrast, the compound malvidin has two methoxy groups joined to phenyl ring B (see Figure 5) that allow its structure to simultaneously

fit in the two hydrophobic functions of the pharmacophore and achieve one of the smallest IC50 values predicted for the phenolic compounds (see Table 5). The conjugated form of malvidin that is detected in humans is malvidin 3-glucoside and, according to the general trends observed for the bioactive forms of anthocyanidins, it is also predicted to be a highly active p110α inhibitor (see Table 6).

Prediction of flavanol activity

Flavanols, a large class of flavonoids that are ubiquitous in plants [8, 40] and widely found in a number of foods [9, 41], are an integral part of the human diet and are considered to be key compounds in the relationship between health and diet. Previous studies have shown that catechin did not inhibit methylglucose uptake (although no information is provided about which catechin isomer is responsible for the observed effect) [16]. In this respect, our results show that the inhibitory activity of catechin and epicatechin on p110α depends on the isomer studied [*i.e.* although (+)-catechin with an IC50 value of 0.87 μM is predicted to be active, (-)-catechin with 42 μM is predicted to be inactive (and the opposite is true for the activities of the two epicatechin isomers; see Table 5)]. Therefore, knowing which flavanol isomer is used in a study is important because the resulting biological activity can be different. Thus, studies using 3T3-L1 cells demonstrated that catechin can inhibit or stimulate lipid accumulation depending on the isomer studied [42].

Our results for the dimeric flavanol structures and the tetramethylated derivate [43] show that these molecules can also inhibit p110α (see Tables 5 and 6). In the case of the trimeric structures, our results show strong discrepancies between the IC50 values derived from conformations obtained from docking and the ones obtained from the BEST conformational search algorithm in Catalyst™ (see Tables 5 and 6). Thus, although the conformations obtained by Catalyst™ are mapped and predicted to be active, the ones generated by docking the ligands into the ATP binding site of the p110α homology model cannot be mapped on the pharmacophore (*i.e.* they cannot be fitted on it). In our opinion, this discrepancy shows, that the ligand conformations that can be mapped in the pharmacophore (*i.e.* the ones obtained with the BEST algorithm) are impossible to achieve in the ligand-binding site of p110α. These results are also supported by preliminary results from our group (that will be published elsewhere) that suggest that trimeric flavanols are not p110α inhibitors because, in fact, they stimulate glucose uptake. Therefore, for the trimer procyanidins, we may have to rely on the results provided by docking-derived conformations, which show that these molecules cannot inhibit p110α. In the case of the procyanidin tetramers, no docking solutions were obtained (see Table 5). This indicates that it is not possible for a tetramer to compete with ATP for its binding site in p110α

because, although they are highly flexible molecules, they are also too big to have a conformation that hinder do not cause any steric hindrance with p110α.

Prediction of isoflavone activity

Experiments have been done to investigate the capacity of genistein (an isoflavone) to inhibit glucose transport [44, 45]. It has been shown to inhibit glucose transport in adipocytes without compromising the insulin-induced recruitment of GLUT4 to the plasma membrane (*i.e.* the number of plasma membrane-associated GLUT4 transporters is not affected by the presence of genistein). It has been suggested, then, that genistein directly binds to GLUT4 and causes a conformational change in the transporter that decreases its intrinsic activity [44]. In accordance with this finding, we predict that p110α is not inhibited by genistein, which explains why GLUT4 translocation is not affected by genistein. The remaining isoflavones analyzed only inhibit p110α when they are glycosylated or methylated (*i.e.* genistin, daidzin, glycitein and glycitin have IC50 values between 0.29 and 7.4 μM; see Figure 5 and Tables 5 and 6).

Prediction of equol and tannic acid activity

Finally, we also tested the compound equol, which is an isoflavandiol metabolized from daidzein by bacterial flora in the intestines and tannic acid. Our results for equol suggest that this molecule is either a moderately active or an inactive p110α inhibitor (see Table 6). Both conformational search methods assayed suggest that tannic acid is inactive as a p110α inhibitor (see Table 5).

CONCLUSIONS

We have developed a pharmacophore that can reproduce known experimental data on the capacity of some phenolic compounds to inhibit p110α or not [14-16, 44, 45]. It is reliable enough to be used to predict the inhibitory potential of other phenolic compounds on p110α. Our study highlights the importance of docking for obtaining the ligand conformations that have to be used to map an existing pharmacophore and shows that conformations obtained by other methods may produce misleading results if they are used to map this pharmacophore and predict activities.

To our knowledge, this is the first time that a systematic study has been made on the extent to which the most frequent naturally-occurring polyphenols and their bioactive forms inhibit p110α. When the docking-derived conformations of these phenolic compounds are mapped into the pharmacophore, we predict that those that can interact with p110α and inhibit its activity are

ISBN:978-84-691-1553-4 /DL: T.151-2008

stilbenes, flavonols (except myricetin), flavones, flavanones, anthocyanidins (except delphinidin and delphinidin 3-glucoside), most flavanol monomers [except (+)-epicatechin, (+/-)-epigallocatechin, (+/-)-gallocatechin, (-)-catechin and 4'-*O*-methyl-(-)-epigallocatechin], all procyanidin dimers and some isoflavones (*i.e.* daidzin, genistin, glycitin and glycitein) (see Tables 5 and 6). Interestingly, when comparing the activities on p110α of naturally-occurring phenolic compounds (see Table 5) and the bioactive forms derived from the absorption and metabolization of these compounds (see Table 6), we can see that glucuronidation, sulfation, glycosylation and methylation do not substantially alter the activity of the original molecule from which they are derived.

Finally, it can also be concluded that although it has been proved that the addition of phenolic compound extracts to food can have an overall benefit on health, it should be taken into account that some of these molecules may exacerbate insulin resistance in susceptible individuals via impaired glucose uptake in muscle and adipose tissues and, therefore, produce an undesirable side effect.

ACKNOWLEDGMENTS

We thank John Bates of our University's Language Service for correcting the manuscript, the Servei de Disseny de Fàrmacs from the Centre de Supercomputació de Catalunya (CESCA) for providing access to Catalyst™ and SimBioSys Inc. (Toronto, Ontario, Canada) for providing us with eHiTS®. This study was supported by grant number CO3/O8 from the Fondo de Investigación Sanitaria (FIS) and AGL2005-04889 from the Comisión Interministerial de Ciencia y Tecnología (CICYT) of the Spanish Government. Montserrat Vaqué is the recipient of a fellowship from grant number CO3/O8.

ISBN:978-84-691-1553-4 /DL: T.151-2008

REFERENCES

[1] N.J. Bryant, R. Govers, D.E. James, Regulated transport of the glucose transporter GLUT4., Nat Rev Mol Cell Biol 3 (2002) 267-277.

[2] S.W. Cushman, L.J. Wardzala, Potential mechanism of insulin action on glucose transport in the isolated rat adipose cell. Apparent translocation of intracellular transport systems to the plasma membrane., J Biol Chem 255 (1980) 4758-4762.

[3] C. Chaussade, G.W. Rewcastle, J.D. Kendall, W.A. Denny, K. Cho, L.M. Grønning, M.L. Chong, S.H. Anagnostou, S.P. Jackson, N. Daniele, P.R. Shepherd, Evidence for functional redundancy of class IA PI3K isoforms in insulin signalling., Biochem J 404 (2007) 449-458.

[4] Z.A. Knight, B. Gonzalez, M.E. Feldman, E.R. Zunder, D.D. Goldenberg, O. Williams, R. Loewith, D. Stokoe, A. Balla, B. Toth, T. Balla, W.A. Weiss, R.L. Williams, K.M. Shokat, A pharmacological map of the PI3-K family defines a role for p110 alpha in insulin signaling, Cell 125 (2006) 733-747.

[5] L.C. Foukas, M. Claret, W. Pearce, K. Okkenhaug, S. Meek, E. Peskett, S. Sancho, A.J.H. Smith, D.J. Withers, B. Vanhaesebroeck, Critical role for the p110 alpha phosphoinositide-3-OH kinase in growth and metabolic regulation, Nature 441 (2006) 366-370.

[6] E. Hirsch, C. Costa, E. Ciraolo, Phosphoinositide 3-kinases as a common platform for multi-hormone signaling., J Endocrinol 194 (2007) 243-256.

[7] A. Scalbert, G. Williamson, Dietary intake and bioavailability of polyphenols., J Nutr 130 (2000) 2073S-2085S.

[8] A. Fleuriet, J.J. Macheix, Phenolic Acids in Fruits and Vegetables, Flavonoids in Health and Disease, 2003, pp. 1-42.

[9] Documentation for the Update of the USDA Database for Flavonoid Content of Selected foods, Release 2.1 (2007).

[10] M. Pinent, C. Bladé, M.J. Salvadó, M. Blay, G. Pujadas, J. Fernández-Larrea, L. Arola, A. Ardévol, Procyanidin effects on adipocyte-related pathologies., Crit Rev Food Sci Nutr 46 (2006) 543-550.

[11] Y. Yilmaz, R.T. Toledo, Major flavonoids in grape seeds and skins: antioxidant capacity of catechin, epicatechin, and gallic acid., J Agric Food Chem 52 (2004) 255-260.

[12] J.A. Ross, C.M. Kasum, Dietary flavonoids: bioavailability, metabolic effects, and safety., Annu Rev Nutr 22 (2002) 19-34.

[13] P.M. Kris-Etherton, K.D. Hecker, A. Bonanome, S.M. Coval, A.E. Binkoski, K.F. Hilpert, A.E. Griel, T.D. Etherton, Bioactive compounds in foods: their role in the prevention of cardiovascular disease and cancer., Am J Med 113 Suppl 9B (2002) 71S-88S.

[14] S. Fröjdö, D. Cozzone, H. Vidal, L. Pirola, Resveratrol is a class IA phosphoinositide 3-kinase inhibitor., Biochem J (2007).

[15] A.W. Harmon, Y.M. Patel, Naringenin inhibits phosphoinositide 3-kinase activity and glucose uptake in 3T3-L1 adipocytes., Biochem Biophys Res Commun 305 (2003) 229-234.

[16] P. Strobel, C. Allard, T. Perez-Acle, R. Calderon, R. Aldunate, F. Leighton, Myricetin, quercetin and catechin-gallate inhibit glucose uptake in isolated rat adipocytes., Biochem J 386 (2005) 471-478.

[17] E.H. Walker, M.E. Pacold, O. Perisic, L. Stephens, P.T. Hawkins, M.P. Wymann, R.L. Williams, Structural determinants of phosphoinositide 3-kinase inhibition by wortmannin, LY294002, quercetin, myricetin, and staurosporine., Mol Cell 6 (2000) 909-919.

ISBN:978-84-691-1553-4 /DL: T.151-2008

[18] C.J. Vlahos, W.F. Matter, K.Y. Hui, R.F. Brown, A specific inhibitor of phosphatidylinositol 3-kinase, 2-(4-morpholinyl)-8-phenyl-4H-1-benzopyran-4-one (LY294002). J Biol Chem 269 (1994) 5241-5248.

[19] J.R. McMullen, F. Amirahmadi, E.A. Woodcock, M. Schinke-Braun, R.D. Bouwman, K.A. Hewitt, J.P. Mollica, L. Zhang, Y. Zhang, T. Shioi, A. Buerger, S. Izumo, P.Y. Jay, G.L. Jennings, Protective effects of exercise and phosphoinositide 3-kinase(p110alpha) signaling in dilated and hypertrophic cardiomyopathy., Proc Natl Acad Sci U S A 104 (2007) 612-617.

[20] J.R. McMullen, P.Y. Jay, PI3K(p110alpha) inhibitors as anti-cancer agents: minding the heart., Cell Cycle 6 (2007) 910-913.

[21] J.J. Macheix, A. Fleuriet, J. Billot, Fruit Phenolics, Boca Raton, FL: CRC Press 1990.

[22] C. Felgines, S. Talavéra, M.P. Gonthier, O. Texier, A. Scalbert, J.L. Lamaison, C. Rémésy, Strawberry anthocyanins are recovered in urine as glucuro- and sulfoconjugates in humans., J Nutr 133 (2003) 1296-1301.

[23] C. Felgines, S. Talavera, O. Texier, A. Gil-Izquierdo, J.L. Lamaison, C. Remesy, Blackberry anthocyanins are mainly recovered from urine as methylated and glucuronidated conjugates in humans., J Agric Food Chem 53 (2005) 7721-7727.

[24] C. Manach, G. Williamson, C. Morand, A. Scalbert, C. Rémésy, Bioavailability and bioefficacy of polyphenols in humans. I. Review of 97 bioavailability studies., Am J Clin Nutr 81 (2005) 230S-242S.

[25] C. Tsang, C. Auger, W. Mullen, A. Bornet, J.M. Rouanet, A. Crozier, P.L. Teissedre, The absorption, metabolism and excretion of flavan-3-ols and procyanidins following the ingestion of a grape seed extract by rats., Br J Nutr 94 (2005) 170-181.

[26] M. Hayakawa, H. Kaizawa, H. Moritomo, T. Koizumi, T. Ohishi, M. Okada, M. Ohta, S. Tsukamoto, P. Parker, P. Workman, M. Waterfield, Synthesis and biological evaluation of 4-morpholino-2-phenylquinazolines and related derivatives as novel PI3 kinase p110alpha inhibitors., Bioorg Med Chem 14 (2006) 6847-6858.

[27] M. Hayakawa, H. Kaizawa, K. Kawaguchi, N. Ishikawa, T. Koizumi, T. Ohishi, M. Yamano, M. Okada, M. Ohta, S. Tsukamoto, F.I. Raynaud, M.D. Waterfield, P. Parker, P. Workman, Synthesis and biological evaluation of imidazo[1,2-a]pyridine derivatives as novel PI3 kinase p110alpha inhibitors., Bioorg Med Chem 15 (2007) 403-412.

[28] A. Smellie, S.L. Teig, P. Towbin, Poling: Promoting conformational variation, J Comput Chem 16 (1995) 171-187.

[29] U. Pieper, N. Eswar, F.P. Davis, H. Braberg, M.S. Madhusudhan, A. Rossi, M. Marti-Renom, R. Karchin, B.M. Webb, D. Eramian, M.Y. Shen, L. Kelly, F. Melo, A. Sali, MODBASE: a database of annotated comparative protein structure models and associated resources., Nucleic Acids Res 34 (2006) D291-295.

[30] N. Guex, A. Diemand, M.C. Peitsch, Protein modelling for all., Trends Biochem Sci 24 (1999) 364-367.

[31] S.A. Aherne, N.M. O'Brien, Dietary flavonols: chemistry, food content, and metabolism., Nutrition 18 (2002) 75-81.

[32] C. Manach, A. Scalbert, C. Morand, C. Rémésy, L. Jiménez, Polyphenols: food sources and bioavailability., Am J Clin Nutr 79 (2004) 727-747.

[33] G. Agullo, L. Gamet-Payrastre, S. Manenti, C. Viala, C. Rémésy, H. Chap, B. Payrastre, Relationship between flavonoid structure and inhibition of phosphatidylinositol 3-kinase: a comparison with tyrosine kinase and protein kinase C inhibition., Biochem Pharmacol 53 (1997) 1649-1657.

[34] L. Gamet-Payrastre, S. Manenti, M.P. Gratacap, J. Tulliez, H. Chap, B. Payrastre, Flavonoids and the inhibition of PKC and PI 3-kinase., Gen Pharmacol 32 (1999) 279-286.

[35] M. Hayakawa, H. Kaizawa, H. Moritomo, T. Koizumi, T. Ohishi, M. Yamano, M. Okada, M. Ohta, S. Tsukamoto, F.I. Raynaud, P. Workman, M.D. Waterfield, P. Parker, Synthesis and biological evaluation of pyrido[3',2':4,5]furo[3,2-d]pyrimidine derivatives as novel PI3 kinase p110alpha inhibitors., Bioorg Med Chem Lett 17 (2007) 2438-2442.

ISBN:978-84-691-1553-4 /DL: T.151-2008

[36] M. Hayakawa, K. Kawaguchi, H. Kaizawa, T. Koizumi, T. Ohishi, M. Yamano, M. Okada, M. Ohta, S. Tsukamoto, F.I. Raynaud, P. Parker, P. Workman, M.D. Waterfield, Synthesis and biological evaluation of sulfonylhydrazone-substituted imidazo[1,2-a]pyridines as novel PI3 kinase p110alpha inhibitors., Bioorg Med Chem 15 (2007) 5837-5844.

[37] T.B. Lanni, K.L. Greene, C.N. Kolz, K.S. Para, M. Visnick, J.L. Mobley, D.T. Dudley, T.J. Baginski, M.B. Liimatta, Design and synthesis of phenethyl benzo[1,4]oxazine-3-ones as potent inhibitors of PI3Kinase gamma, Bioorg Med Chem Lett 17 (2007) 756-760.

[38] F.I. Raynaud, S. Eccles, P.A. Clarke, A. Hayes, B. Nutley, S. Alix, A. Henley, F. Di-Stefano, Z. Ahmad, S. Guillard, L.M. Bjerke, L. Kelland, M. Valenti, L. Patterson, S. Gowan, A. de Haven Brandon, M. Hayakawa, H. Kaizawa, T. Koizumi, T. Ohishi, S. Patel, N. Saghir, P. Parker, M. Waterfield, P. Workman, Pharmacologic characterization of a potent inhibitor of class I phosphatidylinositide 3-kinases., Cancer Res 67 (2007) 5840-5850.

[39] K.C. Ong, H.E. Khoo, Insulinomimetic effects of myricetin on lipogenesis and glucose transport in rat adipocytes but not glucose transport translocation., Biochem Pharmacol 51 (1996) 423-429.

[40] J. Peterson, J. Dwyer, Flavonoids: Dietary occurrence and biochemical activity, Nutr Res 18 (1998) 1995-2018.

[41] W. Mullen, S.C. Marks, A. Crozier, Evaluation of phenolic compounds in commercial fruit juices and fruit drinks., J Agric Food Chem 55 (2007) 3148-3157.

[42] M. Mochizuki, N. Hasegawa, Stereospecific effects of catechin isomers on insulin induced lipogenesis in 3T3-L1 cells., Phytother Res 18 (2004) 449-450.

[43] B. García-Ramírez, J. Fernandez-Larrea, M. Salvadó, A. Ardèvol, L. Arola, C. Bladé, Tetramethylated dimeric procyanidins are detected in rat plasma and liver early after oral administration of synthetic oligomeric procyanidins., J Agric Food Chem 54 (2006) 2543-2551.

[44] R.M. Smith, J.J. Tiesinga, N. Shah, J.A. Smith, L. Jarett, Genistein inhibits insulin-stimulated glucose transport and decreases immunocytochemical labeling of GLUT4 carboxyl-terminus without affecting translocation of GLUT4 in isolated rat adipocytes: additional evidence of GLUT4 activation by insulin., Arch Biochem Biophys 300 (1993) 238-246.

[45] M. Bazuine, P.J. van den Broek, J.A. Maassen, Genistein directly inhibits GLUT4-mediated glucose uptake in 3T3-L1 adipocytes., Biochem Biophys Res Commun 326 (2005) 511-514.

[46] D.J. Boocock, K.R. Patel, G.E. Faust, D.P. Normolle, T.H. Marczylo, J.A. Crowell, D.E. Brenner, T.D. Booth, A. Gescher, W.P. Steward, Quantitation of trans-resveratrol and detection of its metabolites in human plasma and urine by high performance liquid chromatography., J Chromatogr B Analyt Technol Biomed Life Sci 848 (2007) 182-187.

[47] P. Vitaglione, S. Sforza, G. Galaverna, C. Ghidini, N. Caporaso, P. Vescovi, V. Fogliano, R. Marchelli, Bioavailability of trans-resveratrol from red wine in humans., Mol Nutr Food Res 49 (2005) 495-504.

[48] T. Walle, F. Hsieh, M.H. DeLegge, J.E.J. Oatis, U.K. Walle, High absorption but very low bioavailability of oral resveratrol in humans., Drug Metab Dispos 32 (2004) 1377-1382.

[49] V.C.J. de Boer, Towards functional effects of polyphenols, Wageningen University, 2007.

[50] W. Mullen, C.A. Edwards, A. Crozier, Absorption, excretion and metabolite profiling of methyl-, glucuronyl-, glucosyl- and sulpho-conjugates of quercetin in human plasma and urine after ingestion of onions., Br J Nutr 96 (2006) 107-116.

[51] M.S. DuPont, A.J. Day, R.N. Bennett, F.A. Mellon, P.A. Kroon, Absorption of kaempferol from endive, a source of kaempferol-3-glucuronide, in humans., Eur J Clin Nutr 58 (2004) 947-954.

ISBN:978-84-691-1553-4 /DL: T.151-2008

[52] I. Erlund, M.L. Silaste, G. Alfthan, M. Rantala, Y.A. Kesäniemi, A. Aro, Plasma concentrations of the flavonoids hesperetin, naringenin and quercetin in human subjects following their habitual diets, and diets high or low in fruit and vegetables., Eur J Clin Nutr 56 (2002) 891-898.

[53] X. Wu, G. Cao, R.L. Prior, Absorption and metabolism of anthocyanins in elderly women after consumption of elderberry or blueberry., J Nutr 132 (2002) 1865-1871.

[54] T. Frank, M. Netzel, G. Strass, R. Bitsch, I. Bitsch, Bioavailability of anthocyanidin-3-glucosides following consumption of red wine and red grape juice., Can J Physiol Pharmacol 81 (2003) 423-435.

[55] N. Sabolovic, A.C. Humbert, A. Radominska-Pandya, J. Magdalou, Resveratrol is efficiently glucuronidated by UDP-glucuronosyltransferases in the human gastrointestinal tract and in Caco-2 cells., Biopharm Drug Dispos 27 (2006) 181-189.

TABLES AND FIGURES

Table 1. Information and statistical validation for the top ten hypotheses obtained by HypoGen™ and HypoRefine™ runs

Hypotheses	Features	Config. cost	Fixed cost	Total cost	Null–total cost	RMS	Correlation (r)
Hypotheses from HypoGen™ run		16.355	88.116				
1+	HA-HY-HY-RA			97.357	59.246	0.897	0.950
2	HA-HY-HY-RA			99.219	57.384	1.010	0.936
3	HA-HY-HY-RA			100.601	56.002	1.066	0.928
4	HA-HY-HY-RA			102.378	54.225	1.123	0.921
5	HA-HY-HY-RA			105.517	51.086	1.222	0.906
6	HA-HY-RA			106.433	50.170	1.214	0.909
7	HA-HY-RA			107.093	49.510	1.174	0.918
8	HA-HY-HY-RA			107.548	49.055	1.311	0.890
9	HA-HY-RA			110.458	46.145	1.384	0.878
10	HA-HY-HY-RA			110.499	46.104	1.459	0.860
Hypotheses from HypoRefine™ run		16.388	88.148				
1+	HA-HY-HY-RA*			96.307	60.296	0.878	0.952
2+	HA-HY-HY-RA			96.485	60.118	0.856	0.955
3+	HA-HY-HY-RA*			97.535	59.068	0.936	0.945
4	HA-HY-HY-RA*			99.840	56.763	1.023	0.934
5	HA-HY-HY-RA			106.007	50.596	1.287	0.894
6	HA-HY-HY-RA			106.270	50.333	1.270	0.897
7	HA-HA-HY-RA*			106.369	50.234	1.296	0.892
8	HA-HA-HY-RA*			107.457	49.146	1.347	0.882
9	HA-HA-HY-RA			108.921	47.682	1.398	0.873
10	HA-HY-RA*			109.252	47.351	1.300	0.895

Top ten hypotheses obtained from the HypoGen™ and the HypoRefine™ runs that were carried out in the following conditions: (a) training set from Figure 1; (b) the features considered are only HA (hydrogen-bond acceptor), HD (hydrogen-bond donor), HY (hydrophobic group) and RA (aromatic ring); and (c) in a hypothesis, the minimum number of HA, HY and RA functions is one whereas the maximum for all four functions is five. It is worth pointing out that the

hypotheses with identical features differ in the 3D location of their functions and that the null cost is 156.603.

A * superindex symbol beside the hypothesis number indicates a pharmacophore that includes an excluded volume.

A + superindex symbol beside the hypothesis number indicates the pharmacophores that are identified as correct by statistical validation process (where the difference between their total cost and null cost should be approximately 60 bits for at least a 90% probability that the hypothesis is a real correlation with biological activity). To distinguish these validated hypothesis from the rest, we have given them a new name: (a) Hypo1 for the first hypothesis obtained from HypoGen^{TM}; (b) Hypo2 for the first hypothesis obtained from HypoRefine^{TM}; (c) Hypo3 for the second hypothesis obtained from HypoRefine^{TM}; and (d) Hypo4 for the third hypothesis obtained from HypoRefine^{TM}.

Table 2. Validation, by predicting the activity of the training set, of those hypotheses that have been identified as correct by the statistical validation process

The experimental and the estimated IC50 values for the training set molecules are indicated in the **Exp. IC50** and **Est. IC50** columns. The experimental IC50 values were obtained from references [26] and [27]. The estimated IC50 values, on the other hand, were calculated by using the hypothesis indicated. Columns **HA**, **HD**, **HY** and **HA** indicate what hypothesis features are mapped by each molecule (*i.e.* a + sign indicates that the corresponding pharmacophore function is mapped by the molecule whereas the – sign indicates the opposite). The IC50 values span 5 orders of magnitude (ranging from 0.0018 to 60 μM) and they have been classified in the **activity scale** column under the following criteria: (a) IC50 < 0.1 μM (*i.e.* highly active; ++++); (b) 0.1 μM ≤ IC50 < 1 μM (*i.e.* active; +++); (c) 1 μM ≤ IC50 < 10 μM (*i.e.* moderately active; ++); and (d) IC50 >10 μM (*i.e.* inactive; +). The **error column** is computed as the ratio of the experimental IC50 relative to the estimated IC50 (or *vice versa* if the latter is greater than the former). An error with a negative sign indicates that the experimental IC50 is higher than the corresponding estimation. Molecules considered as inactive in the substractive phase of the pharmacophore building are 10f, 5b, 5d, 5c and 5a.

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF
PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

Table 2.

			Hypo1							Hypo2							Hypo3							Hypo4						
Comp.	**Exp. IC50**	**activity scale**	**Est. IC50**	**activity scale**	**Error**	**HA**	**HY**	**HY**	**RA**	**Est. IC50**	**activity scale**	**Error**	**HA**	**HD**	**HY**	**HA**	**Est. IC50**	**activity scale**	**Error**	**HA**	**HY**	**HY**	**RA**	**Est. IC50**	**activity scale**	**Error**	**HA**	**HD**	**HY**	**HA**
2g	0.0018	++++	0.0023	++++	-4.4	+	+	+	+	0.0074	++++	4.1	+	+	+	+	0.0029	++++	1.6	+	+	+	+	0.0082	++++	4.6	+	+	+	+
15e	0.0020	++++	0.002	++++	3.7	+	+	+	+	0.016	++++	7.8	+	+	+	+	0.0082	++++	4.1	+	+	+	+	0.013	++++	6.5	+	+	+	+
2f	0.0031	++++	0.0022	++++	-1.4	+	+	+	+	0.0082	++++	2.6	+	+	+	+	0.0026	++++	-1.2	+	+	+	+	0.0085	++++	2.7	+	+	+	+
15c	0.0130	++++	0.033	++++	2.5	+	+	+	+	0.055	++++	4.2	+	+	+	+	0.023	++++	1.8	+	+	+	+	0.04	++++	3.1	+	+	+	+
14b	0.02	++++	0.089	++++	4.4	+	+	+	+	0.019	++++	-1.1	+	+	+	+	0.05	++++	2.5	+	+	+	+	0.0088	++++	-2.3	+	+	+	+
15	0.0301	++++	0.041	++++	1.3	+	+	+	+	0.017	++++	-1.8	+	+	+	+	0.048	++++	1.5	+	+	+	+	0.012	++++	-2.5	+	+	+	+
15a	0.056	++++	0.032	++++	-1.7	+	+	+	+	0.033	++++	-1.7	+	+	+	+	0.031	++++	-1.8	+	+	+	+	0.045	++++	-1.2	+	+	+	+
11	0.082	++++	0.24	+++	3	+	+	+	+	0.033	++++	-2.5	+	+	+	+	0.13	+++	1.6	+	+	+	+	0.047	++++	-1.7	+	+	+	+
6h	0.1	+++	0.13	+++	1.3	+	+	+	+	0.03	++++	-3.3	+	+	+	+	0.2	+++	2	+	+	+	+	0.045	++++	-2.2	+	+	+	+
8	0.1	+++	0.023	++++	-4.4	+	+	+	+	0.017	++++	-5.9	+	+	+	+	0.024	++++	-4.2	+	+	+	+	0.013	++++	-7.8	+	+	+	+
ly294002	0.63	+++	0.7	+++	1.1	+	+	+	-	0.87	+++	1.4	+	+	+	-	0.72	+++	1.1	+	+	+	-	1.6	++	2.6	+	+	-	+
2a	0.67	+++	0.43	+++	-1.6	+	+	-	+	0.63	+++	-1.1	+	-	+	+	0.68	+++	1	+	+	-	+	0.99	+++	1.5	+	+	-	+
6k	0.93	+++	2.3	++	2.5	-	+	+	+	1.8	++	1.9	-	+	+	+	2.9	++	3.1	-	+	+	+	4	++	4.3	-	+	+	+
10g	2.1	++	0.57	+++	-3.7	+	+	+	-	1	++	-2	+	+	+	-	0.55	+++	-3.8	+	+	+	-	2.2	++	1.1	+	+	+	-
6i	3.7	++	3.6	++	-1	-	+	+	+	2.2	++	-1.7	-	+	+	+	2.7	++	-1.4	+	-	+	+	1.9	++	-1.9	+	+	+	-
10d	3.9	++	3.9	++	1	-	+	+	+	1.2	++	-3.1	-	+	+	+	6	++	1.5	-	+	+	+	1.2	++	-3.2	-	+	+	+
10f	9.8	++	1.1	++	-9.1	+	-	+	+	19	+	1.9	+	-	+	+	1.1	++	-9	+	-	+	+	21	+	2.2	+	+	-	+
5b	12	+	23	+	1.9	+	-	+	+	22	+	1.8	+	+	-	-	20	+	1.7	+	-	+	+	9.1	++	-1.3	+	-	+	+
5d	15	+	9.9	++	-1.5	+	-	+	+	22	+	1.5	+	+	-	-	11	+	-1.4	+	-	+	+	14	+	-1.1	+	-	+	+
5c	21	+	53	+	2.5	-	-	+	+	30	+	1.4	-	+	-	+	60	+	2.9	+	-	+	-	17	+	-1.2	+	-	+	+
5a	60	+	12	+	-5	-	+	+	+	26	+	-2.3	-	+	-	+	13	+	-4.5	-	+	+	+	16	+	-3.7	+	-	+	+

ISBN:978-84-691-1553-4 /DL: T.151-2008

Table 3. Statistical information about the validation of Hypo1, Hypo2, Hypo3 and Hypo4 by means of a randomization test

HypoGen™ selected hypothesis	Total cost	Fixed cost	RMS	Correlation (r)	HypoRefine™ selected hypotheses	Total cost	Fixed cost	RMS	Correlation (r)
Hypo 1	97.357	88.116	0.897	0.950	**Hypo 2**	96.307	88.148	0.878	0.952
					Hypo 3	96.485	88.148	0.856	0.955
					Hypo 4	97.535	88.148	0.936	0.945
HypoGen™ runs after applying catScramble™	**Total cost**	**Fixed cost**	**RMS**	**Correlation (r)**	**HypoRefine™ runs after applying catScramble™**	**Total cost**	**Fixed cost**	**RMS**	**Correlation (r)**
1	129.597	87.202	2.009	0.712	**1**	122.892	82.780	1.928	0.741
2	124.759	81.390	1.969	0.732	**2**	128.061	83.410	1.875	0.784
3	124.334	83.183	1.907	0.752	**3**	120.481	85.969	1.678	0.820
4	151.970	85.633	2.505	0.486	**4**	142.321	86.088	2.281	0.610
5	134.733	83.337	2.123	0.686	**5**	116.679	88.183	1.580	0.837
6	145.164	85.507	2.317	0.606	**6**	132.740	88.854	2.044	0.700
7	138.898	82.408	2.198	0.675	**7**	142.187	81.581	2.354	0.582
8	141.799	85.231	2.271	0.618	**8**	128.492	84.804	2.004	0.717
9	133.303	87.604	2.012	0.686	**9**	133.479	85.162	2.128	0.670
10	129.325	82.649	2.020	0.719	**10**	126.615	87.852	1.864	0.763
11	151.517	83.924	2.536	0.464	**11**	129.586	87.110	1.970	0.729
12	148.957	82.597	2.510	0.481	**12**	147.042	86.577	2.395	0.548
13	131.952	83.214	2.139	0.666	**13**	154.454	82.227	2.622	0.401
14	149.027	82.769	2.504	0.486	**14**	134.988	82.406	2.238	0.623
15	122.578	87.760	1.770	0.789	**15**	142.124	82.662	2.282	0.636
16	135.826	87.351	2.113	0.678	**16**	149.666	82.396	2.465	0.548
17	132.576	86.983	2.084	0.685	**17**	124.469	82.980	1.981	0.722
18	135.936	82.596	2.209	0.643	**18**	132.862	83.386	2.033	0.729
19	118.065	82.262	1.780	0.787	**19**	138.668	85.891	2.228	0.629

Comparison, considering a 95% confidence level, between the statistical information from: (a) hypotheses Hypo1, Hypo2, Hypo3 and Hypo4; and (b) HypoGen™ or HypoRefine™ runs done after catScramble™ performed 19 random reassignations of the activity values among the training set compounds. All 19 HypoGen™ runs that are compared with Hypo1 were carried out in identical experimental conditions (*i.e.* features, parameters) to those used in the HypoGen™ run that provided the Hypo1 hypothesis (and the same is true for all 19 HypoRefine™ runs that are compared with the Hypo2, Hypo3 and Hypo4 hypotheses). Data in the *Total cost*, *RMS* and

Correlation (r) columns from the HypoGen™ or HypoRefine™ runs done after the data had been randomized correspond to the first out of the 10 hypotheses obtained in each run (*i.e.* what is supposed to be the best hyphotesis obtained by the run).

Table 4. Hypothesis validation by predicting the activity of the compounds in the test set

			Hypo1 r = 0.839			Hypo2 r = 0.790			Hypo3 r = 0.851			Hypo4 r = 0.748		
Comp.	**Exp. IC50**	**activity scale**	**Est. IC50**	**activity scale**	**Error**	**Est. IC50**	**activity scale**	**Error**	**Est. IC50**	**activity scale**	**Error**	**Est. IC50**	**activity scale**	**Error**
12	0.0028	++++	0.0066	++++	2.4	0.013	++++	4.5	0.0082	++++	2.9	0.0095	++++	3.4
2e	0.0028	++++	0.0033	++++	1.2	0.012	++++	4.4	0.0037	++++	1.3	0.012	++++	4.3
15d	0.019	++++	0.005	++++	-3.8	0.0091	++++	-2.1	0.0054	++++	-3.5	0.0092	++++	-2.1
15b	0.027	++++	0.12	+++	4.5	0.12	+++	4.6	0.11	+++	4.1	0.12	+++	4.3
6a	0.075	++++	0.033	++++	-2.3	0.024	++++	-3.2	0.041	++++	-1.8	0.031	++++	-2.4
2d	0.28	+++	0.0049	++++	-57	0.012	++++	-24	0.0073	++++	-38	0.012	++++	-24
6j	0.44	+++	1.6	++	3.6	1.5	++	3.3	2.2	++	5.1	1.9	++	4.2
6b	0.6	+++	0.03	++++	-20	0.02	++++	-26	0.3	+++	-18	0.025	++++	-24
2b	0.76	+++	0.13	+++	-5.8	0.02	++++	-35	0.2	+++	-3.7	0.029	++++	-26
1	1.3	++	4	++	3.1	29	+	22	5.6	++	4.3	16	+	12
6c	1.8	++	2.5	++	1.4	1.9	++	1	3.1	++	1.7	1.7	++	-1.1
10b	2.8	++	1.9	++	-1.4	3.3	++	1.2	1.9	++	-1.5	1.6	++	-1.7
10a	2.9	++	3.9	++	1.3	1.3	++	-2.3	6	++	2.1	1.1	++	-2.7
10h	14	+	5	++	-2.8	28	+	2	7.6	++	-1.8	36	+	2.5
10e	30	+	4.5	++	-6.6	23	+	-1.3	7	++	-4.3	1	++	-29

The experimental and the estimated IC50 values for the test set molecules are indicated in the **Exp. IC50** and **Est. IC50** columns. The experimental IC50 values were obtained from references [26] and [27]. The estimated IC50 values, on the other hand, were calculated by using the hypothesis indicated. The value below the name of each hypothesis (*i.e.* r) corresponds to the correlation that is obtained after plotting the experimental versus the predicted IC50 value for each compound and then making a regression analysis. The IC50 values span 5 orders of magnitude (ranging from 0.0028 to 36 μM) and they have been classified in the **activity scale** column under the following criteria: (a) IC50 < 0.1 μM (*i.e.* highly active; ++++); (b) 0.1 μM ≤ IC50 < 1 μM (*i.e.* active; +++); (c) 1 μM ≤ IC50 < 10 μM (*i.e.* moderately active; ++); and (d) IC50 >10 μM (*i.e.* inactive; +). The **error column** is computed as the ratio of the experimental IC50 relative to the estimated IC50 (or the *vice versa* if the latter is greater than the former). An error with a negative sign indicates that the experimental IC50 is higher than the corresponding estimation.

ISBN:978-84-691-1553-4 /DL: T.151-2008

Table 5. IC50 prediction of the activity of the naturally-occurring forms of phenolic compounds most frequently found in plant extracts on p110α by using Hypo3

Naturally-occurring phenolic compounds	Estimated IC50 (μM) docking	activity scale	Estimated IC50 (μM) Catalyst™	activity scale
Non Flavonoids				
Phenolic acids				
m-hydroxyphenylacetic acid	160	+	150	+
3,4-dihydroxyphenilacetic acid	150	+	140	+
p-hydroxyphenylacetic acid	130	+	140	+
m-hydroxybenzoic acid	170	+	180	+
p-hydroxybenzoic acid	170	+	180	+
p-coumaric acid	170	+	170	+
caffeic acid	170	+	170	+
ferulic acid	11	+	13	+
gallic acid	170	+	180	+
hippuric acid	62	+	56	+
m-hydroxyhippuric acid	61	+	53	+
m-hydroxyphenylpropionic acid	150	+	170	+
protocatechuic acid	170	+	180	+
sinapic acid	24	+	43	+
syringic acid	50	+	50	+
vanillic acid	50	+	50	+
Stilbenes				
astringin	0.56	+++	0.62	+++
piceatannol	0.86	+++	0.94	+++
piceid	0.64	+++	0.75	+++
trans-resveratrol	0.92	+++	1.2	++
resveratroloside	0.85	+++	0.85	+++
Flavonoids				
Flavonols				
fisetin	0.85	+++	0.84	+++
galangin	0.92	+++	0.95	+++
isorhamnetin	0.0062	++++	0.0130	++++
kaempferol	0.83	+++	0.89	+++
morin	1.1	++	0.9	+++
myricetin	110	+	160	+
quercetin	0.81	+++	0.85	+++
Flavones				
apigenin	0.83	+++	0.89	+++
chrysin	0.91	+++	1	++
diosmetin	0.83	+++	0.83	+++
luteolin	0.84	+++	0.84	+++

Naturally-occurring phenolic compounds	Estimated IC50 (μM) docking	activity scale	Estimated IC50 (μM) Catalyst™	activity scale
Flavanones				
hesperetin	0.79	+++	0.82	+++
naringenin	0.82	+++	0.87	+++
taxifolin	0.79	+++	0.85	+++
Anthocyanidins				
cyanidin	0.8	+++	0.79	+++
delphinidin	64	+	160	+
malvidin	0.0069	++++	0.0046	++++
peonidin	0.09	++++	0.0062	++++
petunidin	0.86	+++	0.83	+++
Flavanols				
Monomers (catechins)				
(+)-catechin	0.87	+++	0.9	+++
(+)-catechin 3-gallate	2.0	++	0.046	++++
(+)-epicatechin	46	+	0.82	+++
(+)-epicatechin 3-gallate	0.12	+++	0.25	+++
(+)-epigallocatechin	98	+	110	+
(+)-gallocatechin	69	+	65	+
(-)-catechin	42	+	0.82	+++
(-)-catechin 3-gallate	0.18	+++	0.1	+++
(-)-epicatechin	0.82	+++	0.86	+++
(-)-epicatechin 3-gallate	0.8	+++	0.22	+++
(-)-epigallocatechin	99	+	110	+
(-)-gallocatechin	95	+	110	+
Procyanidins				
dimer A1	0.99	+++	0.84	+++
dimer A2	1	++	0.77	+++
dimer B1	0.028	++++	0.14	+++
dimer B2	0.95	+++	0.007	++++
dimer B3	0.76	+++	0.019	++++
dimer B4	0.9	+++	0.025	++++
dimer B5	0.26	+++	0.22	+++
dimer B6	0.81	+++	0.11	+++
dimer B7	0.83	+++	0.038	++++
dimer B8	0.81	+++	0.85	+++
trimer C1	1100	no mapping	0.17	+++
trimer C2	1100	no mapping	0.34	+++
trimer T2	1100	no mapping	0.23	+++
trimer T3	1100	no mapping	0.019	++++

ISBN:978-84-691-1553-4 /DL: T.151-2008

Naturally-occurring phenolic compounds	Estimated IC50 (μM) docking	activity scale	Estimated IC50 (μM) Catalyst™	activity scale
trimer T4	1100	no mapping	0.12	+++
trimer T5	1100	no mapping	0.0410	++++
trimer T6	1100	no mapping	0.12	+++
trimer T7	1100	no mapping	0.073	++++
tetramer 1	no solution	-	0.36	+++
tetramer 2	no solution	-	0.19	+++
Isoflavones				
daidzein	53	+	69	+
daidzin	0.89	+++	7.4	++
genistein	12	+	35	+
genistin	1.5	++	5.2	++
glycitein	0.29	+++	1	++
glycitin	1.8	++	0.42	+++
Other compounds				
tannic acid	1100	no mapping	62	+

The estimated IC50 values in the table were obtained by: (1) generating all possible conformations for a specific phenolic compound; (2) mapping the resulting conformations on Hypo3 by means of the "Fast fit" Catalyst™ option; and (3) selecting the lowest IC50 value (*i.e.* highest inhibitory activity) obtained from the mapping. Two different methods were used to generate the ligand conformations (*i.e.* docking on the p110α model and the BEST conformational search algorithm from Catalyst™) and the lowest IC50 given by each method is reported. The IC50 values of the phenolic compounds that map the pharmacophore span six orders of magnitude (ranging from 0.0046 to 180 μM) and they have been classified in the **activity scale** column under the following criteria: (a) IC50 < 0.1 μM (*i.e.* highly active; ++++); (b) 0.1 μM ≤ IC50 < 1 μM (*i.e.* active; +++); (c) 1 μM ≤ IC50 < 10 μM (*i.e.* moderately active; ++); and (d) IC50 >10 μM (*i.e.* inactive; +).

Table 6. IC50 prediction of the activity of the bioactive forms of phenolic compounds on p110α by using Hypo3

Bioactive forms of phenolic compounds	Estimated IC50 (μM) docking	activity scale	Estimated IC50 (μM) Catalyst™	activity scale	Reference
Non Flavonoids					
Phenolic acids					
m-hydroxyphenylacetic acid‡	100	+	150	+	[24]
p-hydroxyphenylacetic acid*‡	110	+	140	+	[24]
3,4-dihydroxyphenylacetic acid‡	150	+	140	+	[24]
m-hydroxybenzoic acid‡	170	+	180	+	[24]
p-hydroxybenzoic acid*‡	170	+	180	+	[24]
caffeic acid	170	+	170	+	[24]
chlorogenic acid	51	+	50	+	[24]
p-coumaric acid*‡	170	+	170	+	[24]
ferulic acid	11	+	13	+	[24]
gallic acid	170	+	180	+	[24]
gallic 3-glucuronide acid †	160	+	72	+	[24]
gallic 4-glucuronide acid †	74	+	70	+	[24]
4-*O*-methylgallic acid‡	64	+	63	+	[24]
homovanillic acid‡	50	+	50	+	[24]
phenylacetic acid‡	130	+	130	+	[24]
phenylpropionic acid*‡	90	+	210	+	[24]
m-hydroxyphenylpropionic acid‡	150	+	170	+	[24]
p-hydroxyphenylpropionic acid‡	90	+	170	+	[24]
protocatechuic acid*‡	170	+	180	+	[24]
m-hydroxyphenylvaleric acid‡	50	+	50	+	[24]
p-hydroxyphenylvaleric acid	50	+	50	+	[24]
Stilbenes					
trans- resveratrol	0.92	+++	1.2	++	[46]
cis-resveratrol 3-glucuronide +	0.57	+++	0.46	+++	[55]
trans-resveratrol 3-glucuronide	0.99	+++	0.78	+++	[47]
trans-resveratrol 4'-glucuronide	0.81	+++	0.86	+++	[47]
trans-resveratrol 3-sulfate*	0.56	+++	0.65	+++	[48]
Flavonoids					
Flavonols					
Isorhamnetin	0.0062	++++	0.013	++++	[24]
3'-*O*-methylquercetin 3-glucuronide	0.068	++++	0.0066	++++	[24, 49, 50]
quercetin	0,81	+++	0,85	+++	[24]
quercetin 3-glucuronide	0.97	+++	0.82	+++	[24, 50]
quercetin 3-sulfate	0.9	+++	0.83	+++	[24, 50]
kaempferol	0.83	+++	0.89	+++	[51]

ISBN:978-84-691-1553-4 /DL: T.151-2008

Bioactive forms of phenolic compounds	Estimated IC50 (μM) docking	activity scale	Estimated IC50 (μM) Catalyst™	activity scale	Reference
kaempferol 3-glucuronide	0.94	+++	0.81	+++	[51]
Flavanones					
hesperetin	0.79	+++	0.82	+++	[52]
hesperetin 3'-glucuronide †	0.68	+++	0.29	+++	[24]
hesperetin 7-glucuronide †	1	++	0.41	+++	[24]
naringenin	0.82	+++	0.87	+++	[24]
Anthocyanidins					
cyanidin	0.8	+++	0.79	+++	[23]
cyanidin 3-glucoside	0.79	+++	0.86	+++	[23, 24]
cyanidin 3-xyloside	0.82	+++	0.79	+++	[23]
cyanidin 3-sambubioside	0.71	+++	0.85	+++	[53]
cyanidin 3-glucuronide †	0.85	+++	0.86	+++	[23]
cyanidin 3'-glucuronide †	0.83	+++	0.81	+++	[23]
cyanidin 4'-glucuronide †	0.8	+++	0.8	+++	[23]
cyanidin 7-glucuronide †	0.55	+++	0.55	+++	[23]
cyanidin 3-glucoside 3'-glucuronide †	0.82	+++	0.84	+++	[53]
cyanidin 3-glucoside 4'-glucuronide †	0.66	+++	0.79	+++	[53]
cyanidin 3-glucoside-7-glucuronide †	0.38	+++	0.49	+++	[53]
delphinidin 3-glucoside	19	+	79	+	[54]
malvidin 3-glucoside	0.027	++++	0.0071	++++	[24, 54]
pelargonidin	0.83	+++	0.93	+++	[22]
pelargonidin 3-glucoside	0.82	+++	0.9	+++	[22, 24]
pelargonidin 3-glucuronide †	0.83	+++	0.88	+++	[22]
pelargonidin 4'glucuronide †	0.80	+++	0.79	+++	[22]
pelargonidin 7-glucuronide †	0.39	+++	0.52	+++	[22]
Peonidin	0.09	++++	0.0062	++++	[23]
peonidin 3-glucoside	0.034	++++	0.0082	++++	[23]
peonidin 3-sambubioside	0.01	++++	0.011	++++	[53]
peonidin 3-glucuronide †	0.029	++++	0.0072	++++	[22]
peonidin 4'-glucuronide †	0.098	++++	0.0072	++++	[22]
peonidin 7-glucuronide †	0.048	++++	0.0044	++++	[22]
petunidin 3-glucoside	2.7	++	0.92	+++	[54]
Flavanols					
Monomers (catechins)					
(+)-catechin 3'-glucuronide* †	0.84	+++	0.84	+++	[25]
(+)-catechin 3-glucuronide* †	0.62	+++	0.55	+++	[25]
(+)-catechin 4'-glucuronide* †	0.8	+++	0.83	+++	[25]
(+)-catechin 7-glucuronide* †	1.1	++	0.54	+++	[25]
3'-*O*-methyl-(+)-catechin*	0.0043	++++	0.012	++++	[25]
3'-*O*-methyl-(+)-catechin 3-glucuronide* †	0.0054	++++	0.01	++++	[25]
3'-*O*-methyl-(+)-catechin 4'-glucuronide* †	0.39	+++	0.26	+++	[25]

Bioactive forms of phenolic compounds	Estimated IC50 (μM) docking	activity scale	Estimated IC50 (μM) Catalyst™	activity scale	Reference
3'-*O*-methyl-(+)-catechin 7-glucuronide* †	0.045	++++	0.015	++++	[25]
(-)-epicatechin	0.82	+++	0.86	+++	[24]
(-)-epicatechin 3'-glucuronide* †	0.81	+++	0.85	+++	[25]
(-)-epicatechin 3-glucuronide* †	0.81	+++	0.79	+++	[25]
(-)-epicatechin 4'-glucuronide* †	0.79	+++	0.82	+++	[25]
(-)-epicatechin 7-glucuronide* †	1.1	++	0.5	+++	[25]
4'-*O*-methyl-(-)-epicatechin	0.8	+++	0.84	+++	[24]
4'-*O*-methyl-(-)-epicatechin 3-glucuronide	0.67	+++	0.31	+++	[24]
4'-*O*-methyl-(-)-epicatechin 5-glucuronide	0.83	+++	0.37	+++	[24]
4'-*O*-methyl-(-)-epicatechin 7-glucuronide	1	++	0.3	+++	[24]
3'-*O*-methyl-(-)-epicatechin 7-glucuronide	0.46	+++	0.073	++++	[25]
(-)-epigallocatechin	99	+	110	+	[24]
4'-*O*-methyl-(-)-epigallocatechin	72	+	7.3	++	[24]
(-)-epigallocatechingallate (EGCG)	5	++	0.83	+++	[24]
Procyanidins					
tetramethylated dimeric*	0.032	++++	0.011	++++	[43]
dimer B1	0.028	++++	0.14	+++	[24]
dimer B2	0.95	+++	0.007	++++	[24]
dimer B3*	0,76	+++	0.019	++++	[25]
dimer B4*	0.9	+++	0.025	++++	[25]
trimer C2*	1100	no mapping	0.34	+++	[25]
Isoflavones					
daidzein	53	+	69	+	[24]
dihydrodaidzein‡	53	+	1.7	++	[24]
daidzin	0.89	+++	7.4	++	[24]
genistein	12	+	35	+	[24]
dihydrogenistein‡	27	+	9.2	++	[24]
genistin	1.5	++	5.2	++	[24]
glycitein	0.29	+++	1	++	[24]
glycitin	1.8	++	0.42	+++	[24]
Other compounds					
equol‡	50	+	1.2	++	[24]

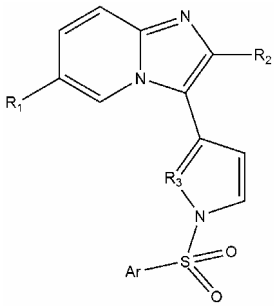
This table contains derivates of phenolic compounds that have been found in plasma or urine in humans. Since the data obtained from humans is limited and there is no information about human bioavailability for all classes of phenolic compounds, we have also included several metabolites that have been determined in rats (indicated by *) or compounds that were detected

ISBN:978-84-691-1553-4 /DL: T.151-2008

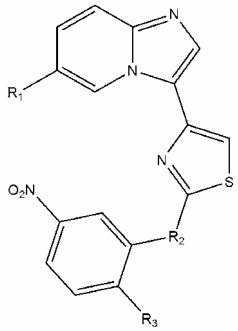
when working with human intestinal cell lines (indicated by +). Metabolites produced by intestinal microflora are also indicated by a ‡ symbol. The bibliografy on the bioactivity of each compound is given. The estimated IC50 values in the table were obtained by: (1) generating all possible conformations for a specific phenolic compound; (2) mapping the resulting conformations on Hypo3 by means of the "Fast fit" Catalyst^{TM} option; and (3) selecting the lowest IC50 value (*i.e.* highest inhibitory activity) obtained from the mapping. Two different methods were used to generate the ligand conformations (*i.e.* docking on the p110α model and the BEST conformational search algorithm from Catalyst^{TM}) and the lowest IC50 given by each method is reported. The IC50 values of the phenolic compounds that map the pharmacophore span six orders of magnitude (ranging from 0.0043 to 210 μM) and they have been classified in the **activity scale** column under the following criteria: (a) IC50 < 0.1 μM (*i.e.* highly active; ++++); (b) 0.1 μM ≤ IC50 < 1 μM (*i.e.* active; +++); (c) 1 μM ≤ IC50 < 10 μM (*i.e.* moderately active; ++); and (d) IC50 >10 μM (*i.e.* inactive; +).

† Several studies have detected glucuronide compounds in plasma or urine, but the exact site of glucuronidation has yet to be determined. Therefore, we have built all the possible structures for further predicting the IC50 of these conjugated compounds.

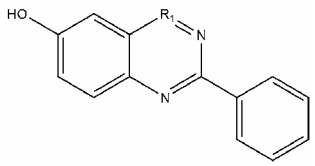
Figure 1. Chemical structures of the molecules used either to generate (*i.e.* training set) or to validate (*i.e.* test set) the pharmacophore candidates obtained with either HypoGen™ or HypoRefine™

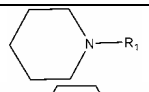
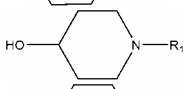
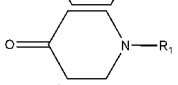
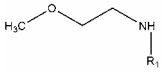
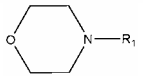


	R1	R2	R3	Ar	IC50 (μM)
2a (training)	H	CH3	N	4-Fluorophenyl	0.67
2f (training)	Br	CH3	N	2-Methyl-5-nitrophenyl	0.0031
2g (training)	Br	H	N	2-Methyl-5-nitrophenyl	0.0018
8 (training)	Cl	H	CH	2-Methyl-5-nitrophenyl	0.10
2b (test)	Cl	CH3	N	4-Fluorophenyl	0.76
2d (test)	Cl	CH3	N	3-Nitrophenyl	0.28
2e (test)	Cl	CH3	N	2-Methyl-5-nitrophenyl	0.008

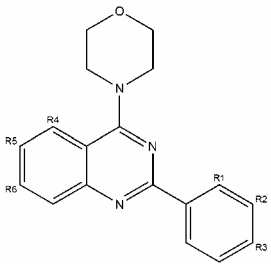


	R1	R2	R3	IC50 (μM)
11 (training)	Cl	-S-	CH3	0.082
15 (training)	Cl	-SO-	CH3	0.031
14b (training)	Cl	-SO-	CH2OH	0.020
12 (test)	Cl	-SO2-	CH3	0.0028

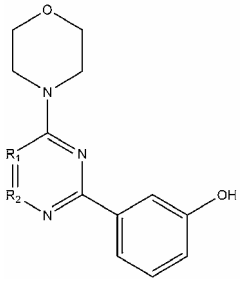


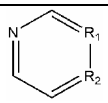
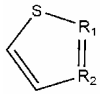
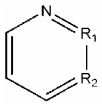
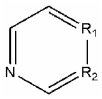
	R1	IC50 (μM)
5a (training)		>60
5b (training)		12
5c (training)		21
5d (training)		15
1 (test)		1.3

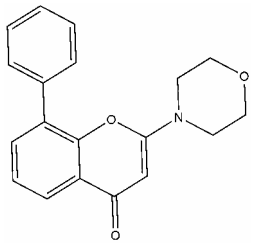
ISBN:978-84-691-1553-4 /DL: T.151-2008



	R1	R2	R3	R4	R5	R6	IC50 (μM)
6h (training)	H	CH2OH	H	H	OH	H	0.10
6i (training)	H	H	CH2OH	H	OH	H	3.7
10d (training)	NH2	H	H	H	OH	H	3.9
6k (training)	H	H	NH2	H	OH	H	0.93
10f (training)	H	H	H	H	H	OH	9.8
10g (training)	H	H	H	H	OCH3	H	2.1
10a (test)	OH	H	H	H	OH	H	2.9
6a (test)	H	OH	H	H	OH	H	0.075
10b (test)	H	H	OH	H	OH	H	2.8
6b (test)	H	OCH3	H	H	OH	H	0.60
6c (test)	H	F	H	H	OH	H	1.8
6j (test)	H	NH2	H	H	OH	H	0.44
10e (test)	H	H	H	OH	H	H	>30
10h (test)	H	H	H	H	H	H	14
15a (test)	H	OH	H	H	H	H	0.056



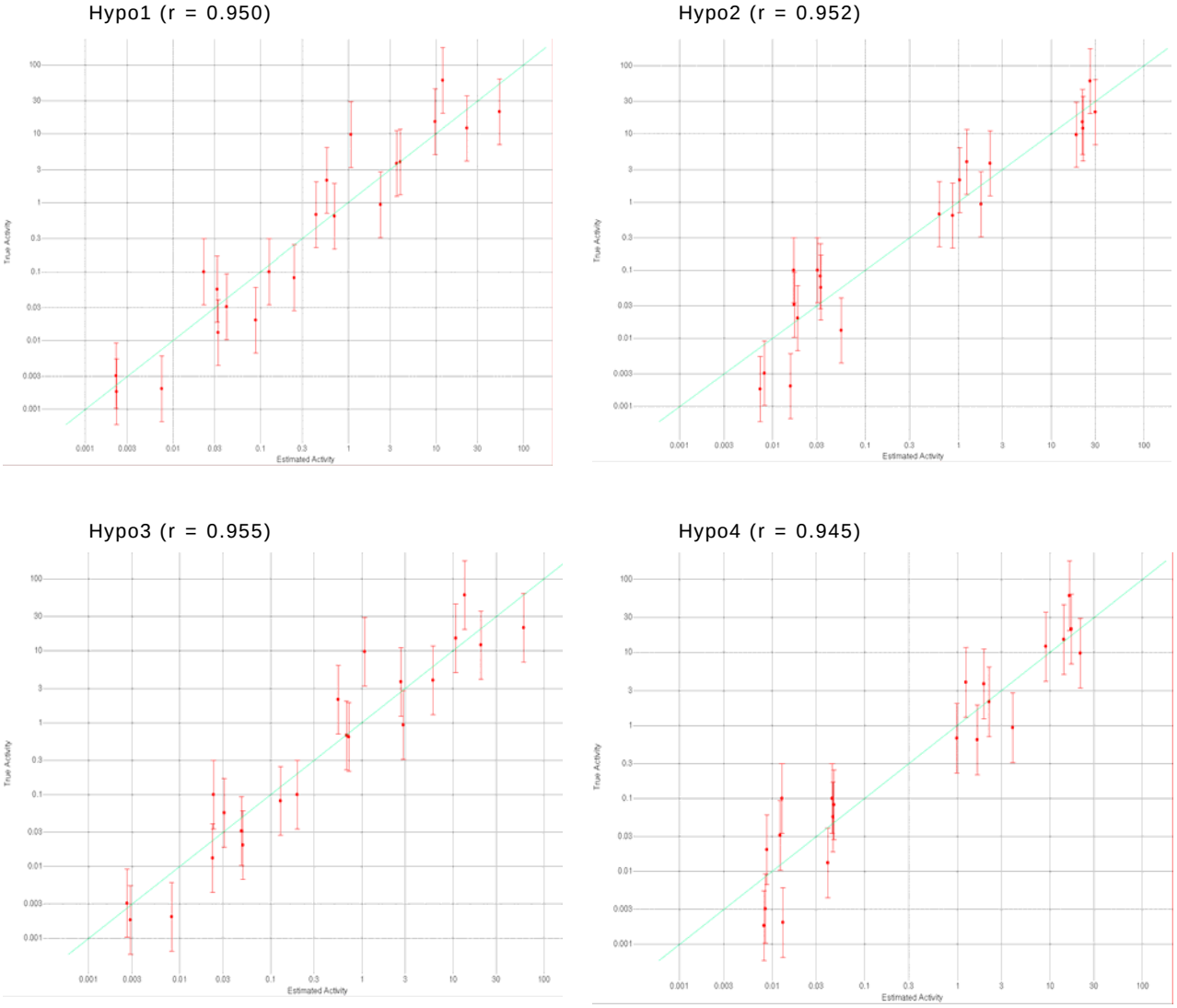
	R1-R2	IC50 (μM)
15c (training)		0.013
15e (training)		0.0020
15b (test))		0.027
15d (test)		0.019



	IC50 (μM)
LY294002 (training)	0.63

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

Figure 2. Regression of experimental versus predicted activities for the training set molecules

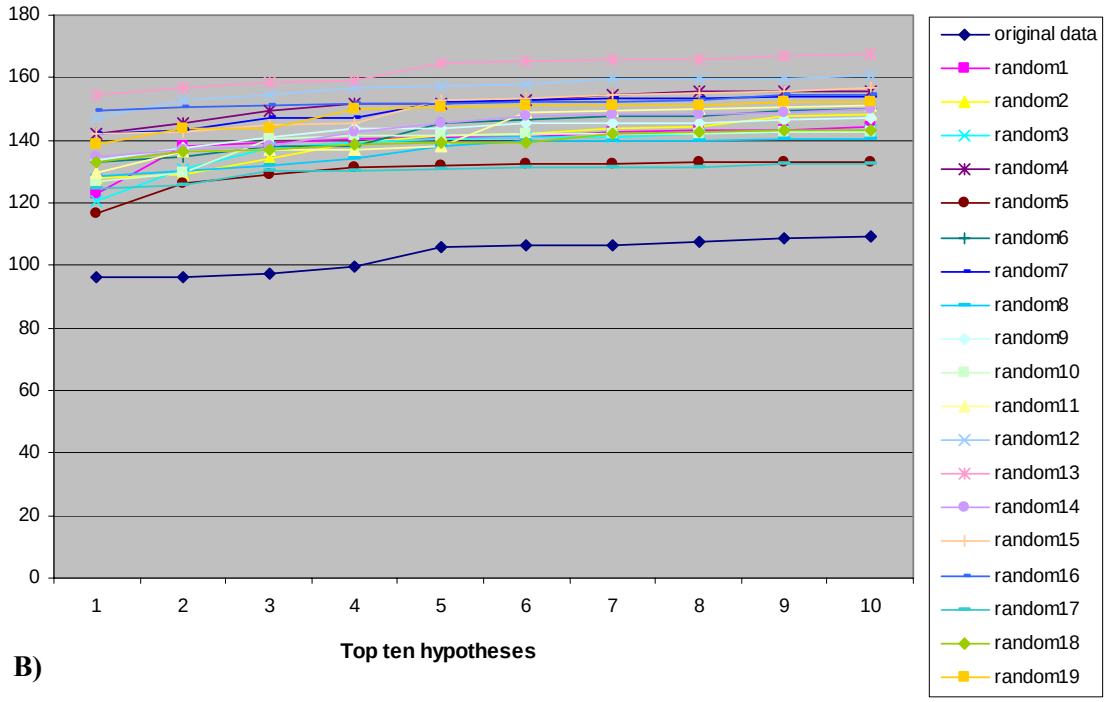


Each graph shows the linear regression that is obtained when the experimental IC50 of the different training set molecules are plotted versus the corresponding IC50 value that is predicted by appliying either Hypo1, Hypo2, Hypo3 or Hypo4.

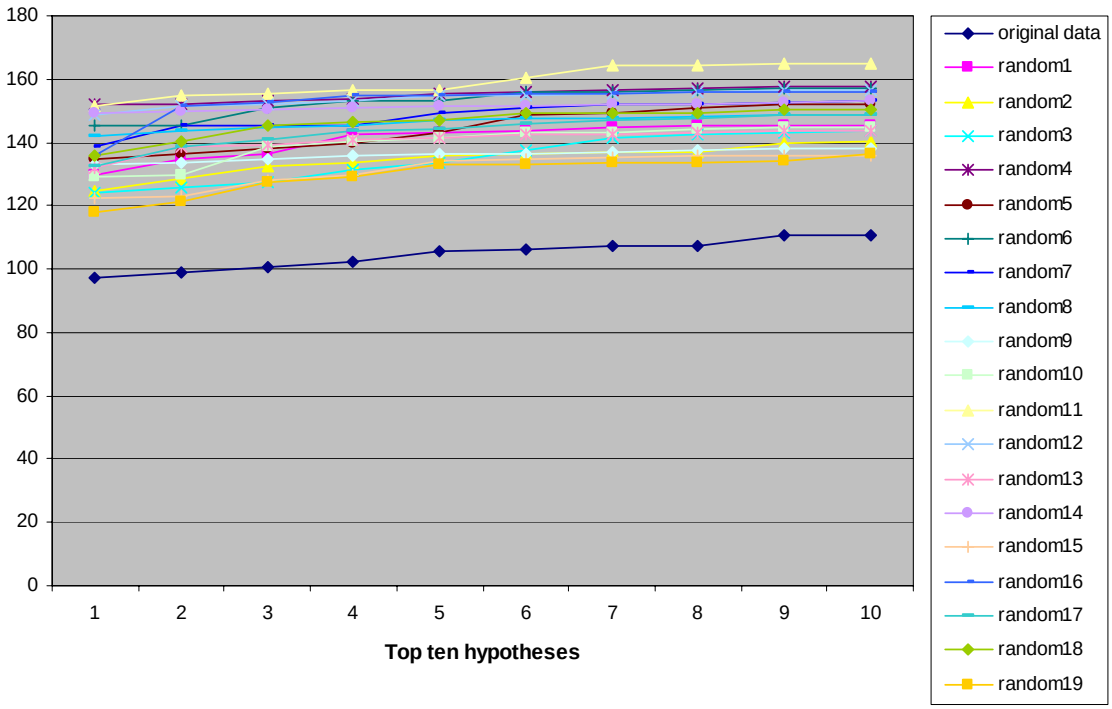
ISBN:978-84-691-1553-4 /DL: T.151-2008

Figure 3. Total cost for the top ten hypotheses obtained after HypoGen™ or HypoRefine™ runs with either the original data or after 19 random data assignments and further HypoGen™ or HypoRefine™ execution

A)

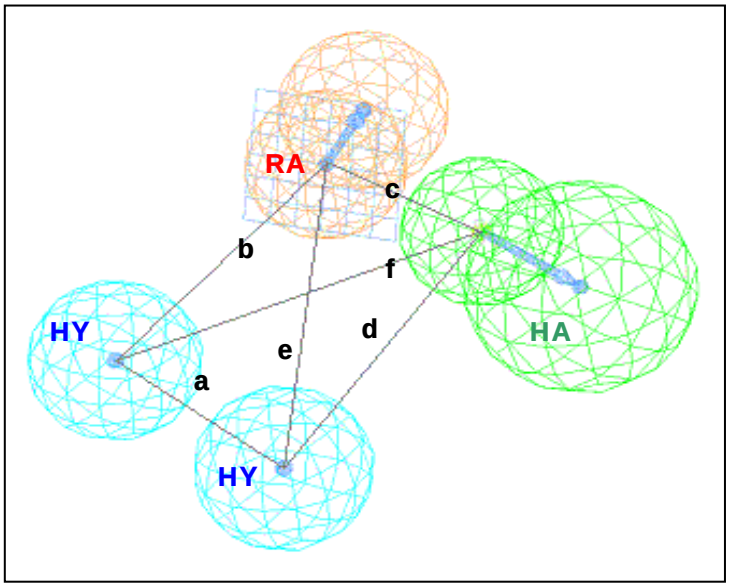


B)



Comparison of the total cost for each of the top ten hypotheses after using either the original data or 19 random-data assignments. The first and second panels show the results after the subsequent HypoGen™ or HypoRefine™ runs, have been performed, respectively. The data was randomly assigned with the catScramble™ utility from Catalyst™ and corresponds to a confidence level, of 95%. The cost for the top ten hypotheses obtained in the same run are joined by a line and color-coded with the same color.

Figure 4. Three-dimensional location of the Hypo3 pharmacophore's features



The spatial location of the pharmacophore features from the best hypothesis (*i.e.* Hypo3) is shown as color-coded spheres where: (a) green corresponds to the hydrogen-bond acceptor (*i.e.* **HA**); (b) light-blue corresponds to the two hydrophobic functions (*i.e.* **HY**); and (c) orange corresponds to the aromatic ring function (*i.e.* **RA**). Additionally, the hydrogen bond acceptor and the aromatic ring features also include a vector that shows the direction of the interaction with the receptor.

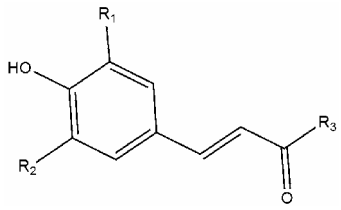
The interval of distances between the pharmacophore functions is reported in Å where **a**, **b**, **c**, **d**, **e** and **f** correspond to intervals [2.993-4.993], [5.621-7.621], [2.355-4.355], [7.058-9.058], [6.978-8.978] and [7.291-9.291], respectively.

ISBN:978-84-691-1553-4 /DL: T.151-2008

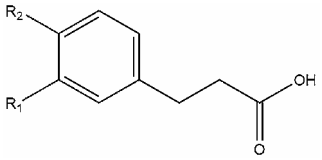
Figure 5. Molecular structures of the phenolic compounds tested

NON FLAVONOIDS

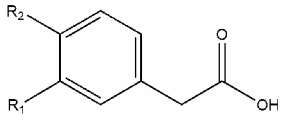
Phenolic acids



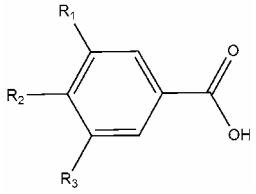
	R1	R2	R3
4-coumaric acid	H	H	OH
caffeic acid	OH	H	OH
ferulic acid	OCH3	H	OH
sinapic acid	OCH3	OCH3	OH
chlorogenic acid	H	OH	Glcde (1)



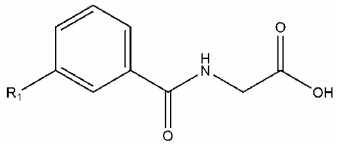
	R1	R2
propionic acid	H	H
m-hydroxyphenylpropionic acid	OH	H
p-hydroxyphenylpropionic acid	H	OH



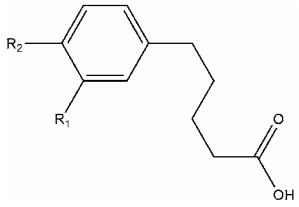
	R1	R2
3,4-dihydroxyphenilacetic acid	OH	OH
m-hydroxyphenylacetic acid	OH	H
p-hydroxyphenylacetic acid	H	OH
phenylacetic acid	H	H
homovanillic acid	OCH3	OH



	R1	R2	R3
gallic acid	OH	OH	OH
4-*O*-methylgallic acid	OH	OCH3	OH
gallic 3-glucuronide	OH	OH	Glcde (1)
gallic 4-glucuronide	OH	Glcde (1)	OH
m-hydroxybenzoic acid	OH	H	H
p-hydroxybenzoic acid	H	OH	H
protocatechuic acid	H	OH	OH
syringic acid	OCH3	OH	OCH3
vanillic acid	H	OH	OCH3

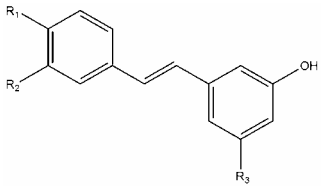


	R1
hippuric acid	H
m-hydroxyhippuric acid	OH

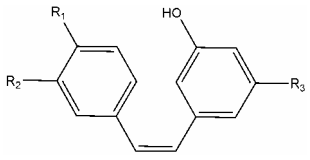


	R1	R2
m-hydroxyphenylvaleric acid	OH	H
p-hydroxyphenylvaleric acid	H	OH

Stilbenes



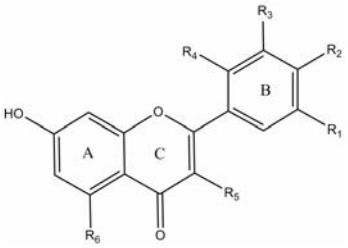
	R1	R2	R3
astringin	OH	OH	Glc (2)
piceatannol	OH	OH	OH
piceid	OH	H	Glc (2)
trans-resveratrol	OH	H	OH
resveratroloside	Glc (2)	H	OH
trans-resveratrol 3-glucuronide	OH	H	Glcde (1)
trans-resveratrol 4'glucuronide	Glcde (1)	H	OH
trans-resveratrol 3-sulfate	OH	H	SO4



	R1	R2	R3
cis-resveratrol 3-glucuronide	OH	H	Glcde (1)

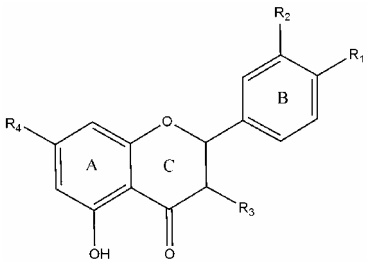
FLAVONOIDS

Flavonols



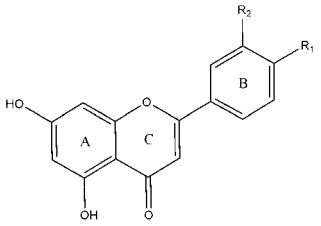
	R1	R2	R3	R4	R5	R6
fisetin	H	OH	OH	H	OH	H
galangin	H	H	H	H	OH	OH
isorhamnetin	H	OH	OCH3	H	OH	OH
kaempferol	H	OH	H	H	OH	OH
morin	H	OH	H	OH	OH	OH
myricetin	OH	OH	OH	H	OH	OH
quercetin	H	OH	OH	H	OH	OH
3'-*O*-methylquercetin 3-glucuronide	H	OH	OCH3	H	Glcde [(1)]	OH
quercetin 3-glucuronide	H	OH	OH	H	Glcde [(1)]	OH
quercetin 3-sulfate	H	OH	SO4	H	OH	OH
kaempferol 3-glucuronide	H	OH	OCH3	H	Glcde [(1)]	OH

Flavanones



	R1	R2	R3	R4
taxifolin	OH	OH	OH	OH
hesperetin	OCH3	OH	H	OH
naringenin	OH	H	H	OH
hesperetin 3'-glucuronide	OCH3	Glcde [(1)]	H	OH
hesperetin 7-glucuronide	OCH3	OH	H	Glcde [(1)]

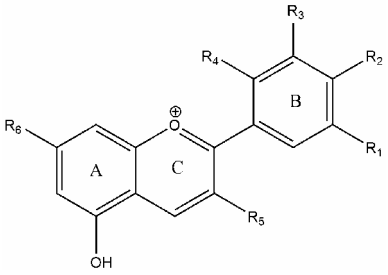
Flavones



	R1	R2
apigenin	OH	H
chrysin	H	H
diosmetin	OCH3	OH
luteolin	OH	OH

ISBN:978-84-691-1553-4 /DL: T.151-2008

Anthocyanidins

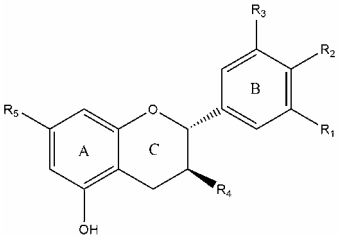


	R1	R2	R3	R4	R5	R6
cyanidin	H	OH	OH	H	OH	OH
delphinidin	OH	OH	OH	H	OH	OH
malvidin	OCH3	OH	OCH3	H	OH	OH
peonidin	OCH3	OH	H	H	OH	OH
petunidin	OH	OH	OCH3	H	OH	OH
cyanidin 3-glucoside	H	OH	OH	H	Glc (2)	OH
cyanidin 3-xyloside	H	OH	OH	H	Xyl (3)	OH
cyanidin 3-sambubioside	H	OH	OH	H	Sam (4)	OH
cyanidin 3-glucuronide	H	OH	OH	H	Glcde (1)	OH
cyanidin 3'-glucuronide	H	OH	OH	H	OH	OH
cyanidin 4'-glucuronide	H	OH	OH	H	OH	OH
cyanidin 7-glucuronide	H	OH	OH	H	OH	OH
cyanidin 3-glucoside 3'-glucuronide	H	OH	Glc (2)	H	OH	OH
cyanidin 3-glucoside 4'-glucuronide	H	Glc (2)	OH	H	Glc (2)	OH
cyanidin 3-glucoside-7-glucuronide	H	OH	OH	H	Glc (2)	Glcde (1)
delphinidin 3-glucoside	OH	OH	OH	H	Glc (2)	OH
malvidin 3-glucoside	OCH3	OH	OCH3	H	Glc (2)	OH
pelargonidin	H	OH	H	H	OH	OH
pelargonidin 3-glucoside	H	OH	H	H	Glc (2)	OH
pelargonidin 3-glucuronide	H	OH	H	H	Glcde (1)	OH
pelargonidin 4'glucuronide	H	Glcde (1)	H	H	OH	OH
pelargonidin 7-glucuronide	H	OH	H	H	OH	Glcde (1)
peonidin 3-glucoside	OCH3	OH	H	H	Glc (2)	OH
peonidin 3-sambubioside	OCH3	OH	H	H	Sam (4)	OH
peonidin 3-glucuronide	OCH3	OH	H	H	Glcde (1)	OH
peonidin 4'-glucuronide	OCH3	Glcde (1)	H	H	OH	OH
peonidin 7-glucuronide	OCH3	OH	H	H	OH	Glcde (1)
petunidin 3-glucoside	OH	OH	OCH3	H	Glc (2)	OH

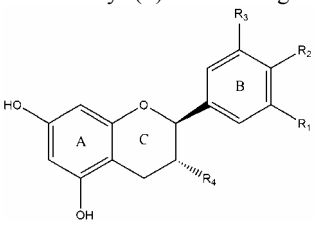
UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

Flavanols

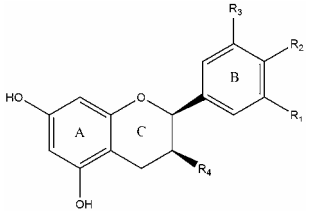
Monomers (*catechins*)



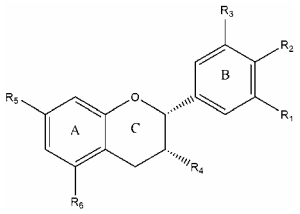
	R1	R2	R3	R4	R5
(+)-catechin	H	OH	OH	OH	OH
(+)-gallocatechin	OH	OH	OH	OH	OH
(+)-catechin 3-gallate	H	OH	OH	Gal [5]	OH
(+)-catechin 3'-glucuronide	H	OH	Glcde [1]	OH	OH
(+)-catechin 3-glucuronide	H	OH	OH	Glcde [1]	OH
(+)-catechin 4'-glucuronide	H	Glcde [1]	OH	OH	OH
(+)-catechin 7-glucuronide	H	OH	OH	OH	Glcde [1]
3'-*O*-methyl-(+)-catechin	H	OH	OCH3	OH	OH
3'-*O*-methyl-(+)-catechin 3-glucuronide	H	OH	OCH3	Glcde [1]	OH
3'-*O*-methyl-(+)-catechin 4'-glucuronide	H	Glcde [1]	OCH3	OH	OH
3'-*O*-methyl-(+)-catechin 7-glucuronide	H	OH	OCH3	OH	Glcde [1]



	R1	R2	R3	R4
(+)-epicatechin	H	OH	OH	OH
(+)-epigallocatechin	OH	OH	OH	OH
(+)-epicatechin 3-gallate	H	OH	OH	Gal [5]



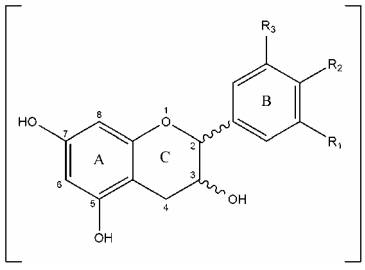
	R1	R2	R3	R4
(-)-catechin	H	OH	OH	OH
(-)-gallocatechin	OH	OH	OH	OH
(-)-catechin 3-gallate	H	OH	OH	Gal [5]



	R1	R2	R3	R4	R5	R6
(-)-epicatechin	H	OH	OH	OH	OH	OH
(-)-epigallocatechin	OH	OH	OH	OH	OH	OH
(-)-epicatechin 3-gallate	H	OH	OH	Gal [5]	OH	OH
(-)-epicatechin 3'-glucuronide* (9)	H	OH	Glcde [1]	OH	OH	OH
(-)-epicatechin 3-glucuronide* (9)	H	OH	OH	Glc [2]	OH	OH
(-)-epicatechin 4'-glucuronide* (9)	H	Glcde [1]	OH	OH	OH	OH
(-)-epicatechin 7-glucuronide* (9)	H	OH	OH	OH	Glcde [1]	OH
4'-*O*-methyl-(-)-epicatechin	H	OCH3	OH	OH	OH	OH
4'-*O*-methyl-(-)-epicatechin 3-glucuronide	H	OCH3	OH	Glcde [1]	OH	OH
4'-*O*-methyl-(-)-epicatechin 5-glucuronide	H	OCH3	OH	OH	OH	Glcde [1]
4'-*O*-methyl-(-)-epicatechin 7-glucuronide	H	OCH3	OH	OH	Glcde [1]	OH
3'-*O*-methyl-(-)-epicatechin 7-glucuronide	H	OH	OCH3	OH	Glcde [1]	OH
4'-*O*-methyl-(-)-epigallocatechin	OH	OCH3	OH	OH	OH	OH
(-)-epigallocatechingallate (EGCG)	OH	OH	OH	Gal [5]	OH	OH

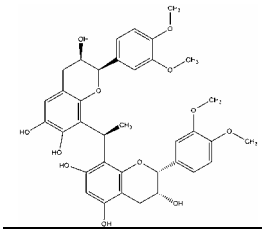
ISBN:978-84-691-1553-4 /DL: T.151-2008

Procyanidins (dimers, trimers and tetramers)



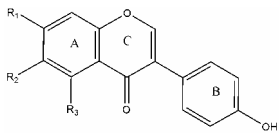
β = R configuration
α = S configuration

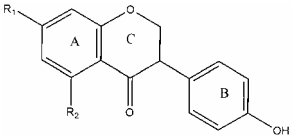
Dimers	
A1	(-)-epicatechin-(4β->8, 2β-*O*-7)-(+)-catechin
A2	(-)-epicatechin-(4β->8, 2β-*O*-7)-(+)-epicatechin
B1	(-)-epicatechin-(4β->8)-(+)-catechin
B2	(-)-epicatechin-(4β->8)-(-)-epicatechin
B3	(+)-catechin-(4α->8)-(+)-catechin
B4	(+)-catechin-(4α->8)-(-)-epicatechin
B5	(-)-epicatechin-(4β->6)-(-)-epicatechin
B6	(+)-catechin-(4α->6)-(+)-catechin
B7	(-)-epicatechin-(4β->6)-(+)-catechin
B8	(+)-catechin-(4α->6)-(-)-epicatechin
Trimers	
C1	(-)-epicatechin-(4β->8)-(-)-epicatechin-(4β->8)-(-)-epicatechin
C2	(+)-catechin-(4α->8)-(+)-catechin-(4α->8)-(+)-catechin
T2	(-)-epicatechin-(4β->8)-(-)-epicatechin-(4β->8)-(+)-catechin
T3	(-)-epicatechin-(4β->8)-(-)-epicatechin-(4β->6)-(+)-catechin
T4	(-)-epicatechin-(4β->6)-(-)-epicatechin-(4β->8)-(-)-epicatechin
T5	(-)-epicatechin-(4β->8)-(-)-epicatechin-(4β->6)-(-)-epicatechin
T6	(-)-epicatechin-(4β->6)-(-)-epicatechin-(4β->8)-(+)-catechin
T7	(+)-catechin-(4α->8)-(+)-catechin-(4α->8)-(-)-epicatechin
Tetramers	
1	(-)-epicatechin-(4β->8)-(-)-epicatechin-(4β->8)-(-)-epicatechin-(4β->8)-(-)-epicatechin
2	(-)-epicatechin-(4β->8)-(-)-epicatechin-(4β->8)-(-)-epicatechin-(4β->8)-(+)-catechin



tetramethylated dimeric

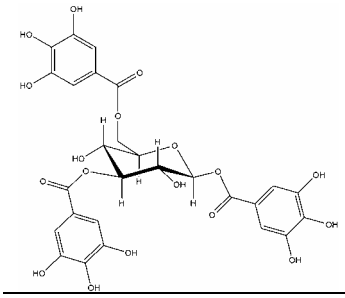
Isoflavones



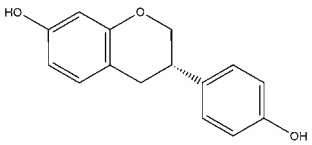


	R1	R2	R3		R1	R2
genistein	OH	H	OH	dihydrogenistein	OH	OH
genistin	Glc [2]	H	OH	dihydrodaidzein	OH	H
daidzein	OH	H	H			
daidzin	Glc [2]	H	H			
glycitein	OH	OCH3	H			
glycitin	Glc [2]	OCH3	H			

OTHER COMPOUNDS



tannic acid



equol

[1] Glucuronide
[2] Glucose
[3] β-xylose
[4] Sambubiose
[5] Gallate

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

Farnesoid X Receptor (FXR)

IV

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

In silico prediction of the activator activity of naturally occurring and bioactive forms of phenolic compounds on the Farnesoid X Receptor

Montserrat Vaqué, Esther Sala, Josep Maria del Bas, Anna Ardévol, Cinta Bladé, M. Josepa Salvadó, Mayte Blay, Juan Fernández-Larrea, Lluís Arola, Gerard Pujadas*

Departament de Bioquímica i Biotecnologia. Universitat Rovira i Virgili,
C/ Marcel·lí Domingo s/n, Campus de Sescelades.
Tarragona 43007, Catalonia (Spain)

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

ABSTRACT

Previous results from our group have shown that the co-incubation of grape seed procyanidin extract (GSPE) with CDCA, a natural ligand of FXR, enhances transcriptional activity of FXR/RXR. This snergy is GSPE dose-dependent and the increase when cells are incubated with 100 mg/L of GSPE and CDCA is two-fold that of treatment with CDCA alone. Other results from the same study suggest that: (1) one of the targets of GSPE is FXR; and (2) GSPE can enhance FXR activity only when the nuclear receptor is activated by CDCA. Therefore, at least some of the dietary phenolic compounds enhance bile acid-bound FXR activity and decrease triglyceride levels by signalling through FXR. Thus, it seems necessary to: (1) evaluate the individual capacity of phenolic compounds to act together with CDCA as FXR agonists; and (2) find a mechanism that explains this synergy. In order to achieve both goals, we have made a 3D quantitative structure-activity relationship study (*i.e.* 3D-QSAR study) that uses the structure and the activity of a large set of synthetic non-steroidal FXR agonists to predict the EC50 (*i.e.* the concentration that activates FXR activity to 50%) for the phenolic compounds most commonly found in plant extracts and for the bioactive structures of the phenolic compounds that are detected in plasma or urine. The resulting model (*i.e.* a 3D representation which identifies functional groups that contribute either positively or negatively to activity) helps to reveal the main functional determinants of non-steroidal FXR agonists and shows that most phenolic compounds (the exception are four bioactive forms) do not fit in the hydrophobic function. This hydrophobic function is described as a region in the 3D-QSAR model that favours activity. It is essential to the active FXR's conformation since it places helix 12 in an active position, which enables co-activator peptides to be recruited. Therefore, on the basis of the pattern of interactions of non-steroidal compounds and on the fact that the natural ligands (*e.g.* CDCA) show some differences from this pattern, we suggest that two molecules (*i.e.* one CDCA and one phenolic compound) occupy the ligand-binding site from FXR concomitantly. Thus, the experimentally-observed synergy would be the result of the favorable interactions that also provide the phenolic compounds for stabilizing the FXR's active conformation and enhance its transcription activity. Our results also suggest that whereas naturally-occurring phenolic compounds can be only moderately active FXR agonists (the predicted activity is highest for such procyanidins as the B5 dimer and the C1, T3, T4, T5 and T6 trimers), four bioactive forms (*i.e.* hesperetin 7-glucuronide, 4'-*O*-methyl-(-)-epicatechin 5-glucuronide, 4'-*O*-methyl-(-)-epicatechin 7-glucuronide and the tetramethylated dimeric) are moderately or highly active FXR agonists.

ISBN:978-84-691-1553-4 /DL: T.151-2008

ABBREVIATIONS

CDCA: Chenodeoxycholic acid; ECDCA: Ethyl-chenodeoxycholic acid; EGCG: Epigallocatechin gallate; FEX: Fexaramine; FXR: Farnesoid X Receptor; GRIP-1: Glucocorticoid Receptor Interacting Protein-1; GSPE: Grape-Seed Procyanidin Extract; LBD: Ligand Binding Domain; NR: Nuclear Receptor; PLS: Partial Least-Square; QSAR: Quantitative Structure Relationship; RMSE: Root Mean-Square Error; RXR: Retinoic X Receptor; SD: Standard Deviation; SD: Standard Deviation; SHP: Small Heterodimer Partner

KEYWORDS

PHASE, protein-ligand docking; eHiTS, procyanidin extract, SAR study, 3D-QSAR study, EC50 prediction

SUPPLEMENTARY INFORMATION

Supplementary materials and additional information for this paper are available at the following URL: http://www.quimica.urv.cat/~pujadas/FXR_01

INTRODUCTION

Nuclear receptors (NRs) have been proposed as targets for the treating such disorders as diabetes, atherosclerosis, osteoporosis and cancer because they have the capacity to modulate a wide batttery of genes [1-3]. In particular, the farnesoid X receptor (FXR) is emerging as a particularly attractive target for treating cholesterol- and bile acid-related diseases [4, 5]. Thus, the identification of a subset of genes regulated by FXR activation has made it possible to discover an elaborate FXR-regulatory cascade that maintains cholesterol and bile acid homeostasis [6]. It has also been suggested that there may be crosstalk between bile-acid metabolism, triglyceride metabolism, and insulin resistance [6]. Thus, FXR is an interesting pharmacological target for liver diseases and related-metabolic disorders such as liver fibrosis, cholestasis, or atherosclerosis.

FXR belongs to a large family of transcription factors (48 members in humans) which share several structural features, including an N-terminal (a highly conserved DNA-binding domain) and a C-terminal ligand-binding domain (LBD). The active and inactive states of the NRs are achieved by small ligands binding to the latter domain [7]. This binding changes the conformation of the LBD and places the activation function-2 (AF-2) helix (*i.e.* helix 12) in an *active* position which allows the recruitment of co-regulatory proteins that act either as co-activators (which contain short NR-binding motifs with the consensus sequence LXXLL, and which are also known as NR boxes) or as co-repressors (which overlap a region of the NR-LBD) [8]. These co-regulators promote or repress the transcription through a complex mechanism that involves the modulation of the chromatin structure. Accordingly, the activity of a particular agonist depends on its ability to stabilize an active conformation while antagonists prevent the AF-2 helix from occupying the active position [8]. On the other hand, partial agonists show a limited capability of promoting an active conformation [8].

The first known activator of FXR was farnesol [9]. Subsequently, a variety of endogenous bile acids were also identified. These findings revealed the key role played by FXR in a variety of physiological processes and pathological conditions related to the regulation of bile acid and lipid metabolism [10-12]. For instance, several bile acid analogs have been synthesized as hypocholesterolemic agents [4, 13]. Moreover, FXR can also be activated by a number of non-steroidal and steroidal compounds that are not structurally related to bile acids [14, 15]. Thus, the most potent endogenous FXR biliar-acid activator is chenodeoxycholic acid (CDCA) [10, 16], but there are others: for example, the synthetic steroid agonist 6-ECDCA [17] and the non-steroidal agonists GW4064 [18] and fexaramine (FEX) [19] (see Figure 1).

Recently, it has been proposed that phenolic compounds in the diet (mainly procyanidins) act as putative FXR ligands or modulators [20, 21]. Del Bas et al. (2005) observed that the small Heterodimer Partner (SHP), an orphan NR that lacks the DNA binding domain, was upregulated in the liver of rats treated with grape seed procyanidin extract (GSPE) concomitantly with the reduction of plasma triglycerides and apolipoprotein B levels [20]. They also demonstrated that SHP is a key mediator of the hypotriglyceridemic actions of procyanidins in HepG2 cells and mice. Since activated transcription of SHP is under the control of FXR/RXR and ERα, procyanidins might act as ligands of FXR, RXR, or ERα. In fact, other results (that will be published elsewhere) supported that one of the putative targets of GSPE is FXR [21]. Thus, the GSPE co-incubation with CDCA, a natural ligand of FXR, enhanced transcriptional activity of FXR/RXR. This snergy is GSPE dose-dependent and the increase when cells are incubated with 100 mg/L of GSPE and CDCA is two-fold that of treatment with CDCA alone. Other results from the same study show that GSPE can enhance FXR activity only when the NR is activated by CDCA [21]. Therefore, the dietary phenolic compounds seem to enhance bile acid-bound FXR activity and can lower triglyceride levels by signalling through FXR [21].

Taking into account these experimental results previously reported by our research group [20, 21], we apply *in silico* approaches to: (a) reveal how phenolic compounds enhance FXR transcriptional activity when CDCA is also present; and (b) predict the individual effect of the different phenolic compounds on FXR activity. During this study, we will assume that polyphenol ligands use the same activation mechanism that has been reported for other non-steroidal FXR agonists [19, 22, 23]. We have made a 3D quantitative structure-activity relationship study (*i.e.* 3D-QSAR study) with the goal of predicting the EC50 (*i.e.* the concentration that activates FXR activity to 50%) of: (a) the phenolic compounds most commonly found in plant extracts [24]; and (b) the bioactive structures of the phenolic compounds detected in plasma or urine [25-28]. The latter group of molecules was metabolized by mechanisms of intestinal and/or hepatic conjugation in which they are glucuronidated, methylated, glycosilated or sulfated, and/or hydrolized by the intestinal microflora. This second set of molecules was added to the study because it was thought that they could contribute to the biological effects of the phenolic compounds on FXR. In order to do the 3D-QSAR study, we obtained a model (*i.e.* a 3D representation that identifies functional groups of the ligands that contribute, positively or negatively, to FXR activity) by using data available on the EC50 and the molecular structure of a relatively large set of non-steroidal FXR agonists [19]. These compounds belong to several classes of FXR agonists, several of which are among the most potent FXR activators reported to date (*e.g.* FEX). Then, the resulting 3D-QSAR model [which described the most important kind of interactions between the receptor (*i.e.* FXR) and the

ligands (*i.e.* FXR agonists)] was used to map all the conformations that are possible for each of the phenolic compounds analyzed and predict their EC50 value.

MATERIALS AND METHODS

FXR structures used in the study

Three experimentally-determined structures of the FXR ligand binding domain (FXR-LBD; where this domain correspond to the 256-474 segment of the complete FXR sequence) are available at present in the Protein Data Bank (PDB; http://www.pdb.org). These FXR-LBD structures corresponds to 1OSV (2.5 Å resolution) and 1OT7 (2.9 Å resolution) [29] for *Rattus norvegicus* and 1OSH for *Homo sapiens* (1.8 Å resolution) [30]. According to the PDB information: (a) 1OSV is a complex of FXR-LBD with the coactivator peptide GRIP-1 and with the semisynthetic bile acid 6α-ethyl-chenodeoxycholic acid (6ECDCA); (b) 1OT7 is a complex of FXR-LBD with the coactivator peptide GRIP-1 and 6-ethyl-iso-ursodeoxycholic acid in chain A or iso-ursodeoxycholic acid in chain B; and (c) 1OSH corresponds to the complex of the 258-280 and 296-486 segment of FXR-LBD with the non-steroidal agonist FEX.

Docking studies

Docking studies were done with the eHiTS® v6.2 software package (SimBioSys Inc., Toronto, Canada; http://www.simbiosys.ca/ehits/) [31, 32]. The aim of these studies was to analyze the suitability of 1OSV and 1OSH as targets for docking the phenolic compounds and, therefore, to analyze how FXR-LBD flexibility can influence docking results. Consequently, redocking (*i.e.* docking 6ECDCA and FEX on 1OSV and 1OSH, respectively) and cross-docking (*i.e.* docking FEX and 6ECDCA on 1OSV and 1OSH, respectively) studies were performed. In order to do these studies 1OSH was first structurally superposed on 1OSV in order to obtain the FEX coordinates in the same coordinate system as 1OSV. The superposition was made with Swiss-PdbViewer/DeepView (http://expasy.org/spdbv/) [33]. The resulting FEX coordinates (together with the original 6ECDCA coordinates from 1OSV) were then used by eHiTS® to define the FXR-LBD region where the redocking and the cross-docking studies were performed (*i.e.* it is defined as the smallest box that can contain all FEX and 6ECDCA atoms plus 7 Å in each dimension). For the rest of the eHiTS® parameters and options, default values were used.

Pharmacophore development

Since our knowledge on how FXR–LBD flexibility influences ligand-binding is limited, but a sufficient number of synthetic agonists (*i.e.* ligands that bind to and activate FXR-LBD) have already been identified (see Figure 2; [19]), we will apply a ligand-based method to try to understand the experimental results obtained in our research group with GSPE and FXR [21]. This methodology consist of two approaches: (a) pharmacophore modelling; and (b) three-dimensional quantitative structural-activity relationships (*i.e.* 3D-QSAR). By these means, we can understand how the agonists interact with and activate the target. Moreover, they can both be used to identify new active ligands. For this reason, we used PHASE™ v2.5 (Schrödinger LLC, Portland, Oregon, USA; http://www.schrodinger.com/) [34] to carry out the study.

We used PHASE™ to develop a pharmacophore model (*i.e.* a spatial arrangement of chemical features common to two or more active ligands) which we propose as an explanation of the key interactions involved in ligand binding and a 3D-QSAR model to correlate the ligand structure with their activity on FXR. This was done in the following five steps: (1) building and preparing the ligands, (2) creating pharmacophore sites from a set of features, (3) finding common pharmacophores, (4) scoring the hypotheses/pharmacophores, and (5) building the 3D-QSAR models and choosing one of them.

Experimental information, ligand structure building and generation of conformers to develop the pharmacophore

In order to build the pharmacophore, we worked with a set of 112 compounds which, according to the literature, are non-steroidal FXR agonists [19] (see Figure 2). The activity data used were EC50 values that span 4 orders of magnitude (ranging from 19 to 10000 nM) and where each order of magnitude contains enough of the compounds to perform our study with PHASE™ (see Figure 2). Furthermore, all the activity values were obtained in the same experimental conditions and by the same scientific team, which ensures the coherence of their EC50 values [19]. At this point, it is worth mentioning that PHASE™ needs to use pEC50 [*i.e.* $-\log_{10}$ (EC50)] values instead of EC50 values because it considers that high activities correspond to active compounds (and low values to the inactive ones).

The development of a pharmacophore requires the 3D structures of the ligands that will be used to build it. Therefore, the 3D-structure ligands required were built and further minimized with ChemDraw Ultra™ v10.0 (CambridgeSoft Corporation, Cambridge, MA, USA; http://www.cambridgesoft.com/). These 3D structures were incorporated in PHASE™ and then cleaned by the PHASE's LipPreg™ module, which generated all possible stereoisomers by

retaining: (a) the user-specified chiralities of the ligands; and (b) their original ionization state. Moreover, in this study, we applied a flexible approach because the pharmacophore model was based on multi-conformer models (*i.e.* each molecule was represented by a set of conformations because the conformation that they adopt when they bind to FXR-LBD is unknown). Thus, the ligand conformations were generated with PHASE's MacroModel™ module by using default parameters in order to (a) avoid redundant conformers; (b) apply a fast ligand torsion search (*i.e.* the molecule is divided into a core and a periphery and then all the core conformations are generated and the conformations of the peripheral groups are sampled one by one); (c) limit the maximum number of conformers to 1000; (d) perform a minimization and a filtering out to discard conformers that are redundant (*i.e.* have a distance cut-off lower than 2 Å) or have an energy value higher than 10 kcal/mol; and (e) apply the OPLS_2005 force field and the distance-dependent dielectric model as the solvation treatment.

Pharmacophore site generation

PHASE™ identifies the chemical features (also known as *pharmacophore features*) in all the agonist structures that will be used to develop the pharmacophore (see Figure 2). In this respect, these features are searched by using patterns of chemical structures and can be classified as: (a) hydrogen bond acceptors (*i.e.* labelled as **A**); (b) hydrogen bond donors (*i.e.* labelled as **D**); (c) hydrophobic groups (*i.e.* labelled as **H**); (d) negatively charged groups (*i.e.* labelled as **N**); (e) positively charged groups (*i.e.* labelled as **P**); and (f) aromatic rings (*i.e.* labelled as **R**). Furthermore, one chemical feature located in a specific place in one conformation of one ligand is called a *pharmacophore site*. Thus, each ligand conformation is represented by a set of points in the 3D space, which are coincident with various chemical features that may facilitate non-covalent binding between the self ligand and its target receptor. Then, all pharmacophore sites found for each ligand's conformation are characterized by type, location and directionality (when it is applicable) and they are recorded for further use.

Other important points that need to be taken into account are that the chemical features identified by PHASE™ in our FXR-agonist structures must be located in the appropriate 3D location and the possibility of redefining or adding new feature types must be evaluated (*i.e.* **X**, **Y**, **Z**) (this must be done only when there is a specific or new feature in the structure of our ligand's set that we think may be important for the intermolecular interaction with the receptor). In this study, we did not redefine or add any new chemical feature types. Therefore, the features identified in our ligand set were: (a) hydrogen bond acceptors (*i.e.* **A**); (b) hydrogen bond donors (*i.e.* **D**); (c) hydrophobic groups (*i.e.* **H**); and (d) aromatic rings (*i.e.* **R**). Moreover, the chemical features are represented with three possible geometries: (a) points (where the pharmacophore site is located on a single atom); (b) vectors (where the pharmacophore site is

ISBN:978-84-691-1553-4 /DL: T.151-2008

located on a single atom and it also has an associated directionality); or (c) groups (where the pharmacophore site is located at the centroid of a group of atoms and they also have an associated directionality). Thus, hydrogen bond acceptors or donors can be represented either by using vector features (*i.e.* an acceptor is represented by indicating the atom in which it carries one or more lone-pair electrons and a vector with an arrow that points in the direction of the lone pairs; and a donor is represented by indicating the hydrogen atom that can be given and a vector feature that is directed along its idealized hydrogen bond axis and with an arrow that points in the direction of the potential lone-pair electrons) or as pure projected points that are located at the complementary positions on a theoretical binding site but without any vector feature. On the other hand, the hydrophobic feature is represented as a pure projected point and ring aromatic functions are represented by the centroid of each aromatic ring.

Common pharmacophore search

Defining the active and inactive set

In order to build the pharmacophore models (*i.e.* a combination of chemical functions in the three-dimensional space that correlates the ligand structure with its effect on the activity of the receptor to which it binds), we had to specify which compounds were active and which were inactive FXR agonists. This is important because the active molecules would then be used to generate the pharmacophore models assuming that the most active ligands either have the strongest binding interaction with FXR or contain the highest number of pharmacophore features that are involved in the binding to the protein target. For this reason, the set of ligands that is used to build the pharmacophore must contain molecules that are diverse enough in structure and activity to ensure the reliability of the pharmacophore models that will be obtained with PHASE™. Thus, if the active agonists that are used have a common receptor binding mode, the resulting pharmacophores should be able to capture the essential features of the agonist structure. On the other hand, the less active agonists would help to penalize the pharmacophore and help us to decide which is the best pharmacophore model.

Therefore, the set of agonists was divided into active or inactive according to their pEC50 values. Thus, we considered those compounds with pEC50 values higher than 7 to be active and those pEC50 lower than 6 to be inactive (see Figure 2).

Application of the tree-based partitioning technique

The goal of PHASE™ is the exhaustive identification of pharmacophores that are common to a set of active compounds and which have a specific number of pharmacophore sites. Therefore, before starting this process, we need to set up: (a) the minimum number of active compounds that are required to build the pharmacophore (but without specifying which ones will be used to

evaluate this); and (b) the number of pharmacophore sites that the common pharmacophore must include. Thus, in this study, we have reduced the total number of required active compounds and the number of pharmacophore sites to 5. At this point, it is important to point out that using this number of required active compounds enables us to relax the search and lower the requirements for all 24 active compounds in our set (see Figure 2). It is also important to use no more than 7 pharmacophore sites to ensure that the correct pharmacophore is found because each site represents a 2-3 kcal/mol energy-requirement for the intermolecular interaction with the receptor.

During the search, the pharmacophores that contained identical sets of chemical features which had a very similar spatial arrangement had to be grouped together. The resulting common pharmacophores (because they may be more than one) are obtained by using a tree-based partitioning technique that groups together the pharmacophores that are similar by using the distances between pairs of sites in the pharmacophore (*i.e.* the so-called *intersite distance*). During this process, a k-point pharmacophore (*i.e.* a pharmacophore with **k** sites) is represented by a vector of **n** distances where $n = (k \cdot (k-1))/2$ and each intersite distance **d** is filtered through a binary decision tree. Thus, PHASE™ filters pharmacophores through a series of nodes whose left and right branches (*i.e.* the binary branches of the decision tree) correspond to whether or not a particular intersite distance present in the pharmacophore is less than some threshold value. By filtering all **n** intersite distances in this manner, the pharmacophore is assigned to an n-dimensional box, whose sides are equal in length to the terminal node width. In other words, the common pharmacophores are identified from a set of variants (where a variant is a set of feature types that define a possible pharmacophore). For instance, the variant AAAHR contains three hydrogen-bond acceptors (*i.e.* **A**), one hydrophobic group (*i.e.* **H**) and one aromatic ring (*i.e.* **R**). Thus, all pharmacophores of a given variant (*e.g.* AAAHR) are enumerated and partitioned into successively smaller high-dimensional boxes according to their intersite distances and pharmacophores that are clustered into the same box are considered to be equivalent and therefore common to the ligands from which they have been obtained. The boxes that contain pharmacophores with the minimum required number of ligands are said to *survive* the partitioning process. Thus, each surviving box contains a set of common pharmacophores, one of which is ultimately singled out as a hypothesis. All these different common pharmacophores are collected to show all desired variants that have been processed. Then, these hypotheses have to be scored (see next step) in order to find which one is best.

By following the above process, we obtained a list of 19 different variants that are the result of the combinations of **A**, **H** and **R** chemical functions.

ISBN:978-84-691-1553-4 /DL: T.151-2008

Scoring hypotheses

PHASE™ includes a procedure for choosing the most appropriate hypothesis or hypotheses from the common pharmacophore or pharmacophores that have been generated. Thus, the hypotheses should: (a) explain how the active molecules bind to the receptor; and (b) be able to discriminate between active and inactive molecules. In order to identify which pharmacophore from each surviving n-dimensional box represents the best alignment of the chosen active molecules, PHASE™ examines all common pharmacophores obtained in the step described above and applies a scoring process. During this scoring process, geometric and heuristic factors (which can be weighted according to the user's preferences) are applied to provide a ranking of the different hypotheses. In other words, this scoring process is based on how well the active ligands superimpose when they are aligned on the chemical features associated with these hypotheses. Although this scoring process is based on the active compounds, PHASE™ can also penalize hypotheses that are not able to distinguish between actives or inactives compounds (therefore, the hypotheses or models identified by PHASE™ only should countain the features that are essential for high-affinity binding with the receptor). Thus, PHASE™ performs a scoring process taking into account the active compounds and it can perform also another one taking into account also the inactive compounds.

Generation of 3D-QSAR models

In order to understand the structural differences between active and inactive compounds we can do a visual inspection of the aligned ligands and quantified these differences by building a 3D-QSAR model that identifies which functional groups contribute, either positively or negatively, to activity.

Hypotheses, training and test set selection

PHASE™ generates 3D-QSAR models by using a set of molecules, which all have been aligned to a common pharmacophore that is associated with a single reference ligand and their activity data. We selected 15 hypotheses or common pharmacophores and the activity data were the EC50 values from our set of 112 non-steroidal FXR's synthetic agonists [19]. From this set of compounds, we selected 54 compounds to generate the 3D-QSAR model (*i.e.* the training set; see Figure 2) and 58 compounds that will be used to validate it (*i.e.* the test set; see Figure 2). The procedure used to select these sets was a random selection. To achieve a useful 3D-QSAR model the set of ligands have to span a range of activities as large as possible (both training and test set spans 4 orders of magitude) and have enough structural diversity (see Figure 2 for the structure and the EC50 values of the test and the training set molecules).

PHASE 's steps to build the 3D-QSAR models

PHASE™ models are 3D-QSAR models in which chemical features of ligand structures are mapped to a cubic 3D grid. For this reason, the first step in 3D-QSAR model generation is the alignment of the ligands to the set of pharmacophore features in the selected hypotheses by using a standard least-squares procedure. After that, a rectangular grid is defined to cover the space occupied by the aligned training set molecules (where this grid divides the occupied space into *N* uniformly-sized cubes; by default 1 Å on each side). Then, PHASE™ can use two alternative ways to generate the structural components that constitute the 3D-QSAR models: (1) one atoms-based that takes all atoms into account; or (2) one pharmacophore features-based that uses the pharmacophore sites that can be matched to the hypothesis.

In the atoms-based 3D-QSAR models, each ligand atom is represented by one sphere whose radius is the van der Waals radius for the corresponding atom type according to MacroModel™. Thus, atoms are divided into six classes: (a) hydrogen-bond donors (**D**) (*i.e.* hydrogens bonded to N, O, P or S); (b) hydrophobic or nonpolar (**H**) (*i.e.* C, H–C, Cl, Br, F or I); (c) negative ionic (**N**) (*i.e.* atoms with a formal negative charge); (d) positive ionic (**P**) (*i.e.* atoms with a formal positive charge); (e) electron-withdrawing (**W**) (*i.e.* N, O; includes hydrogen-bond acceptors); and (e) miscellaneous (**X**) (*i.e.* all other atom types). A given atom can occupy the space of one or more cubes in the grid (*i.e.* a cube is occupied by an atom of a particular class if the center of that cube falls within the radius of this atom). Following with this rule, each ligand is represented by a set of bit values (0 or 1) that indicate which cubes are occupied by atoms of each class. Moreover, a given cube may be occupied by more than one atom, and that occupation may come from atoms from the same or different molecules. The models based on atoms are useful when some features, other than the pharmacophores, are important to activity, such as steric clashes. Furthermore, they are also appropriate when the training set either have a reduced structural diversity or contain a relatively small number of rotatable bonds or some common structural framework.

In pharmacophore-based 3D-QSAR models, the structural components of the ligands are represented by pharmacophore features (*i.e.* by the feature definitions that were previously used to create the hypothesis: **A**, **D**, **H**, **N**, **P** and **R**; see above) with a user-defined radius. Thus, occupation of cube is now deemed to occur if the center of that cube falls within the user-defined radius of a particular pharmacophore site (that is represented by an sphere). A given cube may be occupied by more than one site (but also a given site can occupies the space of one or more cubes in the grid), and that occupation may come from the same or from different molecules. Pharmacophore-based models assume that the activity is explained entirely by the pharmacophore model itself, and therefore cannot predict activities where other features are

ISBN:978-84-691-1553-4 /DL: T.151-2008

important to activity, such as steric clashes. This kind of models are appropriate when the structures in the training set are either chemically diverse (*i.e.* include different chemical families) or highly flexible.

Once the occupancies are determined by using either the atom- or the pharmacophore-based 3D-QSAR models, a regression is performed by using the partial least squares (PLS) method [where the independent variables are the binary-valued occupancies (*i.e.* the values are either 0 or 1) of the cubes by the different structural components]. Thus, in an atom-based 3D-QSAR model, the number of independent variables used are the 6*N* occupancies of the *N* cubes by the six available atom classes (*i.e.* each variable corresponds to a given cube and a given atom class) and the value for each variable can be 0 or 1. In contrast, in a pharmacophore-based 3D-QSAR model, the number of independent variables used are the *mN* occupancies that correspond to the *N* cubes by the *m* types of available pharmacophore features (*i.e.* each variable corresponds to a given cube and a given feature type) and the value for each variable can be 0 or 1. Therefore, the regression involves finding a linear least-squares relationship between the activity data (*i.e.* the dependent variable) and a special set of orthogonal factors that are linear combinations of the bit value variables (*i.e.* the independent variables). The accuracy of the 3D-QSAR models increases with increasing the number of PLS factors until over-fitting starts to occur (where the maximum number of PLS factors is **N**/5 and **N** is the number of ligands). Thus, the PLS facilitates the identification of specific chemical features that tend to increase or decrease the estimated activity. The number of PLS used in the present study was 3.

Since the selection of the best type of 3D-QSAR model to be used in this study was not clear, we generated both types of models. After that, the statistics for the training and test set were analyzed in order to see which of the two approaches produce a 3D-QSAR model with the best predictive power. Since our training set have reduced structural diversity, the better model was obtained by using the atom-based 3D-QSAR models.

3D-QSAR validation and statistics

PHASE™ supports only the use of a true external test set (*i.e.* compounds which have not been used to build the model) to validate the 3D-QSAR model. For this reason we have to analyze the statistics obtained from the training and from the test sets. The main statistical properties that describe the 3D-QSAR model when the training set data is used are: (a) the R-squared or R^2 (*i.e.* the coefficient of determination; which never can be negative); (b) the standard deviation of regression or SD; (c) the F statistic (*i.e.* the overall significance of the model); and (d) the statistical significance or P (*i.e.* the probability that the correlation could occur by chance). Thus, in the case that the independent variables have no statistical relationship with the activity,

R^2 would be 0. On the other hand, the main statistical quantities describing the test set prediction are: (a) the q-squared or q^2 (*i.e.* equivalent to R^2 but now, using the predicted and experimental test set activity values; in contrast to R^2, it can take negative values); (b) the Pearson value or R (*i.e.* the Pearson correlation coefficient); and (c) the root-mean-square error or RMSE. At this point, it is worth to remark that there is no any single parameter that allows to choose the best model. In this sense, we have to consider all the statistic parameters reported by PHASE™ to evaluate the different 3D-QSAR models. Moreover, to do this selection, we have to consider also some other aspects like, for example: (a) the accuracy of our data (about 0.3 pEC50 log units); (b) the quantity of compounds used; and (c) evaluate which model accuracy we need to use for predicting the activiy of other compounds.

3D-QSAR model visualization

Once the 3D-QSAR models have been generated, we have to visualize and analyze them. Thus, to understand how the structures of the ligands contribute, either positively or negatively to the computed activity, we have examined the three-dimensional aspects of the 3D-QSAR model. The visualization allows to view the cubic volume elements occupied by one specific ligand or all the cubes in the 3D-QSAR model (*i.e.* the union of the cubes occupied by all the compounds from the set). In this visualization, the blue cubes indicate regions that are favorable for activity whereas the red cubes indicate regions that are unfavorable for activity.

Searching active agonists in 3D-polyphenol database

Generation of the 3D-polyphenol database

Once a pharmacophore model has been developed and it has been chosen based on its statistical properties and predictive power, it may be used for searching a database in order to find additional active molecules. In this study, our goal is the identification of phenolic compounds which accomplish also the pharmacophore that has been built with synthetic non-steroidal FXR's agonists [19]. For this reason, we created a PHASE™ 3D database with 135 phenolic compounds whose 3D structures were generated and minimizated with ChemDraw Ultra™ v10.0 (CambridgeSoft Corporation, Cambridge, MA, USA; http://www.cambridgesoft.com/). These 3D structures were cleaned by the PHASE's LipPreg™ module that generated all possible stereoisomers by retaining the original specified chiralities and the original states of ionization. Then, we used the MacroModel module of PHASE™ to generate all relevant phenolic compounds conformations because a preliminary cross-docking analysis have shown the decisive influence of FXR-LBD flexibility in docking results (and we do not have any information about the conformation of the FXR-LBD that have to be used to dock our polyphenols database). Hence, this MacroModel run was also done with default options and

ISBN:978-84-691-1553-4 /DL: T.151-2008

parameter values as it was when the conformations of the 112 synthetic non-steroidal FXR's agonists from the training and test sets were obtained. Finally, the phenolic compounds structures were recorded in the database together with their set of conformers and sites (because the sites are automatically created when the conformers are generated). Figure 3 shows the structure of the 135 phenolic compounds that are part of this 3D-database.

Searching in the 3D-polyphenol database

Taking into account that the chemical functions of the 3D-QSAR model describe the most important kind of interactions between the receptor (*i.e.* the FXR-LBD) and the ligands (*i.e.* non-steroidal FXR's agonists [19]), the model could be also used to estimate the activities for the phenolic compounds (assuming that they also could interact with FXR by using the same mechanism as the molecules in the training and test sets; see Figure 2). Therefore, our in-house 3D-polyphenol database has been searched for conformations that match the features of the 3D-QSAR model. In order to proceed with this search, PHASE™ does two steps: *finding* and *fetching*. Thus, in the first step (*i.e.* finding), the database is searched for geometric arrangements of pharmacophore sites that match the site types of the chosen hypothesis and whose intersite distances are *sufficiently* close to them (*i.e.* no further than 2.0 Å). In our case, the hypothesis AAAHR contains three hydrogen-bond acceptors (*i.e.* **A**), one hydrophobic group (*i.e.* **H**) and one aromatic ring (*i.e.* **R**) and, therefore, these five pharmacophore features give rise to ten unique intersite distances (*i.e.* dA_1A_2, dA_1A_3, dA_2A_3, dA_1H, dA_2H, dA_3H, dA_1R, dA_2R, dA_3R, dHR) Consequently, when an occurrence is found in the database (*i.e.* a hit is found), the information about the match is written to a *match file*. Then, in the fetch step, this match file is used as a lookup table to rapidly retrieve the relevant conformers (or hits) from the database and align them to the hypothesis. After hits are fetched, they are ordered according to their decreasing *fitness* score and filtered, so that only a fraction of the total number of matches is further used [where the fitness score measures how well the hit's pharmacophore sites align with those of the hypothesis; how well the hit's vector features (*i.e.* hydrogen-bond acceptors or donors and aromatic rings) overlay with those of the hypothesis; and how well the hit's conformation superimposes, in an overall sense, with the reference ligand conformation]. Finally, the activity of the filtered hits is predicted based on the 3D-QSAR model available.

Since most of the phenolic compounds are not able to simultaneously match in all the 5 sites, we have performed also a *partial matching*. So, we chose to match fewer sites (*i.e.* 3 or 4) than the number in the hypothesis (*i.e.* 5). Accordingly, in that situation, the resulting fitness score has been modified to penalize the hits that do not match all sites.

RESULTS AND DISCUSSION

Docking analysis

At first, a redocking study was made with the PDB complexes 1OSV and 1OSH to test whether eHiTS® can reproduce the experimental conformations of 6ECDCA and FEX in the proteic context of their complexes with FXR-LBD. Therefore, 6ECDCA and FEX were redocked in the proteic parts of 1OSV and 1OSH, respectively (it is worth pointing out that care was taken in using starting conformations for both ligands that were different from the ones that are found in their experimental complexes with FXR-LBD). The results obtained from the redocking show that, under the conditions of these calculations, eHiTS® can reproduce the experimental conformation of both ligands on its FXR-LBD crystallized structures.

We also analyzed how the slightly different FXR-LBD conformations in 1OSV and 1OSH can affect docking results. To do so, we performed a cross-docking study. Therefore, the ligand from 1OSV (*i.e.* 6ECDCA) was docked onto the proteic structure from FXR-LBD in 1OSH and, likewise, the ligand from 1OSH (*i.e.* FEX) was docked onto the proteic structure from FXR-LBD in 1OSV. This cross-docking provided valuable information about the effects of induced fit upon ligand-binding. In this respect, our results confirmed that the FXR-LBD conformational changes that are induced by each ligand are ligand-structure dependent. Accordingly, FEX was unable to interact in the ligand binding site of 1OSV (which corresponds to the induced-fit structure of FXR-LBD upon 6ECDCA binding) and 6ECDCA was unable to fit in the ligand binding site of 1OSH (which corresponds to the induced-fit structure of FXR-LBD upon FEX binding). Therefore, these results show that docking would be a feasible method for obtaining the conformations for the FXR non-steroidal agonists that have been described in the literature [19] only when the FXR-LBD from 1OSH is used as the target of the docking process. In this respect, a 3D-QSAR study has been published recently for a serie of FXR non-steroidal agonists in which their conformations in the ligand-binding site from the FXR-LBD were generated by protein-ligand docking [22]. In this study, docking results were classified in three sets because, apart from those ligands in which docking fails to find any solution, some docking conformations have a pose that is similar to the experimental binding of FEX while others have a pose that disagrees with the experimental results for FEX. Thus, the 3D-QSAR model was built by using only the set of compounds which had a binding to FXR-LDB that is similar to the one found for FEX (interestingly, the less active FXR agonists were not in this set) [22]. Since we wish to evaluate compounds that may have different ligand bindings to that of FEX, we cannot exclude from our study compounds that have low activities or are inefficient even though the docking process may discount them subsequently. Moreover, taking into account the importance for the FXR function of the changes that take place in the FXR-LBD upon ligand-

binding and that these changes are ligand-structure dependent, there is no guarantee that phenolic compounds that are very different to 6ECDCA or FEX will find their correct poses when docked onto the proteic part of 1OSH or 1OSV. At this point, it is worth pointing out that eHiTS® considers the protein target as a rigid body during docking and that other docking programs that are able to include induced-fit changes in the target protein (*i.e.* Glide™/Prime™ [35]) are more focused on reproducing loop changes induced by ligand-binding and not a general change in the complete target structure as happens with FXR-LBD. Therefore, in our study, all the phenolic compound conformations needed were generated with the program PHASE™ and not by docking.

Generation of hypotheses with PHASE™ and analysis of the selected hypothesis in the context of the FXR-FEX experimental complex

In all 112 FXR agonists used to generate the hypotheses [19], PHASE™ was able to identify the following chemical features: (a) hydrogen bond acceptor (*i.e.* **A**); (b) hydrophobic group (*i.e.* **H**); and (c) aromatic ring (*i.e.* **R**). From this initial set, those with a pEC50 higher than 7 were considered to be active (*i.e.* 24 molecules; see Figure 2). Since the active agonists used had a common receptor binding mode, the resulting pharmacophore models should capture the essential features of the agonist structure. From this set of active agonists we obtained a list of 19 different variants that are the result of the combinations of the **A**, **H**, and **R** chemical functions (present in 5 of the 24 active agonists). The hypothesis that we chose belong to a box that had survived the partitioning process and which was characterized by five sites with the variant AAAHR (*i.e.* the hypothesis contained three hydrogen bond acceptors, and one hydrophobic and one aromatic ring) at a specific intersite distance. This hypothesis was chosen on the basis of the active and the inactive compounds (*i.e.* the less active compounds), because the latter were used to penalize. Thus, we chose the hypothesis with the highest survival score after the penalization with the inactive compounds. This hypothesis is shown in Figure 4 in the context of the FEX agonist. Within this context it is possible to see that: (a) the hydrogen bond acceptors map FEX's methyl ester moiety and amide carbonyl oxygen; (b) the aromatic ring maps FEX's outermost benzene ring; and (c) FEX's hydrophobic group maps the dimethylamine moiety. Since we show the hypothesis in the context of FEX, we can compare the chemical features of our hypothesis with the interactions between FEX and the FXR-LBD that are described by the authors of 1OSH (*i.e.* the complex between FEX and FXR-LBD) [30]. In this paper, the interactions were divided between FEX and FXR-LBD in two subsets, which suggests that FEX's potency appears to be mediated by two mechanistic paths [30] (see Figure 5).

Thus, the first subset contains those interactions that stabilize the position of the following FEX groups: (a) the hexyl ring; (b) the outermost first benzene ring; and (c) the methyl ester moiety (see Figure 5A). Thus these interactions are: (a) some minimal van der Waals contacts between the hexyl group and residues around it (*i.e.* Ile339 and Leu344 from helix 5); (b) a hydrophobic surface created by three apolar residues (*i.e.* Met369 and Phe370 from helix 7 and Phe333 from helix 5) behind FEX's central nitrogen and the outermost benzene ring; (c) a hydrophobic surface formed by three apolar amino residues (*i.e.* Leu352 and Ile356 from helix 6 and Met294 from helix 3) and the aliphatic linker between the last benzene ring and the methyl ester moiety; and (d) two hydrogen bonds in the amide carbonyl oxygen made up of residues His298 and Ser336, which stabilize the position of the methyl ester moiety in the neutral groove between helices 3 and 6 (see Figure 6 for the relative helix location in the ligand binding site of FXR-LBD). In our hypothesis (see Figure 4) some of the interactions in this subset are described by: (a) one hydrogen bond acceptor function located at the amide carbonyl oxygen (which includes two vectors and so may perform two hydrogen bonds) and which corresponds to FXR-LBD interaction with FEX described above for residues His298 and Ser336; and (b) one aromatic ring in the outermost benzene ring (which corresponds to FXR-LBD interaction with FEX by means of residues Met294, Leu352 and Ile356).

The second subset of interactions that has been described between FXR-LBD and FEX is thought to be responsible for the stabilization of the biaryl rings and of the dimethylamine moiety of FEX [30] (see Figure 5B). In this respect, it has been reported that the sequential hydrophobic ring structures of FEX penetrate deeper into the ligand binding pocket and increase the number of stable contacts with the LBD [30]. Thus, the interactions that form this second subset are: (a) two hydrophobic surfaces (one at each side of FEX's double ring structure), one of which consist of four residues from helix 3 (*i.e.* Phe288, Leu291, Thr292 and Ala295) and the other of which consist of five residues [four from helix 11 (*i.e.* His451, Met454, Leu455 and Trp458) and one from loop7 (*i.e.* Ile361)]; and (b) a deep hydrophobic pocket that is filled by the biaryl moiety and which is formed by the mentioned hydrophobic surfaces from helix 11 and helix 3 that are bridged by Leu469 and Trp473 from helix 12 and by Phe465 from loop 12. In agreement with this second subset of interactions, our hypothesis proposes a hydrophobic function in the area occupied by one methyl from FEX's dimethylamine moiety (see Figure 4).

Therefore, these results show that our hypothesis can capture the main interactions that have been described for one of the strongest activators of FXR-mediated transcriptional activity (*i.e.* FEX) [30].

ISBN:978-84-691-1553-4 /DL: T.151-2008

Table 1 shows the chemical functions that have been detected in all 112 FXR agonists (either from the training or the test set). The analysis of the results shows that all these molecules (except compounds 136 and 138 from the test set) map the three hydrogen bond acceptors and the ring aromatic functions of the hypothesis. In contrast, the hydrophobic function is not mapped by a significant number of compounds. The lack of the hydrophobic function on two of the most active agonists (*i.e.* 244 and 245; see Table 3) can be attributed to the fact that the hydrophobic site is mapped in both ligands by a cyclic acetal group (see Figure 2), which is not defined as a hydrophobic function by PHASE™.

3D-QSAR model generated by PHASE™

Although the analysis of the alignment of the most and least active compounds (see Figure 7) enabled some common chemical features and functional groups to be identified, we were not able to discern their individual contribution (either positive or negative) to their activity as FXR agonists. For this reason, a 3D-QSAR model was developed.

3D-QSAR model validation

The performance of the three-factor 3D-QSAR model on the training and test set molecules is illustrated in Figure 8 and the results of the statistical parameters obtained for both sets enable it to be validated. Thus, the training set correlation was obtained with three partial least-square (PLS) factors and is characterized by: (a) SD = 0.18; and (b) R^2 = 0.91. These values indicate a good aggreement between predicted and experimental activities (see Figure 8). Moreover, if the predicted and real pEC50 of the training set compounds are classified on an activity scale [*i.e.* "+++" for highly active molecules (pEC50 > 7.0 nM); "++" for moderately active molecules (6.0 nM ≤ pEC50 ≤7.0 nM); and "+" for inactive compounds (pEC50 < 6.0 nM)] then, the 3D-QSAR model correctly predicts the activity category in 43 of the 54 cases (75%) of the training set (and their activity values are predicted within 0.5 log units of the experimental values in all cases; see Table 1). Thus, this confirms that this 3D-QSAR model is a reliable candidate for describing the structure-activity relationships within the training set.

Nevertheless, in order to be of use for our activity prediction goals, it is necessary to demonstrate that this model can also predict the activities of other molecules outside the training set. Therefore, we evaluated this capacity by predicting the activities of 58 other compounds (*i.e.* the so called *test set*; see Figure 2), which are structurally different from the training-set molecules but which have activity values that were obtained in the same experimental conditions as this latter set [19]. Thus, in order to validate the predictive power of this model, the test set correlation was obtained with three PLS factors and is characterized by: (a) a root mean-square error (*i.e.* RMSE) of 0.42; (b) a q^2 value of 0.49; and (c) a Pearson-R value of 0.71.

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

The scatter plot for the test set (see Figure 8) shows a reasonably good correlation between the predicted and experimental activities (r = 0.7). Nevertheless, there appears to be a general overprediction for the molecules with lowest activities. Consequently, the slope of the hypothetical *best-fit* line is less than one and it does not pass through the origin. As a result of this shift, the RMSE of the test set predictions is 0.42 log units (which is roughly twice the model standard deviation of regression).

The pEC50 activity values of the test set compounds range from 5.0 to 7.6 (see Table 1) and have been classified using the same activity scale as the training set. The analysis of this classification again shows that the 3D-QSAR model is more accurate at predicting the activities for ligands with moderate (*i.e.* ++) or high (*i.e.* +++) pEC50 than with low activities (*i.e.* +). In the test set, PHASE™ correctly predicted the scale in the activity interval in 34 out of 58 cases (59%) and the activity is predicted to be within 0.5 log units of the experimental value in all but 10 cases [where the larger differences are concentrated on the prediction of the inactive compounds with a maximum for molecule **220** (*i.e.* 1.11); see Table 1]. Since the activity values are not diverse enough to develop a model that gives a strong correlation without overfitting the data, additional compounds with lower activities (preferably pEC50 values of 5.0 or lower) should be included in the training test. Although these additional molecules are not available and, consequently, we cannot improve the 3D-QSAR model, we have analyzed the activities predicted by it and realized that no highly active agonists were predicted as inactive or *vice versa*. In this respect, the model could only incorrectly predict active agonists to be moderately active (and *vice versa*) or inactive molecules to be moderately active FXR agonists (and *vice versa*).

Analysis of the 3D-QSAR model for FXR

We visualized and analyzed the 3D-QSAR model so that we could understand how the different moieties from the agonist structures contribute, either positively or negatively, to the predicted activity. Thus, Figure 9 shows the three-dimensional aspect of the 3D-QSAR model (Figure 9A) by visualizing also: (1) how the hypothesis features fit in the context of the most active agonist (*i.e.* **245**; Figure 9B), the least active agonist (*i.e.* **222**; Figure 9C) and FEX (*i.e.* **259**; Figure 9D); and (2) the 3Q-QSAR model contributions. This gives additional insights into how each chemical feature can explain the agonist activity. Consequently, the 3D-QSAR model is represented with cubic volume elements occupied by the corresponding ligand where: (a) blue cubes indicate regions that are favorable for activity; and (b) red cubes indicate regions that are unfavorable for activity. Thus, Figure 9 illustrates the most significant favorable and unfavorable interactions that are suggested when the three-factor 3D-QSAR model is applied.

ISBN:978-84-691-1553-4 /DL: T.151-2008

The absence of blue or red regions in the potential hydrogen bond acceptor A_2 (see Figure 9A) suggests that this function does not affect agonist activity. We should take into account that all 112 compounds used (*i.e.* training and test set) are FXR agonists and that their structure always contains this amide carbonyl oxygen (the hypothesis had identified this feature in all agonists, see Table 1 and Figure 2). Consequently, although the 3D-QSAR model does not consider this function to be critical for agonist activity, it may be important for stabilizing the ligands in their binding site on the FXR-LBD. This is coherent with what is found when the experimental complex between FEX and FXR (*i.e.* 1OSH) is analyzed because it has been reported that this hydrogen bond acceptor (*i.e.* A_2) is important for stabilizing the position of the methyl ester group between helices 3 and 6 by means of two hydrogen bonds with His298 and Ser336 (this belongs to the first subset of interactions between FEX and FXR-LBD described above) [30]. Therefore, we consider that this function cannot establish a difference between agonists with high and low activities in our compounds.

When searching for specific chemical features that could enhance agonist activity, we analyzed the blue regions beside A_2 (*i.e.* located around the position of FEX's hexyl ring; see Figure 9D). These blue cubes are occupied either by the functional group $–CH(CH_3)_2$ or $–C_6H_{11}$ in all active compounds (except for **246** which has the $–NHCH(CH_3)_2$ substituent there; see Table 1 and Figure 2). Interestingly, when the activity of compound **246** is compared with that of the other two active compounds, which only differ from it in the location of the -NHCH(CH3)2 substituent. (*i.e.* **244** and **245**), the larger the substituent is, the lower is the activity. This is coherent with the fact that the region that is occupied by the end of these substituents is filled by red cubes in the 3D-QSAR model (see Figures 9A and 9C). Therefore, it seems that the substituents that are bound to the amide carbonyl group are critical for their activity as FXR agonists.

There is another blue region around the hydrophobic feature around the cyclic acetal and the aromatic ring group of the most active ligand (*i.e.* **H;** see Figure 9B). This indicates that this is a favourable region for enhancing the agonist activity. This is supported by the mapping of the least active compound, which only contributes to this hydrophobic region with one chlorine atom (see Figure 9C). Consequently, this hydrophobic region (which corresponds to the second subset of interactions that explains FEX's potency [30]) is also critical for FXR's agonist activity.

The third group of blue cubes is around the hydrogen bond acceptor A_1 (see Figure 9A) and indicates that interactions with the FXR-LBD in this area may increase agonist activity. Thus, when FEX is shown in the context of the 3D-QSAR model it can be seen that the hydrogen

acceptor function located in the methyl ester moiety favours FEX activity (see Figure 9D). This agrees with previous results that show that the methyl ester group provides a significant number of contacts with helix 3 [30]. At this point, it is worth pointing out that, surprisingly, the hydrogen bond acceptor labelled as A_3 seems to make a negative contribution to activity (see Figures 9C and 9D).

Therefore, we can conclude that our 3D-QSAR model explains the activity of the training and test set molecules (see Figure 2 and Table 1) because the most active agonists can occupy the regions with blue cubes while the least active ligands show very limited contributions to these areas (see the paper's website for screen captures that show how the remaining 109 molecules from both the training and the test sets fit in the 3D-QSAR model).

Searching for active FXR agonists in the 3D polyphenolic database

Since our 3D-QSAR model is able to describe the most important kind of interactions between the receptor (*i.e.* FXR) and a large set of non-steroidal agonists, we used it to estimate the activities of a series of phenolic compounds by assuming that if they are only active when the natural ligand (*i.e.* CDCA) is also present, they bind to FXR-LBD in the FEX's binding site (which is located beside CDCA's binding site; see Figure 6). Thus, according to this hypothesis, phenolic compounds interact with FXR by using a mechanism that is similar to the one from the synthetic non-steroidal agonists in the training and test sets [19]. Therefore, we have mapped two sets of phenolic compounds in the 3D-QSAR model: (a) the ones most frequently found in plant extracts [24]; and (b) their derived bioactive structures detected in plasma or urine [25-28].

The results of the search for FXR agonists in our database of phenolic compounds show that these molecules cannot simultaneously match in the five sites or chemical features of the 3D-QSAR model, which means that no phenolic compound is predicted to be highly active FXR agonist. However, when the *partial matching* is performed with the complete set of phenolic compounds (*i.e.* 135 molecules), we found a solution for 114 of them by which, in most cases, they are able to match four sites and a few can match three sites. The remaining 21 compounds that are not able to match in the 3D-QSAR model belong to the subclass of phenolic acids (see Figure 3) and, therefore, are not FXR agonists.

The 114 molecules that match three or four sites in the 3D-QSAR are predicted as either moderately active or inactive compounds (*i.e.* there are no compounds with a pEC50 higher than 7; see Figure 10). Thus, Figure 10 shows the predicted pEC50 values for these 114 compounds together with a RMSE of 0.42 (where the predicted activity values in Figure 10A correspond to the phenolic compounds most frequently found in plant extracts and in Figure

ISBN:978-84-691-1553-4 /DL: T.151-2008

10B to their derived bioactive structures). The analysis of the results in Figure 10A shows that the naturally-occurring phenolic compounds with the highest predicted pEC50 are some procyanidins (*i.e.* the B5 dimer and the C1, T3, T4, T5 and T6 trimers). Nevertheless, they are predicted to be only moderately active (*i.e.* the highest pEC50 value is 6.5 for trimer C1) surely because they are not able to map the hydrophobic feature of the hypothesis (*i.e.* **H**) or occupy the area around it (see at the paper's website for screen captures that show how the remaining of 114 phenolic compounds from our database fit in the 3D-QSAR model). In contrast, although most bioactive compounds were also predicted as moderately active or inactive, four of them (*i.e.* hesperetin 7-glucuronide, 4'-*O*-methyl-(-)-epicatechin 5-glucuronide, 4'-*O*-methyl-(-)-epicatechin 7-glucuronide and the tetramethylated dimeric; see Figure 10B) can also map the hydrophobic function (*i.e.* **H**) and, therefore, are predicted as moderately or highly active agonists (their pEC50 values are between 6.59 and 6.83). Thus, the glucuronidation, and the methylation can, in some cases, lead to substantially greater FXR-agonist activity of the resulting bioactive molecules than that of the original molecules from which they are derived. In fact, this hydrophobic function, which is additionally matched by these four molecules, has been described to be essential to stabilize helix 12 and, thus, enhance the binding affinity of the coactivator [29, 30, 36].

CONCLUSIONS

According to *in vitro* experiments from our group [21], GSPE [which essentially contains phenolic acids (4.22%) and procyanidins (monomers, dimers, trimers, tetramers and oligomers that represent 16.55, 18.77, 16.00, 9.3 and 35.7 % of the GSPE weight] cannot by itself enhance the activity of FXR [21]. However, GSPE co-incubation with CDCA, a natural ligand of FXR, enhanced the transcriptional activity of FXR/RXR [21]. This snergy is GSPE dose-dependent and the increase when cells are incubated with 100 mg/L of GSPE and CDCA is two-fold that of treatment with CDCA alone [21]. These results suggest that GSPE can enhance FXR activity only when the NR is already activated by CDCA.

In order to understand the activation of FXR by GSPE when CDCA is also present, we first analyzed the activation mechanism that has been proposed for CDCA [30]. Since this FXR-natural ligand does not form stable cocrystals with FXR that can be used for a further X-ray analysis, Downes et al. [30] modeled CDCA in the FXR binding pocket of the proteic part of 1OSH (the PDB file of the complex between FXR-LBD and FEX) and proposed that the natural ligand overlays its steroidal backbone on the biaryl group from FEX. Thus, it was predicted that the hydrophobic interactions with CDCA oriented helix 3 similarly to the helix in the complex

between FXR-LBD and FEX [30]. However, FEX's methyl ester group, which provides a significant number of contacts with helix 3, has no partner in its model of the complex between FXR-LBD and CDCA. On the other hand, and at the same time as this model was published, the complex between FXR-LBD and 6ECDCA (a semisynthetic bile acid with a structure closely-related to CDCA; see Figure 1) was obtained (*i.e.* 1OSV [29]). Thus, when the location of 6ECDCA and FEX in their complexes with FXR-LBD is compared, it becomes evident that both ligands are located on binding sites that are close but different (see Figure 6). The analysis of the interactions between 6ECDCA and FXR-LBD shows that the ligand helps helix 12 to be placed against helices 3, 4 and 10 and allows FXR to adopt its activated conformation (whereby helix 12 stabilizes the binding of the coactivator peptide). Consequently, although 6ECDCA cannot directely contact helix 12, the ligand stabilizes the cation-π interaction between His444 (from helix 10/11) and Trp466 (from helix 12) which helps helix 12 to locate in its active conformation [29]. Encouraged by the way in which 6ECDCA activates FXR, Zhang et al. (2007) suggested new ways for designing alternative FEX-based agonists that could make a cation-π interaction with Trp473 (the equivalent residue to Trp466 in human FXR) which was equivalent to the one between His444 and Trp466 in the complex between rat's FXR-LBD and 6ECDCA. As a result, it could also help helix 12 to locate in its active conformation [22]. Thus, on the basis of their 3D-QSAR analysis, they suggested that some agonists can be discovered when the region of the FEX's biaryl rings is properly modified and the modifications favour a cation-π interaction between Trp473 and the agonist.

Therefore, taking into account that FEX's potency appears to be mediated by two mechanisms and that ligands closely related to natural ones (like 6ECDCA) show some differences (but also some similarities) from FEX in their interactions with FXR-LBD, we suggest a mechanism which can explain the experimentally-observed synergic activation of FXR by CDCA and phenolic compounds. This mechanism suggests that two molecules (one from CDCA and the other from a phenolic compound) can simultaneously bind in the FXR ligand's binding site. Since the conformation of the receptor is essential to its activation, we propose a mechanism in which (a) CDCA places helix 12 in a location that favours the binding of the coactivator peptide by means of a hydrophobic or direct contact (therefore, the natural ligand changes the pose that it has in its experimental complex with FXR-LBD for another one that allows it to match in the hydrophobic region of our 3D-QSAR model), and (b) the phenolic compounds make interactions possible that are equivalent to the ones that allow FEX's methyl ester to bridge helix 3 with helix 6 by van der Waals contacts (interaction that is absent in the CDCA binding model). Therefore, the combined action of both molecules would have an effect similar to that of the FEX-derived agonists suggested by Zhang et al. (2007), which can help helix 12 to locate in its active conformation but also make the essential FEX interactions with the FXR-LBD [22].

ISBN:978-84-691-1553-4 /DL: T.151-2008

ACKNOWLEDGMENTS

We thank John Bates of our University's Language Service for correcting the manuscript, the Servei de Disseny de Fàrmacs from the Centre de Supercomputació de Catalunya (CESCA) for providing access to PHASE™ and SimBioSys Inc. (Toronto, Ontario, Canada) for providing us with eHiTS®. This study was supported by grant number CO3/O8 from the Fondo de Investigación Sanitaria (FIS) and AGL2005-04889 from the Comisión Interministerial de Ciencia y Tecnología (CICYT) of the Spanish Government. Montserrat Vaqué is the recipient of a fellowship from grant number CO3/O8.

REFERENCES

[1] C. Carlberg, T.W. Dunlop, An integrated biological approach to nuclear receptor signaling in physiological control and disease., Crit Rev Eukaryot Gene Expr 16 (2006) 1-22.

[2] M.M. Gottardis, E.D. Bischoff, M.A. Shirley, M.A. Wagoner, W.W. Lamph, R.A. Heyman, Chemoprevention of mammary carcinoma by LGD1069 (Targretin): an RXR-selective ligand., Cancer Res 56 (1996) 5566-5570.

[3] R. Mohan, R.A. Heyman, Orphan nuclear receptor modulators., Curr Top Med Chem 3 (2003) 1637-1647.

[4] R. Pellicciari, G. Costantino, S. Fiorucci, Farnesoid X receptor: From structure to potential clinical applications, J. Med. Chem. 48 (2005) 5383-5403.

[5] B. Cariou, B. Staels, FXR: a promising target for the metabolic syndrome?, Trends Pharmacol Sci 28 (2007) 236-243.

[6] S. Fiorucci, G. Rizzo, A. Donini, E. Distrutti, L. Santucci, Targeting farnesoid X receptor for liver and metabolic disorders., Trends Mol Med 13 (2007) 298-309.

[7] P. Germain, B. Staels, C. Dacquet, M. Spedding, V. Laudet, Overview of nomenclature of nuclear receptors., Pharmacol Rev 58 (2006) 685-704.

[8] K.W. Nettles, G.L. Greene, Nuclear receptor ligands and cofactor recruitment: Is there a coactivator "On deck"?, Molecular Cell 11 (2003) 850-851.

[9] B.M. Forman, E. Goode, J. Chen, A.E. Oro, D.J. Bradley, T. Perlmann, D.J. Noonan, L.T. Burka, T. McMorris, W.W. Lamph, R.M. Evans, C. Weinberger, Identification of a nuclear receptor that is activated by farnesol metabolites., Cell 81 (1995) 687-693.

[10] S.A. Kliewer, J.M. Lehmann, T.M. Willson, Orphan nuclear receptors: shifting endocrinology into reverse., Science 284 (1999) 757-760.

[11] D.J. Parks, S.G. Blanchard, R.K. Bledsoe, G. Chandra, T.G. Consler, S.A. Kliewer, J.B. Stimmel, T.M. Willson, A.M. Zavacki, D.D. Moore, J.M. Lehmann, Bile acids: natural ligands for an orphan nuclear receptor., Science 284 (1999) 1365-1368.

[12] H. Wang, J. Chen, K. Hollister, L.C. Sowers, B.M. Forman, Endogenous bile acids are ligands for the nuclear receptor FXR/BAR., Mol Cell 3 (1999) 543-553.

[13] T. Claudel, B. Staels, F. Kuipers, The Farnesoid X receptor: a molecular link between bile acid and lipid and glucose metabolism., Arterioscler Thromb Vasc Biol 25 (2005) 2020-2030.

[14] D. Duran-Sandoval, B. Cariou, J.C. Fruchart, B. Staels, Potential regulatory role of the farnesoid X receptor in the metabolic syndrome., Biochimie 87 (2005) 93-98.

[15] F. Kuipers, T. Claudel, E. Sturm, B. Staels, The Farnesoid X Receptor (FXR) as modulator of bile acid metabolism., Rev Endocr Metab Disord 5 (2004) 319-326.

[16] M. Makishima, A.Y. Okamoto, J.J. Repa, H. Tu, R.M. Learned, A. Luk, M.V. Hull, K.D. Lustig, D.J. Mangelsdorf, B. Shan, Identification of a nuclear receptor for bile acids., Science 284 (1999) 1362-1365.

[17] R. Pellicciari, S. Fiorucci, E. Camaioni, C. Clerici, G. Costantino, P.R. Maloney, A. Morelli, D.J. Parks, T.M. Willson, 6alpha-ethyl-chenodeoxycholic acid (6-ECDCA), a potent and selective FXR agonist endowed with anticholestatic activity., J Med Chem 45 (2002) 3569-3572.

[18] P.R. Maloney, D.J. Parks, C.D. Haffner, A.M. Fivush, G. Chandra, K.D. Plunket, K.L. Creech, L.B. Moore, J.G. Wilson, M.C. Lewis, S.A. Jones, T.M. Willson, Identification of a chemical tool for the orphan nuclear receptor FXR., J Med Chem 43 (2000) 2971-2974.

[19] K.C. Nicolaou, R.M. Evans, A.J. Roecker, R. Hughes, M. Downes, J.A. Pfefferkorn, Discovery and optimization of non-steroidal FXR agonists from natural product-like libraries, Org Biomol Chem 1 (2003) 908-920.

ISBN:978-84-691-1553-4 /DL: T.151-2008

[20] J.M. Del Bas, J. Fernández-Larrea, M. Blay, A. Ardèvol, M.J. Salvadó, L. Arola, C. Bladé, Grape seed procyanidins improve atherosclerotic risk index and induce liver CYP7A1 and SHP expression in healthy rats., FASEB J 19 (2005) 479-481.

[21] J.M. Del Bas, Modulation of hepatic lipoprotein metabolism by dietary procyanidins, Departament de Bioquímica i Biotecnologia, Universitat Rovira i Virgili, Tarragona, 2007, p. 218.

[22] T. Zhang, J.H. Zhou, L.W. Shi, R.X. Zhu, M.B. Chen, 3D-QSAR studies with the aid of molecular docking for a series of non-steroidal FXR agonists., Bioorg Med Chem Lett 17 (2007) 2156-2160.

[23] K.M. Honório, R.C. Garratt, I. Polikarpov, A.D. Andricopulo, 3D QSAR comparative molecular field analysis on nonsteroidal farnesoid X receptor activators., J Mol Graph Model 25 (2007) 921-927.

[24] J. Macheix, A. Fleuriet, J. Billot, Fruit Phenolics, Boca Raton, FL: CRC Press 1990.

[25] C. Felgines, S. Talavéra, M. Gonthier, O. Texier, A. Scalbert, J. Lamaison, C. Rémésy, Strawberry anthocyanins are recovered in urine as glucuro- and sulfoconjugates in humans., J Nutr 133 (2003) 1296-1301.

[26] C. Felgines, S. Talavera, O. Texier, A. Gil-Izquierdo, J. Lamaison, C. Remesy, Blackberry anthocyanins are mainly recovered from urine as methylated and glucuronidated conjugates in humans., J Agric Food Chem 53 (2005) 7721-7727.

[27] C. Manach, G. Williamson, C. Morand, A. Scalbert, C. Rémésy, Bioavailability and bioefficacy of polyphenols in humans. I. Review of 97 bioavailability studies., Am J Clin Nutr 81 (2005) 230S-242S.

[28] C. Tsang, C. Auger, W. Mullen, A. Bornet, J. Rouanet, A. Crozier, P. Teissedre, The absorption, metabolism and excretion of flavan-3-ols and procyanidins following the ingestion of a grape seed extract by rats., Br J Nutr 94 (2005) 170-181.

[29] L.Z. Mi, S. Devarakonda, J.M. Harp, Q. Han, R. Pellicciari, T.M. Willson, S. Khorasanizadeh, F. Rastinejad, Structural basis for bile acid binding and activation of the nuclear receptor FXR., Mol Cell 11 (2003) 1093-1100.

[30] M. Downes, M.A. Verdecia, A.J. Roecker, R. Hughes, J.B. Hogenesch, H.R. Kast-Woelbern, M.E. Bowman, J.L. Ferrer, A.M. Anisfeld, P.A. Edwards, J.M. Rosenfeld, J.G. Alvarez, J.P. Noel, K.C. Nicolaou, R.M. Evans, A chemical, genetic, and structural analysis of the nuclear bile acid receptor FXR., Mol Cell 11 (2003) 1079-1092.

[31] Z. Zsoldos, D. Reid, A. Simon, B.S. Sadjad, A.P. Johnson, eHiTS: an innovative approach to the docking and scoring function problems., Curr Protein Pept Sci 7 (2006) 421-435.

[32] Z. Zsoldos, D. Reid, A. Simon, S.B. Sadjad, A.P. Johnson, eHiTS: a new fast, exhaustive flexible ligand docking system., J Mol Graph Model 26 (2007) 198-212.

[33] N. Guex, M.C. Peitsch, SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling., Electrophoresis 18 (1997) 2714-2723.

[34] S.L. Dixon, A.M. Smondyrev, E.H. Knoll, S.N. Rao, D.E. Shaw, R.A. Friesner, PHASE: a new engine for pharmacophore perception, 3D QSAR model development, and 3D database screening: 1. Methodology and preliminary results., J Comput Aided Mol Des 20 (2006) 647-671.

[35] W. Sherman, T. Day, M.P. Jacobson, R.A. Friesner, R. Farid, Novel procedure for modeling ligand/receptor induced fit effects., J Med Chem 49 (2006) 534-553.

[36] K.W. Nettles, G.L. Greene, Nuclear receptor ligands and cofactor recruitment: is there a coactivator "on deck"?, Mol Cell 11 (2003) 850-851.

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

TABLES AND FIGURES

Table 1. Parameters for the analysis of the hypothesis generated and 3D-QSAR model

Training set											Test set									
	Hypothesis							3D-QSAR model				Hypothesis							3D-QSAR model	
Comp.	A_1	A_2	A_3	H	R	Exp. pEC50	activity scale	Est. pEC50	activity scale	residual	Comp.	A_1	A_2	A_3	H	R	Exp. pEC50	activity scale	Est. pEC50	activity scale	residual
245	+	+	+	-	+	7.72	+++	7.37	+++	0.35	**259***	+	+	+	+	+	7.60	+++	6.98	++	0.62
244	+	+	+	-	+	7.42	+++	7.19	+++	0.23	**235**	+	+	+	+	+	7.59	+++	6.93	++	0.66
192	+	+	+	+	+	7.19	+++	7.10	+++	0.09	**216**	+	+	+	+	+	7.29	+++	6.88	++	0.41
247	+	+	+	+	+	7.18	+++	7.11	+++	0.07	**260**	+	+	+	+	+	7.24	+++	6.89	++	0.35
193	+	+	+	+	+	7.15	+++	7.25	+++	-0.10	**198**	+	+	+	-	+	7.19	+++	6.84	++	0.35
202	+	+	+	+	+	7.15	+++	7.24	+++	-0.09	**125**	+	+	+	+	+	7.16	+++	6.88	++	0.28
213	+	+	+	+	+	7.14	+++	6.85	++	0.29	**224**	+	+	+	+	+	7.14	+++	7.01	+++	0.13
233	+	+	+	+	+	7.11	+++	7.06	+++	0.05	**126**	+	+	+	+	+	7.11	+++	7.09	+++	0.02
201	+	+	+	+	+	7.07	+++	6.90	++	0.17	**256**	+	+	+	-	+	7.11	+++	6.67	++	0.44
257	+	+	+	-	+	7.02	+++	7.13	+++	-0.11	**229**	+	+	+	+	+	7.05	+++	6.59	++	0.46
246	+	+	+	-	+	7.02	+++	6.71	++	0.31	**232**	+	+	+	+	+	7.03	+++	7.08	+++	-0.05
223	+	+	+	+	+	7.00	++	7.10	+++	-0.10	**230**	+	+	+	+	+	7.01	+++	6.63	++	0.38
236	+	+	+	+	+	6.93	++	7.07	+++	-0.14	**190**	+	+	+	+	+	6.97	++	7.08	+++	-0.11
205	+	+	+	-	+	6.92	++	6.85	++	0.07	**238**	+	+	+	+	+	6.96	++	7.06	+++	-0.10
196	+	+	+	+	+	6.83	++	6.94	++	-0.11	**248**	+	+	+	+	+	6.89	++	6.95	++	-0.06
177	+	+	+	+	+	6.82	++	6.62	++	0.20	**226**	+	+	+	+	+	6.83	++	6.77	++	0.06
239	+	+	+	+	+	6.79	++	7.03	+++	-0.24	**261**	+	+	+	+	+	6.79	++	6.72	++	0.07
227	+	+	+	+	+	6.76	++	6.72	++	0.04	**181**	+	+	+	+	+	6.79	++	7.18	+++	-0.39
178	+	+	+	+	+	6.71	++	6.87	++	-0.16	**204**	+	+	+	-	+	6.73	++	6.31	++	0.42
144	+	+	+	+	+	6.69	++	6.92	++	-0.23	**157**	+	+	+	+	+	6.70	++	6.34	++	0.36
145	+	+	+	+	+	6.67	++	6.41	++	0.26	**128**	+	+	+	+	+	6.69	++	6.93	++	-0.24
137	+	+	+	+	+	6.66	++	6.69	++	-0.03	**122**	+	+	+	-	+	6.68	++	6.43	++	0.25
242	+	+	+	+	+	6.65	++	6.82	++	-0.17	**251**	+	+	+	+	+	6.66	++	6.87	++	-0.21
127	+	+	+	-	+	6.64	++	6.45	++	0.19	**210**	+	+	+	+	+	6.64	++	6.88	++	-0.24
156	+	+	+	-	+	6.63	++	6.50	++	0.13	**211**	+	+	+	+	+	6.64	++	7.08	+++	-0.44
143	+	+	+	+	+	6.63	++	6.34	++	0.29	**241**	+	+	+	+	+	6.63	++	6.47	++	0.16
129	+	+	+	-	+	6.59	++	6.60	++	-0.01	**136**	-	+	+	+	+	6.61	++	5.92	+	0.69
250	+	+	+	+	+	6.58	++	6.69	++	-0.11	**254**	+	+	+	+	+	6.61	++	6.82	++	-0.21
207	+	+	+	-	+	6.51	++	6.36	++	0.15	**187**	+	+	+	+	+	6.57	++	6.39	++	0.18
206	+	+	+	-	+	6.46	++	6.60	++	-0.14	**208**	+	+	+	-	+	6.51	++	6.68	++	-0.17
253	+	+	+	+	+	6.38	++	6.61	++	-0.23	**102**	+	+	+	+	+	6.48	++	5.84	+	0.64

ISBN:978-84-691-1553-4 /DL: T.151-2008

Training set											Test set									
Hypothesis								3D-QSAR model			Hypothesis								3D-QSAR model	
Comp.	A_1	A_2	A_3	H	R	Exp. pEC50	activity scale	Est. pEC50	activity scale	residual	Comp.	A_1	A_2	A_3	H	R	Exp. pEC50	activity scale	Est. pEC50	activity scale	residual
237	+	+	+	+	+	6.35	++	6.37	++	-0.02	**65**	+	+	+	+	+	6.45	++	6.52	++	-0.07
118	+	+	+	-	+	6.22	++	6.19	++	0.03	**194**	+	+	+	+	+	6.37	++	6.56	++	-0.19
188	+	+	+	+	+	6.03	++	5.92	++	0.11	**197**	+	+	+	+	+	6.37	++	6.45	++	-0.08
110	+	+	+	-	+	6.00	++	6.33	++	-0.33	**203**	+	+	+	+	+	6.33	++	5.91	+	0.42
113	+	+	+	-	+	6.00	++	6.21	++	-0.21	**258**	+	+	+	-	+	6.25	++	6.17	++	0.08
132	+	+	+	-	+	6.00	++	6.17	++	-0.17	**111**	+	+	+	-	+	6.17	++	5.77	+	0.40
134	+	+	+	+	+	6.00	++	6.01	++	-0.01	**112**	+	+	+	-	+	6.00	++	6.44	++	-0.44
146	+	+	+	+	+	6.00	++	6.09	++	-0.09	**114**	+	+	+	-	+	6.00	++	6.3	++	-0.30
148	+	+	+	-	+	6.00	++	5.74	+	0.26	**115**	+	+	+	-	+	6.00	++	5.89	+	0.11
153	+	+	+	-	+	6.00	++	5.99	+	0.01	**135**	+	+	+	+	+	6.00	++	6.13	++	-0.13
154	+	+	+	-	+	6.00	++	6.21	++	-0.21	**147**	+	+	+	-	+	6.00	++	6.07	++	-0.07
161	+	+	+	+	+	6.00	++	6.08	++	-0.08	**150**	+	+	+	-	+	6.00	++	6.27	++	-0.27
163	+	+	+	+	+	6.00	++	5.87	+	0.13	**151**	+	+	+	+	+	6.00	++	6.64	++	-0.64
165	+	+	+	+	+	6.00	++	5.99	+	0.01	**155**	+	+	+	+	+	6.00	++	6.34	++	-0.34
166	+	+	+	+	+	6.00	++	5.92	+	0.08	**160**	+	+	+	+	+	6.00	++	6.2	++	-0.20
240	+	+	+	+	+	5.88	+	5.86	+	0.02	**162**	+	+	+	+	+	6.00	++	6.24	++	-0.24
175	+	+	+	-	+	5.85	+	5.95	+	-0.10	**164**	+	+	+	+	+	6.00	++	6.13	++	-0.13
185	+	+	+	+	+	5.74	+	5.88	+	-0.14	**167**	+	+	+	+	+	6.00	++	6.03	++	-0.03
184	+	+	+	+	+	5.71	+	5.79	+	-0.08	**225**	+	+	+	+	+	5.86	+	6.07	++	-0.21
249	+	+	+	+	+	5.52	+	5.83	+	-0.31	**234**	+	+	+	+	+	5.85	+	5.83	+	0.02
176	+	+	+	-	+	5.45	+	5.38	+	0.07	**183**	+	+	+	+	+	5.83	+	6.07	++	-0.24
191	+	+	+	+	+	5.40	+	5.05	+	0.35	**228**	+	+	+	+	+	5.63	+	6.62	++	-0.99
222	+	+	+	+	+	5.00	+	5.24	+	-0.24	**138**	+	+	-	+	+	5.55	+	6.23	++	-0.68
											243	+	+	+	+	+	5.51	+	5.96	+	-0.45
											220	+	+	+	-	+	5.40	+	6.51	++	-1.11
											252	+	+	+	+	+	5.12	+	6.17	++	-1.05
											255	+	+	+	+	+	5.00	+	5.93	+	-0.93

In the hypothesis columns **A1**, **A2**, **A3**, **H** and **R** indicate which hypothesis features are mapped by each molecule (*i.e.* a **+** sign indicates that the corresponding pharmacophore function is mapped by the molecule whereas the **–** sign indicates the opposite; **A**, **H** and **R** stand for hydrogen bond acceptor, hydrophobic group and aromatic ring, respectively; **A1**, **A2** and **A3** identify the three different hydrogen bond acceptors of the hypothesis). See Figure 4 for the location of the hypothesis features relative to the experimental pose of FEX when it binds to FXR-LBD.

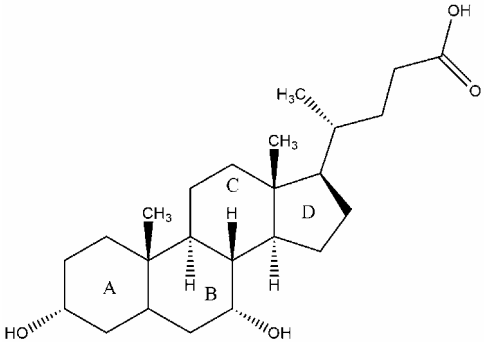
The 3D-QSAR model is characterized by its statistical values (SD = 0.18, R^2 = 0.91, F = 159.1 and P = 1.45e-25) and by predicting the activity of the test set (RMSE = 0.42, q^2 = 0.49 and Pearson-R = 0.71).

The experimental and the estimated pEC50 values for the training and test set molecules are indicated in the corresponding **Exp. pEC50** and **Est. pEC50** columns. The experimental pEC50 values were obtained from the bibliography [19]. The estimated pEC50 values, on the other hand, were calculated by using the 3D-QSAR model obtained. The EC50 values span four orders of magnitude (with pEC50 values ranging from 5.00 to 7.72 nM) and they have been classified in the **activity scale** column under the following criteria: (a) "+++" for highly active molecules (*i.e.* pEC50 > 7.0 nM); (b) "++" for moderately active molecules (*i.e.* 6.0 nM ≤ pEC50 ≤7.0 nM); (c) and "+" for inactive compounds (*i.e.* pEC50 < 6.0 nM). The **residual** column is computed as the difference between the experimental pEC50 and the estimated pEC50. An error with a negative sign indicates that the estimated pEC50 is higher than the corresponding experimental value.

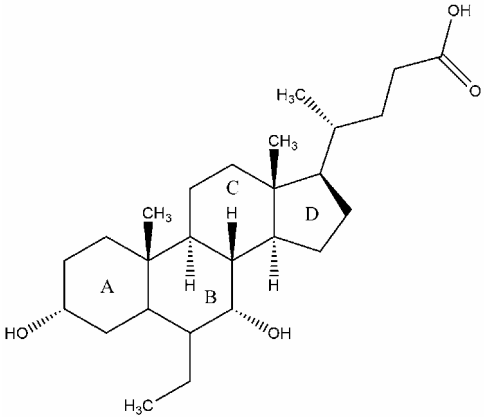
* The compound labelled 259 corresponds to FEX.

ISBN:978-84-691-1553-4 /DL: T.151-2008

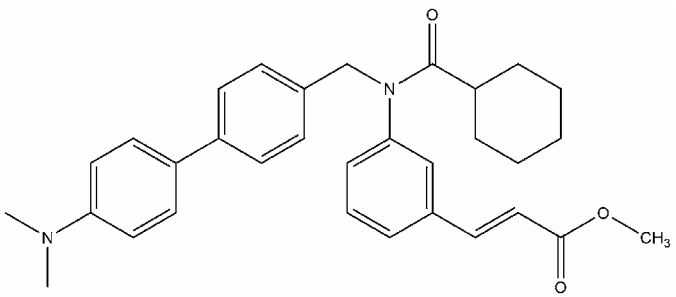
Figure 1. Endogenous and synthetic FXR ligands



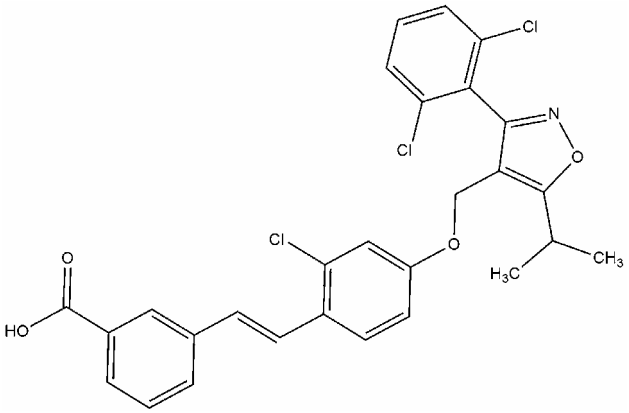
chenodeoxycholic acid or CDCA (low affinity endogenous agonist)



6α-ethyl-chenodeoxycholic acid or 6ECDCA (semisynthetic bile acid)

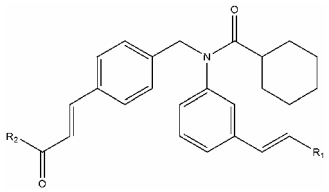


fexaramine or FEX

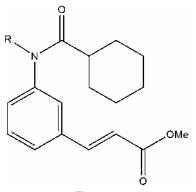


GW4064 (high affinity agonist)

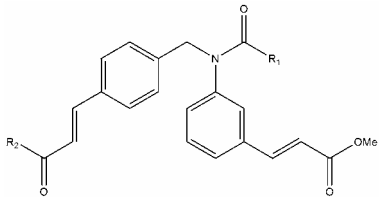
Figure 2. Chemical structures of the molecules used either to generate (*i.e.* training set) or to validate (*i.e.* test set) the pharmacophore candidates obtained with PHASE™



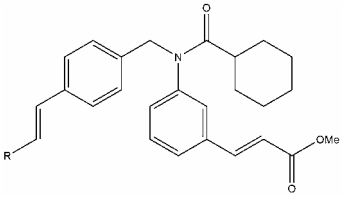
	R1	R2	EC_{50}nM	pEC_{50}
134 (training)	COO^{t}Bu	^{t}BuO	>1000	6.000
135 (test)	$CONH_2$	^{t}BuO	>1000	6.000
136 (test)	CH_2OM_e	^{t}BuO	243	6.614
137 (training)	CH_2OEt	^{t}BuO	220	6.658
138 (test)	CH_2OPh	^{t}BuO	2830	5.548



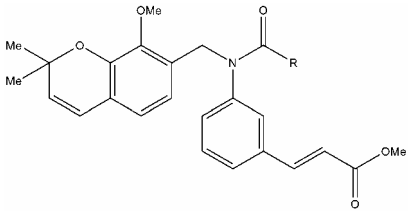
	R	EC_{50}nM	pEC_{50}
132 (training)	H	>1000	6.000
147 (test)	methyl	>1000	6.000
110 (training)	benzyl	>1000	6.000
111 (test)	2-napthyl	680	6.167
114 (test)	2-bromobenzyl	>1000	6.000
113 (training)	3-bromobenzyl	>1000	6.000
112 (test)	4-bromobenzyl	>1000	6.000
148 (training)	4-*tert*-butylbenzyl	>1000	6.000
115 (test)	3-methoxybenzyl	>1000	6.000
118 (training)	3,5-dimethoxybenzyl	606	6.218



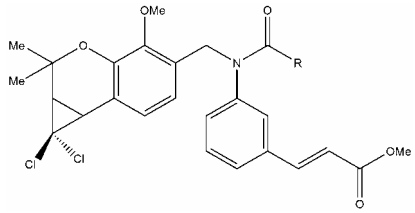
	R1	R2	EC_{50}nM	pEC_{50}
143 (training)	phenyl	^{t}BuO	236	6.627
144 (training)	2-furyl	^{t}BuO	205	6.688
145 (training)	isopropylamino	^{t}BuO	212	6.674
146 (training)	benzylamino	^{t}BuO	>1000	6.000



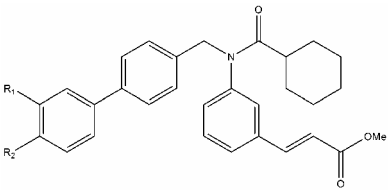
	R	EC_{50}nM	pEC_{50}
150 (test)	COOMe	>1000	6.000
151(test)	COOEt	>1000	6.000
153 (training)	COOBn	>1000	6.000
154 (training)	$CONMe_2$	>1000	6.000
155 (test)	CONH^{t}Bu	>1000	6.000
156 (training)	CH_2OMe	233	6.633
157 (test)	CH_2OEt	198	6.703



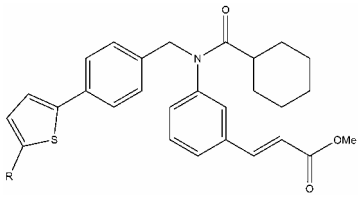
	R	EC_{50}nM	pEC_{50}
65 (test)	cyclohexyl	358	6.446



	R	EC_{50}nM	pEC_{50}
102 (test)	phenyl	333	6.478

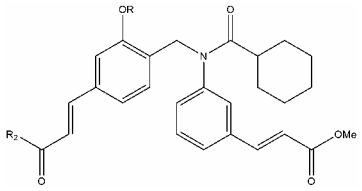


	R1	R2	EC_{50}nM	pEC_{50}
126 (test)	OMe	H	77	7.114
127 (training)	C(O)Me	H	227	6.644
125 (test)	H	SMe	69	7.161

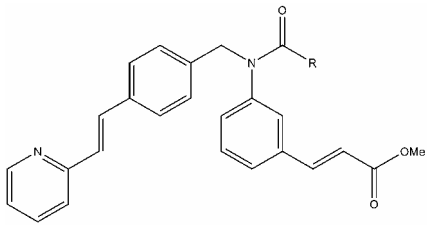


	R	EC_{50}nM	pEC_{50}
128 (test)	Me	206	6.686
129 (training)	C(O)Me	256	6.592

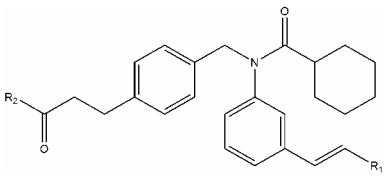
ISBN:978-84-691-1553-4 /DL: T.151-2008



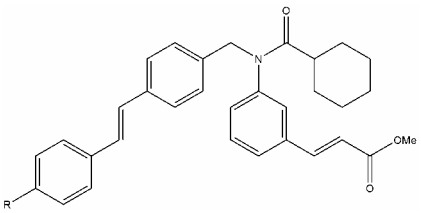
	R1	R2	EC_{50}nM	pEC_{50}
161 (training)	H	^{t}BuO	>1000	6.000
162 (test)	Me	^{t}BuO	>1000	6.000
163 (training)	Bn	^{t}BuO	>1000	6.000
164 (test)	MeC(O)	^{t}BuO	>1000	6.000
165 (training)	$C_6H_5C(O)$	^{t}BuO	>1000	6.000
166 (training)	$MeS(O_2)$	^{t}BuO	>1000	6.000
167 (test)	$EtOOCCH_2$	^{t}BuO	>1000	6.000



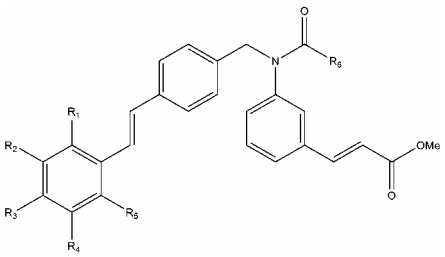
	R	EC_{50}nM	pEC_{50}
207 (training)	$-C_6H_{11}$	309	6.510
208 (test)	$-CH(CH_3)_2$	310	6.509



	R1	R2	EC_{50}nM	pEC_{50}
160 (test)	COO^tBu	^{t}BuO	>1000	6.000

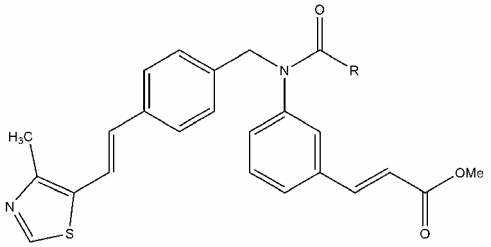


	R	EC_{50}nM	pEC_{50}
122 (test)	OMe	208	6.682

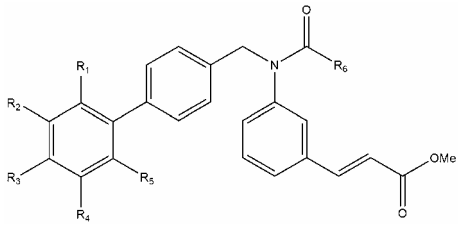


	R1	R2	R3	R4	R5	R6	EC_{50}nM	pEC_{50}
175 (training)	H	H	Me	H	H	$-CH(CH_3)_2$	1410	5.851
176 (training)	H	H	Me	H	H	$-NHCH(CH_3)_2$	3570	5.447
177 (training)	Cl	H	H	H	Cl	$-C_6H11$	150	6.824
178 (training)	Cl	H	H	H	Cl	$-CH(CH_3)_2$	195	6.710
181 (test)	H	Cl	H	H	H	$-CH(CH_3)_2$	164	6.785
183 (test)	H	CF_3	H	CF_3	H	$-C_6H_{11}$	1470	5.833
184 (training)	H	CF_3	H	CF_3	H	$-CH(CH_3)_2$	1950	5.710
185 (training)	H	CF_3	H	CF_3	H	$-NHCH(CH_3)_2$	1830	5.738
187 (test)	H	CF_3	H	H	H	$-CH(CH_3)_2$	267	6.573
188 (training)	H	CF_3	H	H	H	$-NHCH(CH_3)_2$	932	6.031
190 (test)	F	H	H	H	F	$-CH(CH_3)_2$	108	6.967
191 (training)	F	H	H	H	F	$-NHCH(CH_3)_2$	4020	5.396
192 (training)	F	H	H	H	H	$-C_6H^{11}$	64	7.194
193 (training)	F	H	H	H	H	$-CH(CH_3)_2$	70	7.155
194 (test)	F	H	H	H	H	$-NHCH(CH_3)_2$	431	6.366
196 (training)	Me	H	Me	H	Me	$-CH(CH_3)_2$	149	6.827
197 (test)	Me	H	Me	H	Me	$-NHCH(CH_3)_2$	431	6.366
198 (test)	H	H	H	H	H	$-CH(CH_3)_2$	65	7.187
201(training)	H	F	H	H	H	$-C_6H_{11}$	86	7.066
202 (training)	H	F	H	H	H	$-CH(CH_3)_2$	71	7.149
203 (test)	H	F	H	H	H	$-NHCH(CH_3)_2$	467	6.331
204 (test)	H	H	F	H	H	$-C_6H_{11}$	185	6.733
205 (training)	H	H	F	H	H	$-CH(CH_3)_2$	120	6.921
206 (training)	H	H	F	H	H	$-NHCH(CH_3)_2$	348	6.458

ISBN:978-84-691-1553-4 /DL: T.151-2008



	R	EC_{50}nM	pEC_{50}
210 (test)	$-C_6H_{11}$	227	6.644
211 (test)	$-CH(CH_3)_2$	228	6.642



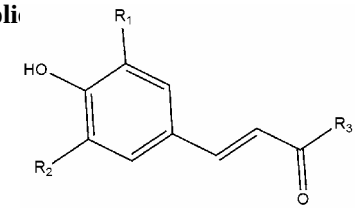
	R1	R2	R3	R4	R5	R6	EC_{50}nM	pEC_{50}
213 (training)	H	F	F	H	H	$-C_6H_{11}$	72	7.143
216 (test)	H	H	SMe	H	H	$-CH(CH_3)_2$	51	7.292
220 (test)	OMe	H	H	H	H	$-NHCH(CH_3)_2$	4010	5.397
222 (training)	H	Cl	H	Cl	H	$-NHCH(CH_3)_2$	>10000	5.000
223 (training)	H	OMe	H	H	H	$-C_6H_{11}$	101	6.996
224 (test)	H	OMe	H	H	H	$-CH(CH_3)_2$	72	7.143
225 (test)	H	OMe	H	H	H	$-NHCH(CH_3)_2$	1370	5.863
226 (test)	H	OEt	H	H	H	$-C_6H_{11}$	147	6.833
227 (training)	H	OEt	H	H	H	$-CH(CH_3)_2$	173	6.762
228 (test)	H	OEt	H	H	H	$-NHCH(CH_3)_2$	2350	5.629
229 (test)	H	H	OMe	H	H	$-C_6H_{11}$	89	7.051
230 (test)	H	H	OMe	H	H	$-CH(CH_3)_2$	97	7.013
232 (test)	H	Cl	H	H	H	$-C_6H_{11}$	94	7.027
233 (training)	H	Cl	H	H	H	$-CH(CH_3)_2$	77	7.114
234 (test)	H	Cl	H	H	H	$-NHCH(CH_3)_2$	1400	5.854
235 (test)	H	H	Me	H	H	$-C_6H_{11}$	26	7.585
236 (training)	H	H	Me	H	H	$-CH(CH_3)_2$	118	6.928
237 (training)	H	H	Me	H	H	$-NHCH(CH_3)_2$	449	6.348
238 (test)	H	Me	H	H	H	$-C_6H_{11}$	109	6.963
239 (training)	H	Me	H	H	H	$-CH(CH_3)_2$	163	6.788
240 (training)	H	Me	H	H	H	$-NHCH(CH_3)_2$	1330	5.876
241 (test)	OMe	H	H	Cl	H	$-C_6H_{11}$	233	6.633
242 (training)	OMe	H	H	Cl	H	$-CH(CH_3)_2$	226	6.646
243 (test)	OMe	H	H	Cl	H	$-NHCH(CH_3)_2$	3080	5.511
244 (training)	H	$-OCH_2O-$		H	H	$-C_6H_{11}$	38	7.420
245 (training)	H	$-OCH_2O-$		H	H	$-CH(CH_3)_2$	19	7.721
246 (training)	H	$-OCH_2O-$		H	H	$-NHCH(CH_3)_2$	96	7.018
247 (training)	H	Cl	F	H	H	$-C_6H_{11}$	66	7.180
248 (test)	H	Cl	F	H	H	$-CH(CH_3)_2$	129	6.889
249 (training)	H	Cl	F	H	H	$-NHCH(CH_3)_2$	3050	5.516
250 (training)	H	H	OCF_3	H	H	$-C_6H_{11}$	264	6.578
251 (test)	H	H	OCF_3	H	H	$-CH(CH_3)_2$	219	6.660
252 (test)	H	H	OCF_3	H	H	$-NHCH(CH_3)_2$	7530	5.123
253 (training)	H	OCF_3	H	H	H	$-C_6H_{11}$	420	6.377
254 (test)	H	OCF_3	H	H	H	$-CH(CH_3)_2$	247	6.607
255 (test)	H	OCF_3	H	H	H	$-NHCH(CH_3)_2$	>10000	5.000
256 (test)	OMe	H	H	H	OMe	$-C_6H_{11}$	77	7.114
257 (training)	OMe	H	H	H	OMe	$-CH(CH_3)_2$	95	7.022
258 (test)	OMe	H	H	H	OMe	$-NHCH(CH_3)_2$	561	6.251
259 (test)	H	H	NMe_2	H	H	$-C_6H_{11}$	25	7.602
260 (test)	H	H	NMe_2	H	H	$-CH(CH_3)_2$	57	7.244
261 (test)	H	H	NMe_2	H	H	$-NHCH(CH_3)_2$	162	6.790

ISBN:978-84-691-1553-4 /DL: T.151-2008

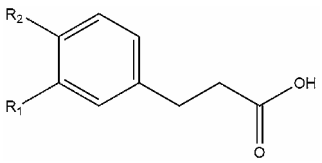
Figure 3. Molecular structures of the phenolic compounds tested

NON FLAVONOIDS

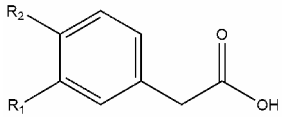
Phenoli



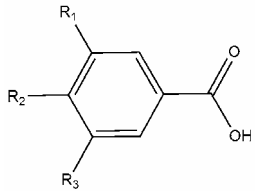
	R1	R2	R3
4-coumaric acid	H	H	OH
caffeic acid	OH	H	OH
ferulic acid	OCH3	H	OH
sinapic acid	OCH3	OCH3	OH
chlorogenic acid	H	OH	Glcde (1)



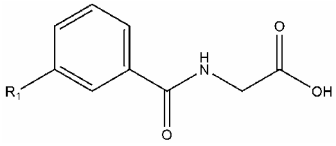
	R1	R2
propionic acid	H	H
m-hydroxyphenylpropionic acid	OH	H
p-hydroxyphenylpropionic acid	H	OH



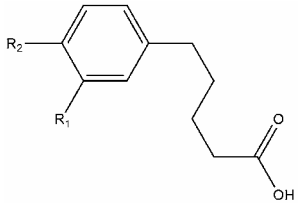
	R1	R2
3,4-dihydroxyphenilacetic acid	OH	OH
m-hydroxyphenylacetic acid	OH	H
p-hydroxyphenylacetic acid	H	OH
phenylacetic acid	H	H
homovanillic acid	OCH3	OH



	R1	R2	R3
gallic acid	OH	OH	OH
4-*O*-methylgallic acid	OH	OCH3	OH
gallic 3-glucuronide	OH	OH	Glcde (1)
gallic 4-glucuronide	OH	Glcde (1)	OH
m-hydroxybenzoic acid	OH	H	H
p-hydroxybenzoic acid	H	OH	H
protocatechuic acid	H	OH	OH
syringic acid	OCH3	OH	OCH3
vanillic acid	H	OH	OCH3

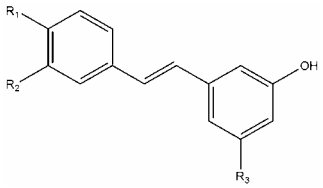


	R1
hippuric acid	H
m-hydroxyhippuric acid	OH

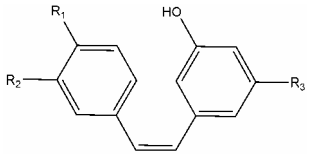


	R1	R2
m-hydroxyphenylvaleric acid	OH	H
p-hydroxyphenylvaleric acid	H	OH

Stilbenes



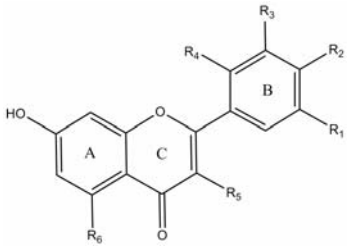
	R1	R2	R3
astringin	OH	OH	Glc (2)
piceatannol	OH	OH	OH
piceid	OH	H	Glc (2)
trans-resveratrol	OH	H	OH
resveratroloside	Glc (2)	H	OH
trans-resveratrol 3-glucuronide	OH	H	Glcde (1)
trans-resveratrol 4'glucuronide	Glcde (1)	H	OH
trans-resveratrol 3-sulfate	OH	H	SO4



	R1	R2	R3
cis-resveratrol 3-glucuronide	OH	H	Glcde (1)

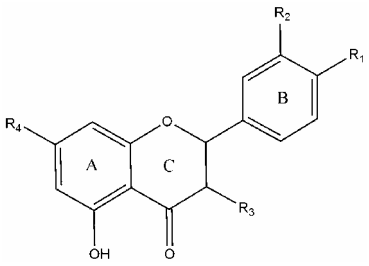
FLAVONOIDS

Flavonols



	R1	R2	R3	R4	R5	R6
isorhamnetin	H	OH	OCH3	H	OH	OH
kaempferol	H	OH	H	H	OH	OH
myricetin	OH	OH	OH	H	OH	OH
quercetin	H	OH	OH	H	OH	OH
3'-*O*-methylquercetin 3-glucuronide	H	OH	OCH3	H	Glcde [(1)]	OH
quercetin 3-glucuronide	H	OH	OH	H	Glcde [(1)]	OH
quercetin 3-sulfate	H	OH	SO4	H	OH	OH
kaempferol 3-glucuronide	H	OH	OCH3	H	Glcde [(1)]	OH

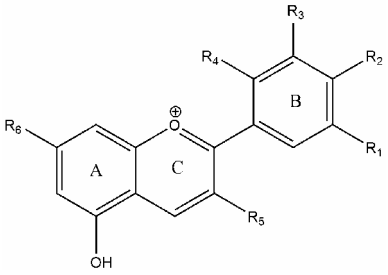
Flavanones



	R1	R2	R3	R4
naringenin	OH	H	H	OH
hesperetin 3'-glucuronide	OCH3	Glcde [(1)]	H	OH
hesperetin 7-glucuronide	OCH3	OH	H	Glcde [(1)]

ISBN:978-84-691-1553-4 /DL: T.151-2008

Anthocyanidins

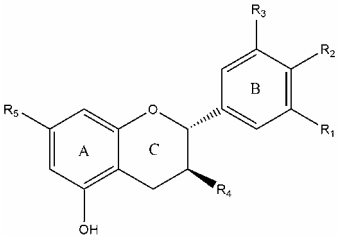


	R1	R2	R3	R4	R5	R6
cyanidin	H	OH	OH	H	OH	OH
delphinidin	OH	OH	OH	H	OH	OH
malvidin	OCH3	OH	OCH3	H	OH	OH
peonidin	OCH3	OH	H	H	OH	OH
petunidin	OH	OH	OCH3	H	OH	OH
cyanidin 3-glucoside	H	OH	OH	H	Glc (2)	OH
cyanidin 3-xyloside	H	OH	OH	H	Xyl (3)	OH
cyanidin 3-sambubioside	H	OH	OH	H	Sam (4)	OH
cyanidin 3-glucuronide	H	OH	OH	H	Glcde (1)	OH
cyanidin 3'-glucuronide	H	OH	OH	H	OH	OH
cyanidin 4'-glucuronide	H	OH	OH	H	OH	OH
cyanidin 7-glucuronide	H	OH	OH	H	OH	OH
cyanidin 3-glucoside 3'-glucuronide	H	OH	Glc (2)	H	OH	OH
cyanidin 3-glucoside 4'-glucuronide	H	Glc (2)	OH	H	Glc (2)	OH
cyanidin 3-glucoside-7-glucuronide	H	OH	OH	H	Glc (2)	Glcde (1)
delphinidin 3-glucoside	OH	OH	OH	H	Glc (2)	OH
malvidin 3-glucoside	OCH3	OH	OCH3	H	Glc (2)	OH
pelargonidin	H	OH	H	H	OH	OH
pelargonidin 3-glucoside	H	OH	H	H	Glc (2)	OH
pelargonidin 3-glucuronide	H	OH	H	H	Glcde (1)	OH
pelargonidin 4'glucuronide	H	Glcde (1)	H	H	OH	OH
pelargonidin 7-glucuronide	H	OH	H	H	OH	Glcde (1)
peonidin 3-glucoside	OCH3	OH	H	H	Glc (2)	OH
peonidin 3-sambubioside	OCH3	OH	H	H	Sam (4)	OH
peonidin 3-glucuronide	OCH3	OH	H	H	Glcde (1)	OH
peonidin 4'-glucuronide	OCH3	Glcde (1)	H	H	OH	OH
peonidin 7-glucuronide	OCH3	OH	H	H	OH	Glcde (1)
petunidin 3-glucoside	OH	OH	OCH3	H	Glc (2)	OH

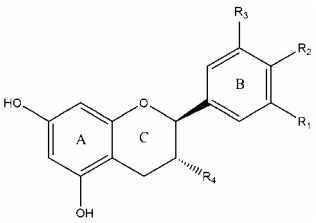
UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

Flavanols

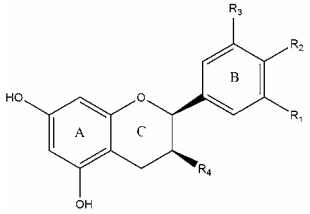
Monomers (*catechins*)



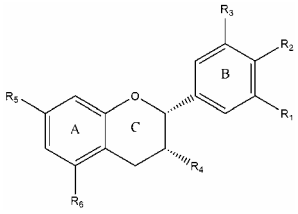
	R1	R2	R3	R4	R5
(+)-catechin	H	OH	OH	OH	OH
(+)-gallocatechin	OH	OH	OH	OH	OH
(+)-catechin 3'-glucuronide	H	OH	Glcde [(1)]	OH	OH
(+)-catechin 3-glucuronide	H	OH	OH	Glcde [(1)]	OH
(+)-catechin 4'-glucuronide	H	Glcde [(1)]	OH	OH	OH
(+)-catechin 7-glucuronide	H	OH	OH	OH	Glcde [(1)]
3'-*O*-methyl-(+)-catechin	H	OH	OCH3	OH	OH
3'-*O*-methyl-(+)-catechin 3-glucuronide	H	OH	OCH3	Glcde [(1)]	OH
3'-*O*-methyl-(+)-catechin 4'-glucuronide	H	Glcde [(1)]	OCH3	OH	OH
3'-*O*-methyl-(+)-catechin 7-glucuronide	H	OH	OCH3	OH	Glcde [(1)]



	R1	R2	R3	R4
(+)-epicatechin	H	OH	OH	OH
(+)-epigallocatechin	OH	OH	OH	OH



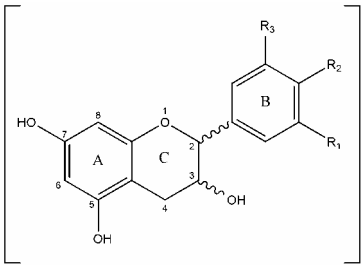
	R1	R2	R3	R4
(-)-catechin	H	OH	OH	OH
(-)-gallocatechin	OH	OH	OH	OH
(-)-catechin 3-gallate	H	OH	OH	Gal [(5)]



	R1	R2	R3	R4	R5	R6
(-)-epicatechin	H	OH	OH	OH	OH	OH
(-)-epigallocatechin	OH	OH	OH	OH	OH	OH
(-)-epicatechin 3-gallate	H	OH	OH	Gal [(5)]	OH	OH
(-)-epicatechin 3'-glucuronide* (9)	H	OH	Glcde [(1)]	OH	OH	OH
(-)-epicatechin 3-glucuronide* (9)	H	OH	OH	Glc [(2)]	OH	OH
(-)-epicatechin 4'-glucuronide* (9)	H	Glcde [(1)]	OH	OH	OH	OH
(-)-epicatechin 7-glucuronide* (9)	H	OH	OH	OH	Glcde [(1)]	OH
4'-*O*-methyl-(-)-epicatechin	H	OCH3	OH	OH	OH	OH
4'-*O*-methyl-(-)-epicatechin 3-glucuronide	H	OCH3	OH	Glcde [(1)]	OH	OH
4'-*O*-methyl-(-)-epicatechin 5-glucuronide	H	OCH3	OH	OH	OH	Glcde [(1)]
4'-*O*-methyl-(-)-epicatechin 7-glucuronide	H	OCH3	OH	OH	Glcde [(1)]	OH
3'-*O*-methyl-(-)-epicatechin 7-glucuronide	H	OH	OCH3	OH	Glcde [(1)]	OH
4'-*O*-methyl-(-)-epigallocatechin	OH	OCH3	OH	OH	OH	OH
(-)-epigallocatechingallate (EGCG)	OH	OH	OH	Gal [(5)]	OH	OH

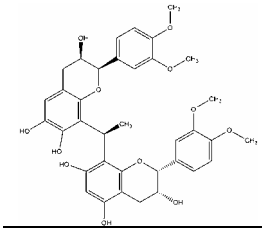
ISBN:978-84-691-1553-4 /DL: T.151-2008

Procyanidins (dimers, trimers and tetramers)



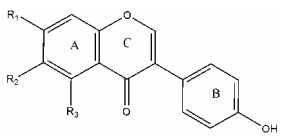
β = R configuration
α = S configuration

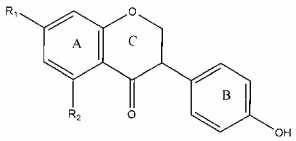
Dimers	
A1	(-)-epicatechin-(4β->8, 2β-*O*-7)-(+)-catechin
A2	(-)-epicatechin-(4β->8, 2β-*O*-7)-(+)-epicatechin
B1	(-)-epicatechin-(4β->8)-(+)-catechin
B2	(-)-epicatechin-(4β->8)-(-)-epicatechin
B3	(+)-catechin-(4α->8)-(+)-catechin
B4	(+)-catechin-(4α->8)-(-)-epicatechin
B5	(-)-epicatechin-(4β->6)-(-)-epicatechin
B6	(+)-catechin-(4α->6)-(+)-catechin
B7	(-)-epicatechin-(4β->6)-(+)-catechin
B8	(+)-catechin-(4α->6)-(-)-epicatechin
Trimers	
C1	(-)-epicatechin-(4β->8)-(-)-epicatechin-(4β->8)-(-)-epicatechin
C2	(+)-catechin-(4α->8)-(+)-catechin-(4α->8)-(+)-catechin
T2	(-)-epicatechin-(4β->8)-(-)-epicatechin-(4β->8)-(+)-catechin
T3	(-)-epicatechin-(4β->8)-(-)-epicatechin-(4β->6)-(+)-catechin
T4	(-)-epicatechin-(4β->6)-(-)-epicatechin-(4β->8)-(-)-epicatechin
T5	(-)-epicatechin-(4β->8)-(-)-epicatechin-(4β->6)-(-)-epicatechin
T6	(-)-epicatechin-(4β->6)-(-)-epicatechin-(4β->8)-(+)-catechin
T7	(+)-catechin-(4α->8)-(+)-catechin-(4α->8)-(-)-epicatechin
Tetramers	
1	(-)-epicatechin-(4β->8)-(-)-epicatechin-(4β->8)-(-)-epicatechin-(4β->8)-(-)-epicatechin
2	(-)-epicatechin-(4β->8)-(-)-epicatechin-(4β->8)-(-)-epicatechin-(4β->8)-(+)-catechin



tetramethylated dimeric

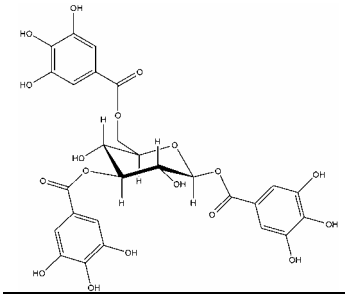
Isoflavones



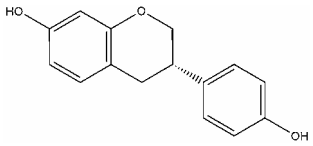


	R1	R2	R3		R1	R2
genistein	OH	H	OH	dihydrogenistein	OH	OH
genistin	Glc [2]	H	OH	dihydrodaidzein	OH	H
daidzein	OH	H	H			
daidzin	Glc [2]	H	H			
glycitein	OH	OCH3	H			
glycitin	Glc [2]	OCH3	H			

OTHER COMPOUNDS



tannic acid

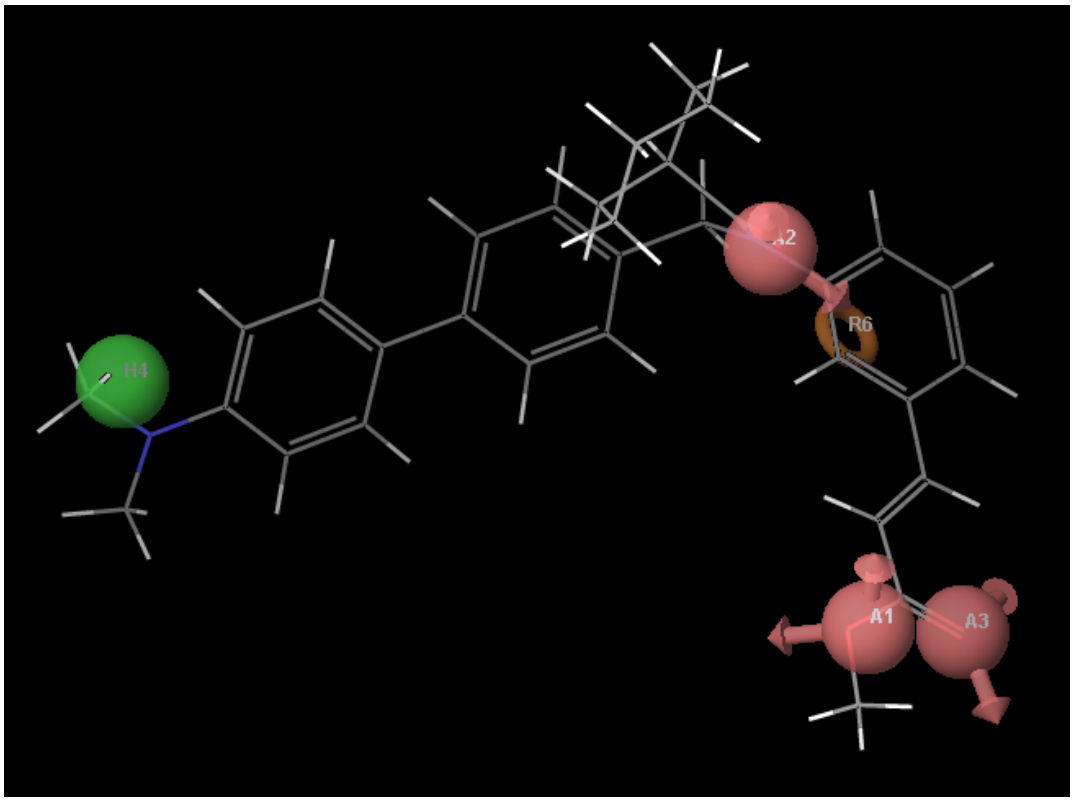


equol

[1] Glucuronide
[2] Glucose
[3] β-xylose
[4] Sambubiose
[5] Gallate

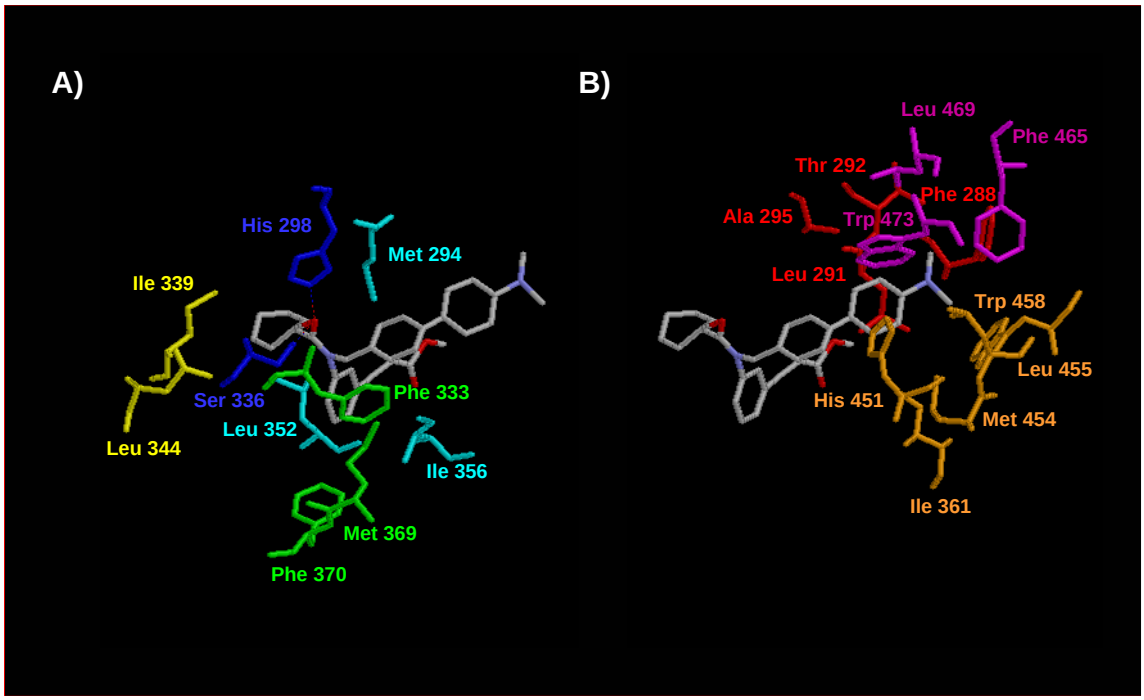
ISBN:978-84-691-1553-4 /DL: T.151-2008

Figure 4. Three-dimensional representation of the hypothesis in the context of FEX.



In this schemathic representation of the hypothesis, the hydrogen bond acceptors ($\mathbf{A_1}$ on the methyl ester oxygen; $\mathbf{A_2}$ on amide carbonyl oxygen; and $\mathbf{A_3}$ on the other methyl ester oxygen) are shown by using pink vector features (*i.e.* indicating the atom in which it carries one or more lone-pair electrons and a vector with an arrow that points in the direction of these lone pairs); the hydrophobic feature (**H**) is represented as a pure projected point in green; and the ring aromatic function (**R**) is represented with a centroid of the aromatic ring in brown.

Figure 5. Interactions between FXR-LBD and FEX according to Downes et al. (2003)



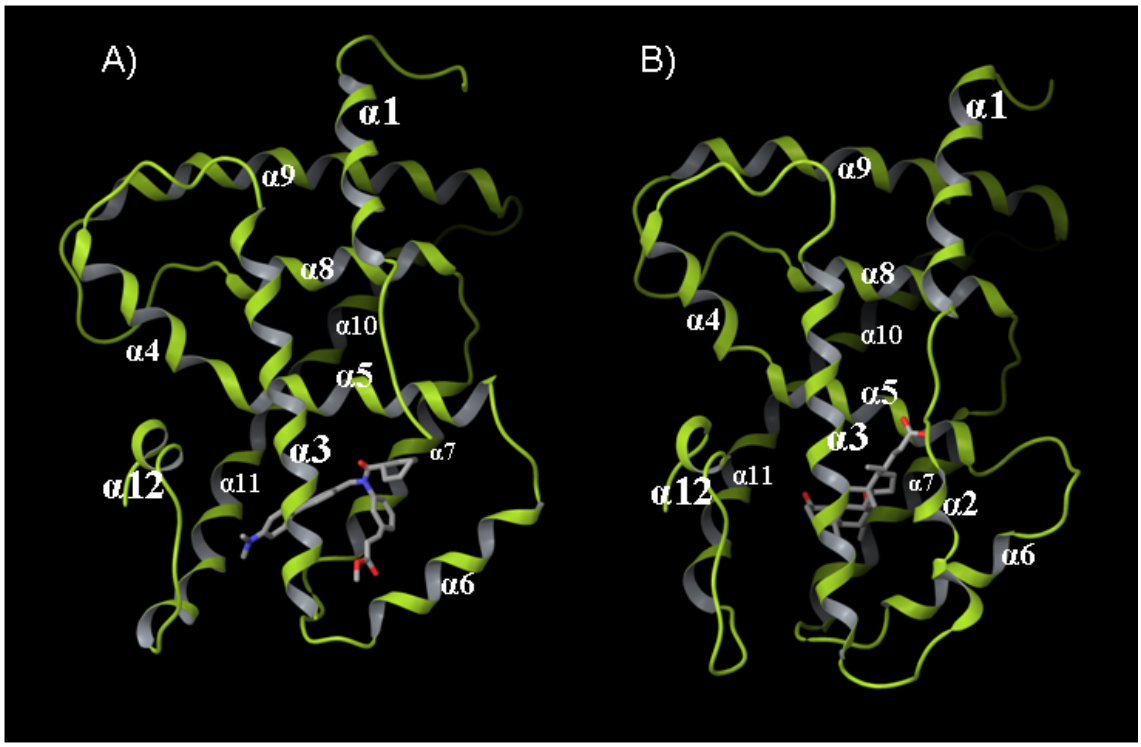
This figure shows the interaction between FEX (wireframe model in CPK colors) and FXR-LBD residues according to Downes at al. (2003) [30]. The color of each residue shows the groups to which it belongs.

The interactions between FEX and FXR-LBD have been divided into two subsets, which suggests that the potency of FEX appears to be mediated by two mechanistic paths. Thus, the first subset contains those interactions that stabilize the position of the following FEX groups: (a) the hexyl ring; (b) the outermost first benzene ring; and (c) the methyl ester moiety (see panel A). These interactions are: (a) some minimal van der Waals contacts between the hexyl group and residues around it (*i.e.* Ile339 and Leu344 from helix 5; in yellow); (b) a hydrophobic surface created by three apolar residues (*i.e.* Met369 and Phe370 from helix 7 and Phe333 from helix 5; in green) behind FEX's central nitrogen and the outermost benzene ring; (c) a hydrophobic surface formed by three apolar amino residues (*i.e.* Leu352 and Ile356 from helix 6 and Met294 from helix 3; in cyan) and the aliphatic linker between the last benzene ring and the methyl ester moiety; and (d) two hydrogen bonds in the amide carbonyl oxygen made up of residues His298 and Ser336, which stabilize the position of the methyl ester moiety in the neutral groove between helices 3 and 6 (in blue). The second subset of interactions that has been described between FXR-LBD and FEX, is thought to be responsible for the stabilization of the biaryl rings and of the dimethylamine moiety of FEX (see panel B). In this respect, it has been reported that the sequential hydrophobic ring structures of FEX penetrate deeper into the ligand

ISBN:978-84-691-1553-4 /DL: T.151-2008

binding pocket and increase the number of stable contacts with the LBD. Thus, the interactions that form this second subset are: (a) two hydrophobic surfaces (one at each side of FEX's double ring structure), one of which consists of four residues from helix 3 (*i.e.* Phe288, Leu291, Thr292 and Ala295; in red) and the other of which consists of five residues [four from helix 11 (*i.e.* His451, Met454, Leu455 and Trp458) and one from loop7 (*i.e.* Ile361); in orange]; and (b) a deep hydrophobic pocket that is filled by the biaryl moiety and which is formed by the mentioned hydrophobic surfaces from helix 11 and helix 3 that are bridged by Leu469 and Trp473 from helix 12 and by Phe465 from loop 12 (in magenta). This figure has been drawn with RasMol (http://www.openrasmol.org/) and the PDB file 1OSH.

Figure 6. Crystal structure of FXR-LBD bound to FEX and to 6ECDCA



Cartoon models from: **(A)** the FXR-LBD (obtained from the PDB file 1OSH) that contain coordinates for residues 245 to 270 and 286 to 476 from the complete human FXR sequence and complexed with the high-affinity FEX agonist; **(B)** the FXR-LBD (obtained from the PDB file 1OSV) that contain coordinates for residues 240 to 468 from the complete *Rattus norvegicus* FXR sequence and complexed with the semisynthetic bile acid 6ECDCA agonist. The ligands are shown in wireframe format and colored according to the CPK criteria. The helices are numbered according to the canonical structure for the LBD of nuclear receptors (the size of the label also indicates the proximity of each helix to the reader). This figure has been drawn with Maestro (http://www.schrodinger.com/) and the PDB files 1OSH (panel A) and 1OSV (panel B)

ISBN:978-84-691-1553-4 /DL: T.151-2008

Figure 7. Structural alignment of the FXR's agonists used to generate the 3D-QSAR model

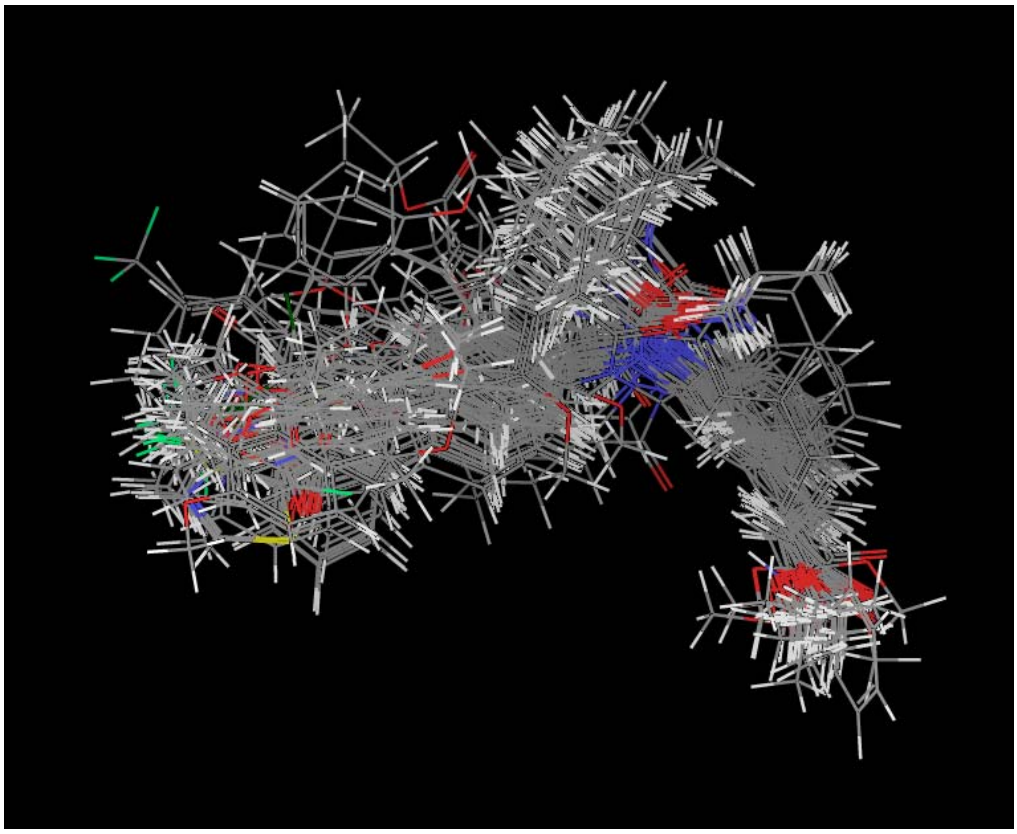
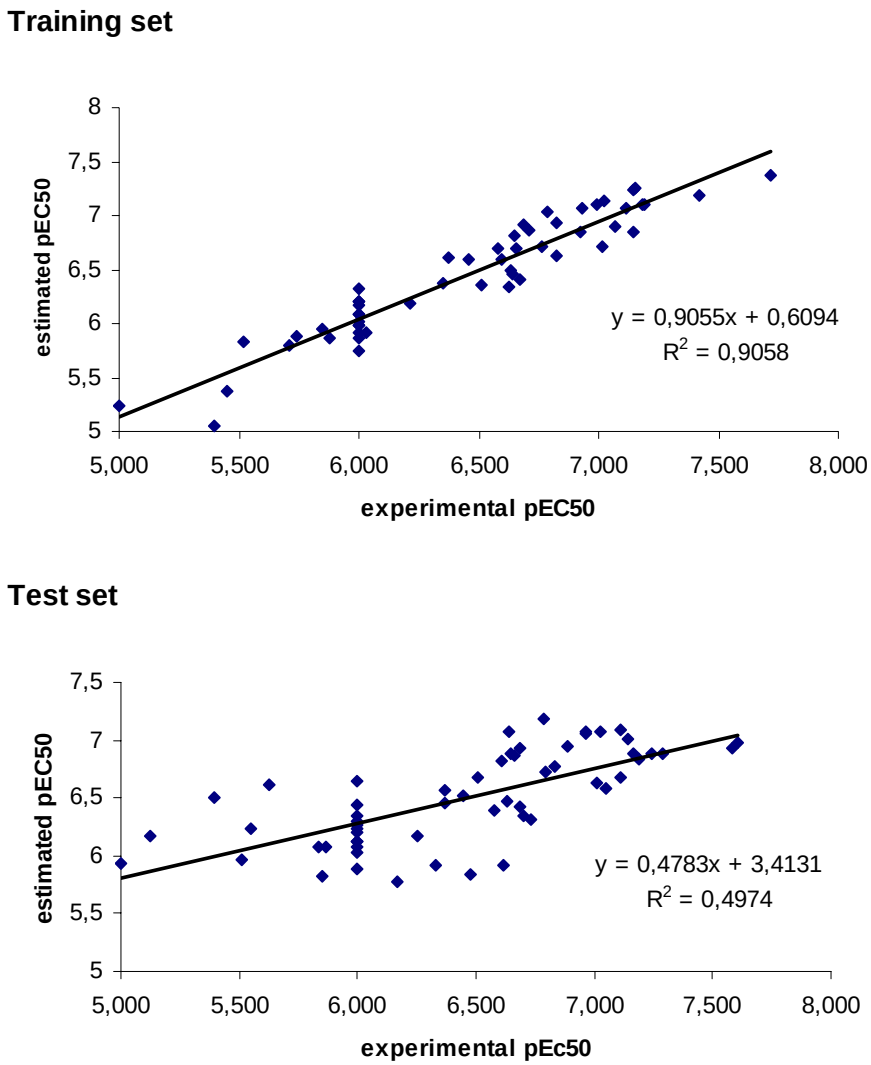


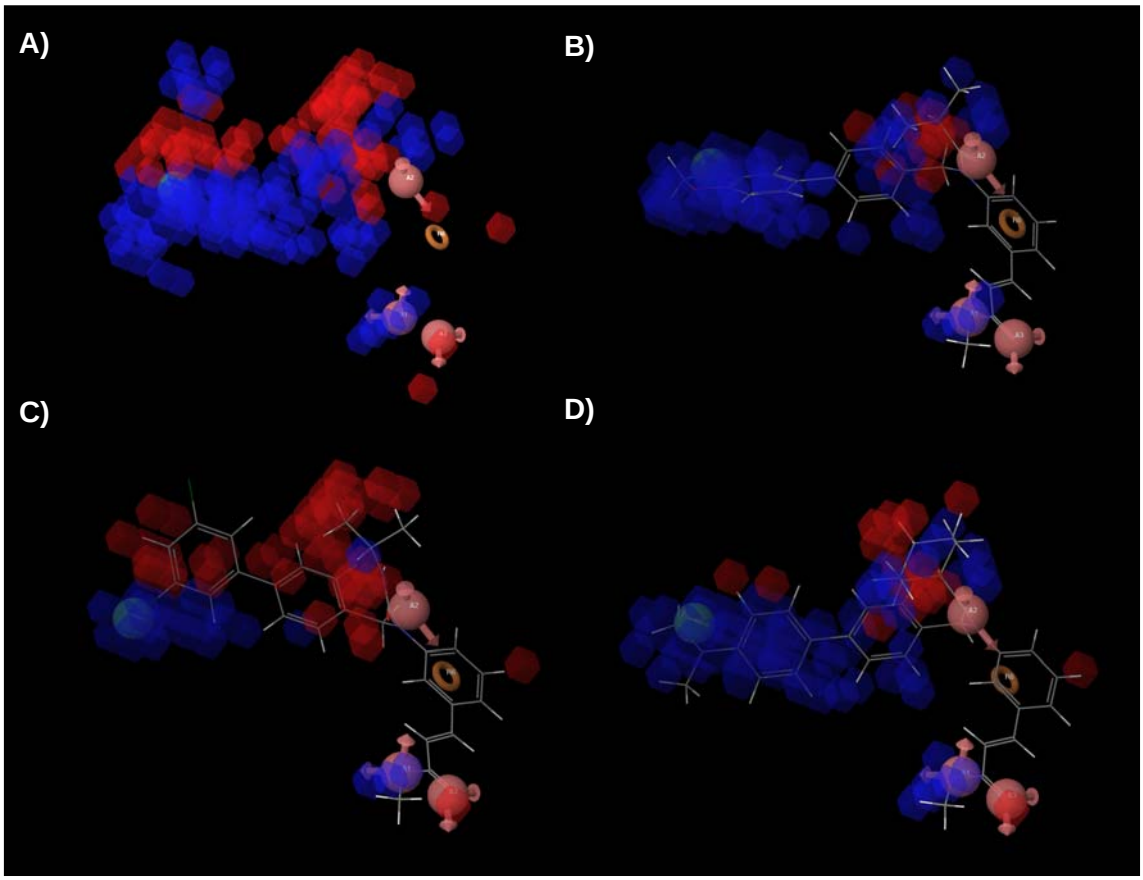
Figure 8. Regression of predicted versus experimental activities for the training and test set molecules



Each graph shows the linear regression that is obtained when the pEC50 value predicted by appliying the 3D-QSAR model is plotted versus the corresponding experimental pEC50. The training set correlation has been obtained with three partial least-square (PLS) factors and is characterized by: (a) SD = 0.18; and (b) R^2 = 0.91. On the other hand, the test set correlation has been also obtained with three PLS factors and is characterized by: (a) root mean-square error (RMSE) = 0.42; (b) q^2 = 0.49; and (c) Pearson-R = 0.71

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marquès
ISBN:978-84-691-1553-4 /DL: T.151-2008

Figure 9. Spatial comparison of the 3D-QSAR model, the hypothesis and the poses for three ligands with different FXR-agonist activity

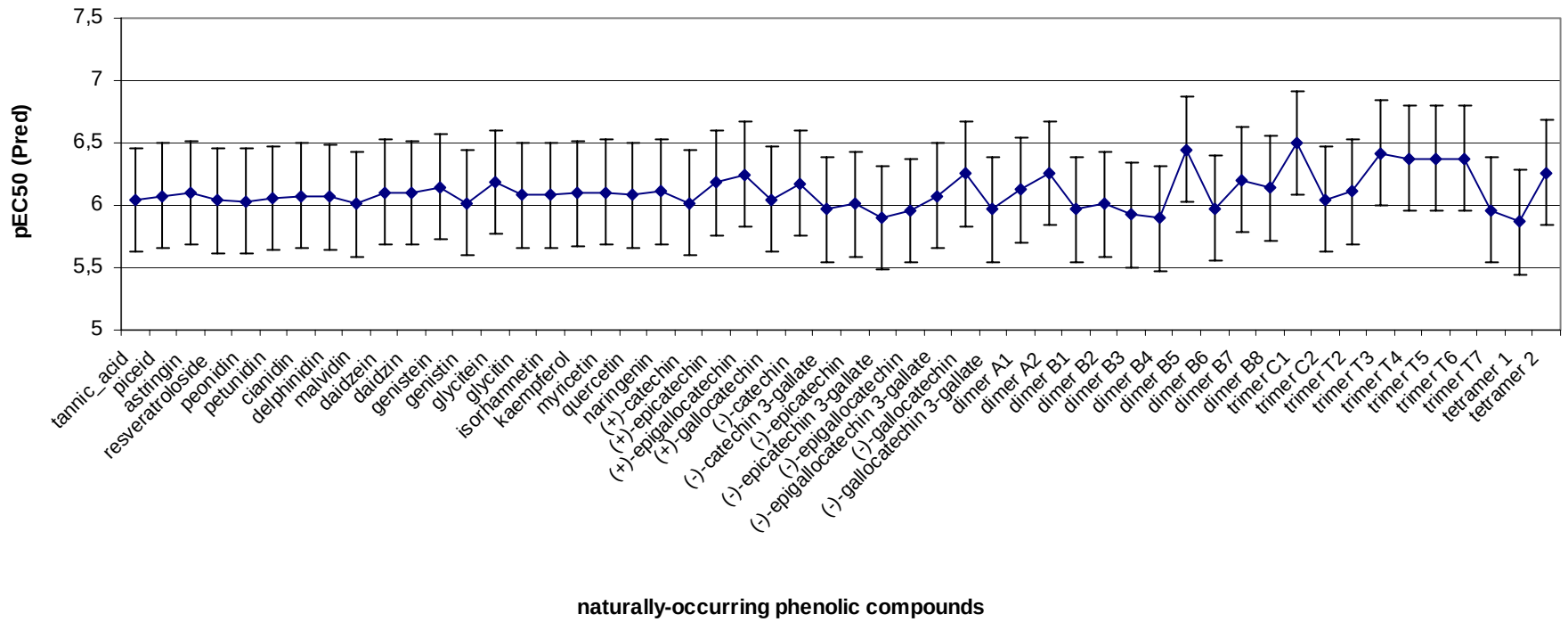


Schemathic representation of the 3D-QSAR model (panel A) and the cubic volume-elements that are occupied by three selected ligands [*i.e.* the most active (**245**; panel B); the least active (**222**; panel C); and FEX (**259**; panel D)]. Hence, blue cubes indicate regions that are favorable for activity whereas red cubes indicate regions that are unfavorable for activity. The schemathic representation of the hypothesis allows to identify also the relative locations of the functions that have been described by the active compounds. See Figure 4 for the interpretation of the hypothesis features.

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF
PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

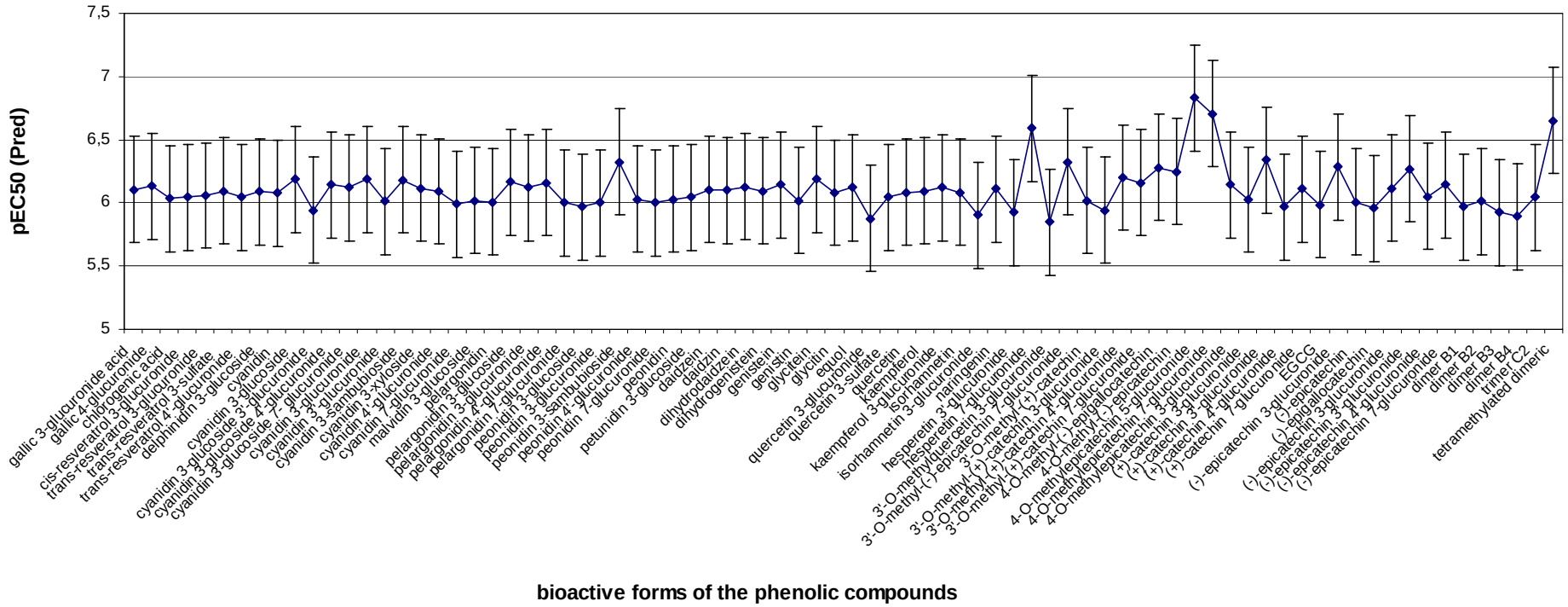
Figure 10. Predicted pEC50 of the molecules from our phenolic compounds database when their FXR-agonist activity is calculated with the 3D-QSAR model shown in Figure 9.

A



UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF
PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

B



bioactive forms of the phenolic compounds

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

General discussion

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

This thesis was written with the aim of applying computational methods that have already been developed for molecular design and simulation (*i.e.* pharmacophore generation and protein-ligand docking) to nutrigenomics. So, *in silico* tools that are routinely used by the pharmaceutical industry to develop drugs [1] have been used to understand, at the molecular level, how natural products such as phenolic compounds (*i.e.* molecules that are commonly found in fruits and vegetables [2-4]) can improve health and prevent diseases [5-7].

Therefore, we first focused on predicting the structure of protein-ligand complexes. In this respect, under highly expert control time-consuming techniques such as X-ray crystallography or NMR enable the experimental structures of the protein/ligand complexes to be obtained. However, they are not suitable for routinely screening the possible interaction between one receptor and the thousands of bioactive ligands of interest for the functional food industry. In this respect, *in silico* tools (*i.e.* docking algorithms) can use the individual structures from receptor and ligand to predict (1) whether they can form a complex and (2) if so, the structure of the resulting complex. This prediction can be made, for instance, with AutoGrid/AutoDock [8-11], the most cited docking software in the literature [12]. It is a suite of automated docking tools which was designed to predict how small molecules, such as substrates or drug candidates, bind to a receptor of known 3D structure. The automation of AutoGrid/AutoDock is not trivial for tasks such as (1) the virtual screening of a library of ligands against a set of possible receptors; (2) the use of receptor flexibility and (3) making a blind-docking experiment with the whole receptor surface. Therefore, in order to circumvent these limitations, we have designed BDT (*i.e.* **b**lind-**d**ocking **t**ester; http://www.quimica.urv.cat/~pujadas/BDT), an easy-to-use graphic interface for using AutoGrid/AutoDock [13]. BDT is a Tcl/Tk graphic front-end application that runs on top of four Fortran programs (*i.e.* **make_grids**, **combine_grids**, **make_docks** and **analyze**, one under each BDT window tab) and which controls the conditions of the AutoGrid and AutoDock runs. BDT therefore provides easy access to AutoGrid/AutoDock docking tools and to the use of sophisticated docking strategies. It also extends the use of this kind of *in silico* experiments to research teams in the fields of Chemistry and the Life Sciences who are interested in docking but do not have enough programming skills.

Classical studies using phenolic compounds have attributed their beneficial effect on health to their influence on the regulation of such processes as the modulation of glucose and cholesterol metabolism and changes in the lipid plasmatic profile [14, 15]. As far as the modulation of the glucose metabolism is concerned, several *in vivo* and *in vitro* results obtained by our group have shown that grape seed procyanidin extracts (GSPE) stimulate glucose uptake in 3T3-L1 adipocytes and thus help to maintain their glucose homeostasis [14]. In contrast, it is also well known that although some phenolic compounds (*i.e.* catechin and gallic acid) do not affect

ISBN:978-84-691-1553-4 /DL: T.151-2008

glucose uptake, others (*i.e.* quercetin, myricetin, catechin-gallate, resveratrol and naringenin) reversibly inhibit it in isolated rat adipocytes, 3T3-L1 adipocytes, rat L6 myotubes and human muscle-derived cell lines [16-18]. Moreover, for at least some of these phenolic compounds, this inhibition is the result of their competition with ATP for the ATP-binding site in p110α (*i.e.* the α isoform of the catalytic subunit of phosphoinositide 3-kinase or PI3Kα) [16]. Furthermore, recent studies with isoform-specific inhibitors have identified p110α as the crucial isoform for insulin-stimulated glucose-uptake in such cell lines as 3T3-L1 adipocytes, L6 myotubes and CHO-IR [19, 20]. Therefore, although it has been proved that the addition of phenolic compound extracts to food can have an overall benefit on health, it should be taken into account that some of these molecules may exacerbate insulin resistance in susceptible individuals via impaired glucose uptake in muscle and adipose tissues and, therefore, produce an undesirable side effect. In this context, we have applied computational approaches (*i.e.* protein-ligand docking and 3D-QSAR) to predict the IC50 (*i.e.* the concentration that reduces the p110α activity to 50%) for: (a) the most frequent phenolic compounds found in plant extracts [21]; and (b) the bioactive structures of the phenolic compounds detected in plasma or urine [22-25] (these molecules have been added to the study because they may contribute to the biological effects of the phenolic compounds on p110α). Our results are in good aggreement with previous experimental results [16-18] and predict that stilbenes, flavonols (except myricetin), flavones, flavanones, anthocyanidins (except delphinidin and delphinidin 3-glucoside), most flavanol monomers [except (+)-epicatechin, (+/-)-epigallocatechin, (+/-)-gallocatechin, (-)-catechin and 4'-*O*-methyl-(-)-epigallocatechin], all procyanidin dimers and some isoflavones (*i.e.* daidzin, genistin, glycitin and glycitein) can map most of the features of a pharmacophore built with data from synthetic p110α inhibitors [26, 27]. Therefore, they are potential inhibitors of this enzyme. However, phenolic acids, the anthocyanidins delphinidin and delphinidin 3-glucoside, some flavanol monomers [*i.e.* (+)-epicatechin, (+/-)-epigallocatechin, (+/-)-gallocatechin and (-)-catechin and 4'-*O*-methyl-(-)-epigallocatechin], all procyanidin trimers and tetramers, some isoflavones (*i.e.* genistein, dihydrogenistein, daidzein and dihydrodaidzein), tannic acid and the isoflavandiol equol do not significantly inhibit p110α activity. Remarkably, these results show that the bioactive forms found in plasma or urine [2, 22-25, 28-31] do not substantially alter the activity on p110α of the original molecule from which they are derived.

The positive health effects of phenolic compounds have also been related to mechanisms that modify transcriptional activities [15]. Thus, recent results in our research group have demonstrated that the phenolic compounds in GSPE increase the activity of the farnesoid X receptor (*i.e.* FXR; a nuclear receptor involved in bile acid metabolism and the control of cholesterol and triglyceride metabolism [32, 33]) in a dose-dependent way when the natural ligand of FXR (*i.e.* CDCA) is also present [34]. In this respect, the phenolic compounds might

induce specific conformational changes that increase FXR activity and then contribute to cardioprotection through mechanisms that are independent of their intrinsic antioxidant capacities [35] but that involve direct interaction with FXR to modulate gene expression. Therefore, taking into account this hypothesis and the previously mentioned experimental results, a 3D-QSAR analysis was made in an attempt to understand how phenolic compounds activate FXR. So, our results explain why phenolic compounds cannot activate FXR by themselves and how they can add new interactions to stabilize the active conformation of FXR when its natural ligand (*i.e.* CDCA) is present. So, none of the phenolic compounds can simultaneously fit in all five sites or chemical features that are defined by the 3D-QSAR model derived from the FXR-activator capacity of a set of synthetic non-steroidal FXR agonists, which bind on a binding site that is close to but different from that of the steroidal ligand (6CDCA, CDCA, etc.) [36]. This would explain why phenolic compounds are not predicted as highly active agonists. Thus, the solutions of the "partial matching" show that the phenolic compounds can map four out of five chemical functions of the 3D-QSAR model. Interestingly they cannot map the hydrophobic surface which has been described to be essential for: (1) achieving the proper conformation of the ligand binding pocket; (2) stabilizing helix 12; and (3) enhancing the binding affinity of the coactivator peptide [37-39]. Therefore, we proposed a mechanism of FXR activation by dietary phenolic compounds in which they may enhance bile acid-bound FXR activity. This mechanism suggests how they can lower triglyceride levels when the natural ligand is also present (which would be the natural situation in the cell). Interestingly, our results also show that some bioactive forms of the phenolic compounds have a pattern of interactions with FXR that is different from that of the original molecules from which they are derived. Thus, the pattern of interactions with FXR is different in four bioactive forms (*i.e.* hesperetin 7-glucuronide, 4'-*O*-methyl-(-)-epicatechin 5-glucuronide, 4'-*O*-methyl-(-)-epicatechin 7-glucuronide and tetramethylated dimeric). These forms derive from the absorption and metabolization of the naturally-occurring phenolic compounds, and are predicted to be highly/moderately active agonists. At this point, it is worth pointing out that although our 3D-QSAR model might explain the increase in the FXR activity when the natural agonist and a phenolic compound are both bound to FXR, we have no information about the conformational changes that might affect the nuclear receptor. Therefore, we are suggesting a theoretical model that should be confirmed by additional approaches for predicting how the binding of two ligands affects the structure of the ligand binding domain of FXR.

ISBN:978-84-691-1553-4 /DL: T.151-2008

Therefore, this PhD demonstrates the utility of the application of such *in silico* approaches in nutritional sciences as nutrigenomics and nutrigenetics. Thus, they can be used to help us to: (1) understand how bioactive molecules in food improve health conditions and prevent diseases like diabetes, obesity, cardiovascular pathologies and cancer; and (2) predict which non-experimentally tested phytochemistry ligands would be most effective against a pre-defined protein or gene target [40]. Moreover, new methodologies and strategies based on computational approaches must be developed and used to create nutritional products (*i.e.* food supplements) and semisynthetic analogs that have a substantial protective capacity and produce minimal adverse side effects.

References

[1] T. Langer, R.D. Hoffmann, Pharmacophores and Pharmacophore Searches, in: W. WILEY-VCH Verlag GmbH & Co. KGaA, Germany (Ed.), vol. 32, 2006.

[2] A. Scalbert, G. Williamson, Dietary intake and bioavailability of polyphenols., J Nutr 130 (2000) 2073S-2085S.

[3] A. Fleuriet, J.J. Macheix, Phenolic Acids in Fruits and Vegetables, Flavonoids in Health and Disease, 2003, pp. 1-42.

[4] Documentation for the Update of the USDA Database for Flavonoid Content of Selected foods, Release 2.1 (2007).

[5] S.E. Rasmussen, H. Frederiksen, K. Struntze Krogholm, L. Poulsen, Dietary proanthocyanidins: occurrence, dietary intake, bioavailability, and protection against cardiovascular disease., Mol Nutr Food Res 49 (2005) 159-174.

[6] W.G. Li, X.Y. Zhang, Y.J. Wu, X. Tian, Anti-inflammatory effect and mechanism of proanthocyanidins from grape seeds., Acta Pharmacol Sin 22 (2001) 1117-1120.

[7] X. Terra, J. Valls, X. Vitrac, J.M. Mérrillon, L. Arola, A. Ardèvol, C. Bladé, J. Fernandez-Larrea, G. Pujadas, J. Salvadó, M. Blay, Grape-seed procyanidins act as antiinflammatory agents in endotoxin-stimulated RAW 264.7 macrophages by inhibiting NFkB signaling pathway., J Agric Food Chem 55 (2007) 4357-4365.

[8] D.S. Goodsell, A.J. Olson, Automated docking of substrates to proteins by simulated annealing., Proteins 8 (1990) 195-202.

[9] D.S. Goodsell, G.M. Morris, A.J. Olson, Automated docking of flexible ligands: applications of AutoDock., J Mol Recognit 9 (1996) 1-5.

[10] G.M. Morris, D.S. Goodsell, R. Huey, A.J. Olson, Distributed automated docking of flexible ligands to proteins: parallel applications of AutoDock 2.4., J Comput Aided Mol Des 10 (1996) 293-304.

[11] G.M. Morris, D.S. Goodsell, R.S. Halliday, R. Huey, W.E. Hart, R.K. Belew, A.J. Olson, Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function, J Comput Chem 19 (1998) 1639-1662.

[12] S.F. Sousa, P.A. Fernandes, M.J. Ramos, Protein-ligand docking: Current status and future challenges, Proteins 65 (2006) 15-26.

[13] M. Vaque, A. Arola, C. Aliagas, G. Pujadas, BDT: an easy-to-use front-end application for automation of massive docking tasks and complex docking strategies with AutoDock, Bioinformatics 22 (2006) 1803-1804.

[14] M. Pinent, M. Blay, M.C. Bladé, M.J. Salvadó, L. Arola, A. Ardévol, Grape seed-derived procyanidins have an antihyperglycemic effect in streptozotocin-induced diabetic rats and insulinomimetic activity in insulin-sensitive cell lines., Endocrinology 145 (2004) 4985-4990.

[15] J.M. Del Bas, J. Fernández-Larrea, M. Blay, A. Ardèvol, M.J. Salvadó, L. Arola, C. Bladé, Grape seed procyanidins improve atherosclerotic risk index and induce liver CYP7A1 and SHP expression in healthy rats., FASEB J 19 (2005) 479-481.

[16] S. Fröjdö, D. Cozzone, H. Vidal, L. Pirola, Resveratrol is a class IA phosphoinositide 3-kinase inhibitor., Biochem J (2007).

[17] A.W. Harmon, Y.M. Patel, Naringenin inhibits phosphoinositide 3-kinase activity and glucose uptake in 3T3-L1 adipocytes., Biochem Biophys Res Commun 305 (2003) 229-234.

[18] P. Strobel, C. Allard, T. Perez-Acle, R. Calderon, R. Aldunate, F. Leighton, Myricetin, quercetin and catechin-gallate inhibit glucose uptake in isolated rat adipocytes., Biochem J 386 (2005) 471-478.

[19] C. Chaussade, G.W. Rewcastle, J.D. Kendall, W.A. Denny, K. Cho, L.M. Grønning, M.L. Chong, S.H. Anagnostou, S.P. Jackson, N. Daniele, P.R. Shepherd, Evidence for functional redundancy of class IA PI3K isoforms in insulin signalling., Biochem J 404 (2007) 449-458.

[20] Z.A. Knight, B. Gonzalez, M.E. Feldman, E.R. Zunder, D.D. Goldenberg, O. Williams, R. Loewith, D. Stokoe, A. Balla, B. Toth, T. Balla, W.A. Weiss, R.L. Williams, K.M.

ISBN:978-84-691-1553-4 /DL: T.151-2008

Shokat, A pharmacological map of the PI3-K family defines a role for p110 alpha in insulin signaling, Cell 125 (2006) 733-747.

[21] J.J. Macheix, A. Fleuriet, J. Billot, Fruit Phenolics, Boca Raton, FL: CRC Press 1990.

[22] C. Felgines, S. Talavéra, M.P. Gonthier, O. Texier, A. Scalbert, J.L. Lamaison, C. Rémésy, Strawberry anthocyanins are recovered in urine as glucuro- and sulfoconjugates in humans., J Nutr 133 (2003) 1296-1301.

[23] C. Felgines, S. Talavera, O. Texier, A. Gil-Izquierdo, J.L. Lamaison, C. Remesy, Blackberry anthocyanins are mainly recovered from urine as methylated and glucuronidated conjugates in humans., J Agric Food Chem 53 (2005) 7721-7727.

[24] C. Manach, G. Williamson, C. Morand, A. Scalbert, C. Rémésy, Bioavailability and bioefficacy of polyphenols in humans. I. Review of 97 bioavailability studies., Am J Clin Nutr 81 (2005) 230S-242S.

[25] C. Tsang, C. Auger, W. Mullen, A. Bornet, J.M. Rouanet, A. Crozier, P.L. Teissedre, The absorption, metabolism and excretion of flavan-3-ols and procyanidins following the ingestion of a grape seed extract by rats., Br J Nutr 94 (2005) 170-181.

[26] M. Hayakawa, H. Kaizawa, H. Moritomo, T. Koizumi, T. Ohishi, M. Okada, M. Ohta, S. Tsukamoto, P. Parker, P. Workman, M. Waterfield, Synthesis and biological evaluation of 4-morpholino-2-phenylquinazolines and related derivatives as novel PI3 kinase p110alpha inhibitors., Bioorg Med Chem 14 (2006) 6847-6858.

[27] M. Hayakawa, H. Kaizawa, K. Kawaguchi, N. Ishikawa, T. Koizumi, T. Ohishi, M. Yamano, M. Okada, M. Ohta, S. Tsukamoto, F.I. Raynaud, M.D. Waterfield, P. Parker, P. Workman, Synthesis and biological evaluation of imidazo[1,2-a]pyridine derivatives as novel PI3 kinase p110alpha inhibitors., Bioorg Med Chem 15 (2007) 403-412.

[28] S.A. Aherne, N.M. O'Brien, Dietary flavonols: chemistry, food content, and metabolism., Nutrition 18 (2002) 75-81.

[29] C. Manach, A. Scalbert, C. Morand, C. Rémésy, L. Jiménez, Polyphenols: food sources and bioavailability., Am J Clin Nutr 79 (2004) 727-747.

[30] G. Williamson, C. Manach, Bioavailability and bioefficacy of polyphenols in humans. II. Review of 93 intervention studies., Am J Clin Nutr 81 (2005) 243S-255S.

[31] F. Saura-Calixtoa, J. Serrano, I. Goñi, Intake and bioaccessibility of total polyphenols in a whole diet, Food Chem 101 (2007) 492.

[32] R. Pellicciari, G. Costantino, S. Fiorucci, Farnesoid X receptor: from structure to potential clinical applications., J Med Chem 48 (2005) 5383-5403.

[33] B. Cariou, B. Staels, FXR: a promising target for the metabolic syndrome?, Trends Pharmacol Sci 28 (2007) 236-243.

[34] J.M. Del Bas, Modulation of hepatic lipoprotein metabolism by dietary procyanidins, Departament de Bioquímica i Biotecnologia, Universitat Rovira i Virgili, Tarragona, 2007, p. 218.

[35] K.E. Heim, A.R. Tagliaferro, D.J. Bobilya, Flavonoid antioxidants: chemistry, metabolism and structure-activity relationships., J Nutr Biochem 13 (2002) 572-584.

[36] K.C. Nicolaou, R.M. Evans, A.J. Roecker, R. Hughes, M. Downes, J.A. Pfefferkorn, Discovery and optimization of non-steroidal FXR agonists from natural product-like libraries, Org Biomol Chem 1 (2003) 908-920.

[37] M. Downes, M.A. Verdecia, A.J. Roecker, R. Hughes, J.B. Hogenesch, H.R. Kast-Woelbern, M.E. Bowman, J.L. Ferrer, A.M. Anisfeld, P.A. Edwards, J.M. Rosenfeld, J.G. Alvarez, J.P. Noel, K.C. Nicolaou, R.M. Evans, A chemical, genetic, and structural analysis of the nuclear bile acid receptor FXR., Mol Cell 11 (2003) 1079-1092.

[38] L.Z. Mi, S. Devarakonda, J.M. Harp, Q. Han, R. Pellicciari, T.M. Willson, S. Khorasanizadeh, F. Rastinejad, Structural basis for bile acid binding and activation of the nuclear receptor FXR., Mol Cell 11 (2003) 1093-1100.

[39] K.W. Nettles, G.L. Greene, Nuclear receptor ligands and cofactor recruitment: is there a coactivator "on deck"?, Mol Cell 11 (2003) 850-851.

[40] J.M. Rollinger, T. Langer, H. Stuppner, Integrated in silico tools for exploiting the natural products' bioactivity., Planta Med 72 (2006) 671-678.

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

Conclusions

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

- We have designed BDT, an easy-to-use graphic interface for AutoGrid/AutoDock (which is currently the most cited protein-ligand docking software in the literature). The automation of AutoGrid/AutoDock is not trivial for tasks such as (1) the virtual screening of a library of ligands against a set of possible receptors; (2) the use of receptor flexibility; and (3) making a blind docking experiment with the whole receptor surface. So we have developed BDT, a Tcl/Tk graphic front-end application that runs on top of four Fortran programs (*i.e.* **make_grids**, **combine_grids**, **make_docks** and **analyze**; one under each BDT window tab), which control the conditions of AutoGrid and AutoDock runs. Therefore, BDT can be used by research teams in the fields of Chemistry and the Life Sciences who are interested in conducting this kind of protein-ligand docking experiments but do not have enough programming skills to do so.

- Generating of phenolic compound conformations by docking enables conformations that are incompatible with the ligand-binding site structure to be discounted. Our study highlights the importance of docking for obtaining the conformations of the ligands that can be used to map an existing pharmacophore. Thus, the *in vacuo* generated conformations of some ligands (*e.g.* procyanidin tetramers) map the pharmacophore quite well but docking demonstrates that they cannot bind to p110α because they are too large.

- The homology model of p110α is valid for predicting how phenolic compounds bind to its ATP-binding site. The model of p110α incorporates receptor-ligand steric clashes that, when taken into account during the docking: (a) produce ligand conformations that can predict experimentally reported effects on glucose uptake; and (b) explain preliminary results from our research group that suggest that trimeric flavanols are not p110α inhibitors because, in fact, they stimulate glucose uptake (which is not predicted if the steric hindrance of the receptor is not considered).

- Although it has been proved that the addition of phenolic compound extracts to food can have an overall benefit on health, it should be taken into account that some of these molecules may exacerbate insulin resistance in susceptible individuals via impaired glucose uptake in muscle and adipose tissues and, therefore, produce an undesirable side effect. Stilbenes, flavonols (except myricetin), flavones, flavanones, anthocyanidins (except delphinidin and delphinidin 3-glucoside), most flavanol monomers [except (+)-epicatechin, (+/-)-epigallocatechin, (+/-)-gallocatechin, (-)-catechin and 4'-*O*-methyl-(-)-epigallocatechin], all procyanidin dimers and some isoflavones (*i.e.* daidzin, genistin, glycitin and glycitein) can map

ISBN:978-84-691-1553-4 /DL: T.151-2008

most of the features of a pharmacophore built with data from synthetic p110α inhibitors and, therefore, they are potential inhibitors of this enzyme.

- The bioactive forms of the phenolic compounds do not substantially alter the activity of the original molecules from which they are derived when the activities on p110α are compared. Thus, the glucuronidation, sulfation, glycosilation and methylation does not substantially alter the activity on p110α of the original molecule from which the bioactive forms are derived.

- Phenolic compounds cannot activate FXR by themselves, but they might add new interactions that stabilize the active conformation of FXR when its natural ligand is also present. None of the phenolic compounds fit in the five sites or chemical features defined by the 3D-QSAR model, which meant that phenolic compounds could not act as highly active agonists. Therefore, the "partial matching" solutions meant that the phenolic compounds can perform some of the interactions in the 3D-QSAR model (*i.e.* they can fit in four out of the five sites) but they cannot map the hydrophobic surface which has been described to be essential for: (1) achieving the proper conformation of the ligand binding pocket; (2) stabilizing helix 12; and (3) enhancing the binding affinity of the coactivator peptide.

- Some bioactive forms of the phenolic compounds have interaction patterns with FXR that are different from the ones in the original molecule from which they are derived. In four bioactive forms (*i.e.* hesperetin 7-glucuronide, 4'-*O*-methyl-(-)-epicatechin 5-glucuronide, 4'-*O*-methyl-(-)-epicatechin 7-glucuronide and tetramethylated dimeric) the pattern of interactions with FXR is different from that of the molecules from which they are derived.

- *In silico* approaches can be useful for analyzing the intermolecular interactions between phenolic compounds (*i.e* natural products) and proteins (*i.e.* enzymes such as p110α or nuclear receptors such as FXR) and analyzing how these interactions modulate the corresponding target function. Computational tools such as pharmacophore generation and protein-ligand docking allow us to correlate a compound's biological activity (*i.e.* IC50 or EC50) with structural information for deriving 3D quantitative structure-activity relationships (*i.e.* 3D-QSAR studies) and gain insights into the ligand-structural requirements for an increased target affinity and/or selectivity. Therefore, these *in silico* tools together with classical *in vivo* and *in vitro* methods can be applied in research into natural products to study their efficacy in the development of functional foods.

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

Annexes

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

Pharmacophore Generation Software Tool

Index

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

1. What is Catalyst™

Catalyst™ program, v4.11 (Accelrys Software, San Diego, CA, http://www.accelrys.com) is a tool for automated pharmacophore pattern recognition in a collection of compounds based on chemical features correlated with three-dimensional structure and biological activity data.

A Catalyst™ pharmacophore (so called hypothesis or model) consists of sets of chemical features arranged at certain positions in the three-dimensional space. The features definitions are designed to cover different types of functional groups that allow interactions between ligand and target (*e.g.* hydrophobic, H-bond donor, H-bond acceptor, positive ionizable, negative ionizable) [1]. These features are surrounded by certain spatial tolerance spheres, assessing the area in space that should be assigned by the corresponding chemical function of the matched molecule. Hydrogen bond acceptors, donors, and ring aromatic features additionally include a vector, indicating the direction of the interaction [2].

In order to generate a pharmacophore, Catalyst™ has implemented some modules (*i.e.* HypoGen™, HypoRefine™, HipHop™). In this tutorial, only HypoGen™ and HypoRefine™ are used to generate the pharmacophore because with these modules it is possible to predict the biological activity quantitatively.

HypoGen™, is an algorithm implemented in Catalyst™, that uses biological data experimentally obtained (*e.g.* IC50 or EC50) to derive hypotheses that can predict quantitatively the activity of compounds. This module tries to find hypotheses that are common among the active compounds of the training set but do not reflect the inactive ones [2]. HypoGen™ consist of three phases: constructive, subtractive and optimization [1]. In the constructive phase, the training set (*i.e.* the set of molecules that are used to develop the pharmacophore) is divided into one “active” and one “inactive” compounds subsets. Then, the first two most active compounds are used to identify all the pharmacophore candidates that they share. This is done by overlaying systematically all their conformations. After that, only those hypothesis candidates that fit a minimum subset of the chemical features that are present in the remaining active compounds are kept. In the substractive phase, the program inspects the remaining hypotheses and removes those most common to the inactive subset. Compounds are considered inactive when their activity lies 3.5 logarithmic units (this value is user-adjustable) below that of the most active compound. In the optimization phase, a simulated annealing (*i.e.* a technique to find a good solution to an optimization problem by trying random variations of the current solution) is used to improve the predictive power of the hypotheses. Thus, small changes are made to the models

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

and they are scored according to the accuracy in activity estimation tests. Finally, the simplest models that correctly estimate activity are selected and the top N solutions are reported to the user. The method has been described in more detail elsewhere: Guner et al. (2000) and Kurogi et al.(2001) [3, 4]

An important assumption that is made within HypoGen™ is that the higher number of contacts with the receptor (*i.e.* the higher number of chemical features), the higher the resulting activity. Nevertheless, it is well known from practice that often this is not true (*e.g.* large and feature-rich compounds may be barely active because of unfavourable steric interactions). Then, the HypoRefine™ module is an extension of HypoGen™ algorithm, that tries to solve this problem by placing exclusion volumes in key locations that are derived from atoms of well-fitting but inactive compounds [1]. The constructive phase of HypoRefine™ is identical to the one in HypoGen™ whereas the subtractive one is not performed. Then, excluded volume spheres are included in the simulated annealing optimization process. Finally, pharmacophore models that also fit inactive molecules are automatically penalized (*i.e.* their total cost is increased)

In this tutorial will be applied both HypoGen™ and HypoRefine™ modules in order to evaluate whether the low activity of the least active compounds can be explain with steric clashes of these compounds with the receptor or not.

(For more information see http://www.accelrys.com/doc/life/catalyst410/index.html)

2. How to execute Catalyst™

First of all, you have to make a reservation of the Catalyst™ package at **CESCA - Reserves SCF** on the webpage:http://reserves-scf.cesca.es:8085/. Thus, you need to reserve the basic module (called *Visualizer™*) and the modules that you need to build the pharmacophore. In this tutorial we explain how to obtain a pharmacophore by using –HypoGen™ and HypoRefine™ so, you should also reserve both modules.

Note:

For license or technical problems you should contact with Ingrid Barcena i Roig from CESCA (Centre de Supercomputació de Catalunya; e-mail: ingrid@cesca.es; Dept. Assistència Tècnica Tel: +34 93 205 6464, Fax: +34 93 205 6979, Edifici Nexus, Gran Capità, 2-4, 08034; Barcelona). For scientific doubts or questions about the algorithm, you should contact with

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

Katalin Nadassy (katalinn@accelrys.com) from the support services of Accelrys either directly or through Ingrid Barcena i Roig from CESCA.

2.1. Executing Catalyst™ modules remotely on CESCA

To connect with CESCA in order to execute the program, you have to open a terminal and connect to the server where Catalyst™ is installed (*i.e.* encantat.cesca.es) by using the ssh protocol with the -X option. In order to do that, type the following command:

```
$ ssh -X username@encantat.cesca.es
```

(where username needs to be a registered user for encantat.cesca.es)

where pujadas@encantat.cesca.es is the computer's IP where catalyst has been installed in CESCA

2.2 Executing Catalyst™ modules locally

Catalyst™ have to be executed with the C shell (csh). Therefore, in the user's account where Catalyst™ is installed you should have the C shell as the default shell. Then, if you don't have this shell in your computer, you have to install it with the following commands:

```
$ sudo -s
# aptitude install csh
```

After that, you have to change the shell in the user's account where Catalyst™ is installed. For instance, this can be done in Ubuntu by going to the control panel where new user accounts are created and then, going to the Catalyst's account and selecting the csh from the pull-down menu where the user's shell can be selected. Now, you can use Catalyst™ either from a remote computer or from the self Catalyst's account. For the former situation, you have to login into the Catalyst's account by using the ssh protocol with the -X option

```
$ ssh -X catalyst_account@IP
```

where IP is the computer's IP where Catalyst™ has been installed and catalyst_account is the computer's account where the program is installed.

3. How to start Catalyst™

3.1 Creating the operating directory

Firstly, you should create a working directory to run Catalyst™ (except if you want to work with data obtained in a previous Catalyst™ run):

```
% mkdir directory_name
```

where "%" is the C shell prompt

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

3.2 Starting catalyst

Starting Catalyst™ by typing the next command in the terminal:

```
% catalyst
```

VERY IMPORTANT NOTE 1: if you want to work with results of a previous Catalyst™ run, it is important to start the program in the same directory where the older run was started and where its data is saved.

When you run Catalyst™ for the first time, you have to chose between these two options:

(a) ***installing training data***. With this option, Catalyst™ copies the files and directories needed for the tutorials into a subdirectory called *cattrain* that is created under your current directory. With this option, the Stockroom (*i.e.* the Catalyst™ top level window) will contain molecules, hypotheses, etc.

(b) ***not to install***. With this option, the Stockroom will contain only an empty Stockroom.

VERY IMPORTANT NOTE 2:

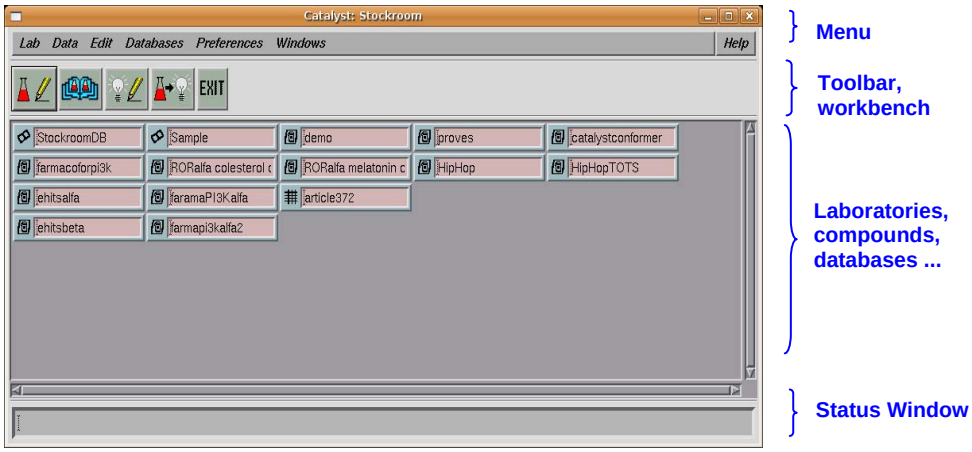
If Catalyst™ accidentally hangs, a warning will appear the next time you start Catalyst™. The warning tells you that the ".lckCatalyst" invisible file (which is inside the catdata folder in the working directory), has to be removed. To do this, write the following command in the terminal:

```
% rm -rf home/name_of_working_directory/catdata/.lckCatalyst
```

4. How to use Catalyst™

4.1. Catalyst™ windows

Catalyst



UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

- ***Catalyst Stockroom:*** is the top level Catalyst™ window and opens when Catalyst™ is started
- ***Menu:*** contains Lab, Data, Edit, Database, Preferences, Windows and Help
- ***Catalyst Toolbar*** (Instrument Buttons): contains the ***Workbench*** (a specific window) that can be launched by clicking with the mouse to a tool button (or an object) in the Catalyst™ Stockroom. Therefore, from left to right, we found:
  - ***View Compound:*** button that access the View Compound Workbench
  - ***View Database:*** button that access the View Database Workbench
  - ***View Hypothesis*:** button that access the View Hypothesis Generation Workbench
  - ***Hypothesis Generation:*** button that access the Hypothesis Generation Workbench
- ***Status Window:*** shows a simple description of the button (or the object) you are pointing with the mouse in the Catalyst™ Stockroom.

4.2. Using Catalyst™ Windows

Catalyst ™ contains different kind of windows to work with. You can move, expand, iconify, save and close these windows:

- Moving and resizing: the title bar and borders of Catalyst™ windows and shell windows act in the same way.
- Expanding: the full screen toggle button acts in the same way.
- Iconifying: to represent all Catalyst™ windows with one icon and not have to close it while you are working in another workbench.
- Closing: click Exit in the Stockroom to leave Catalyst™ and remove all Catalyst™ windows.
- Save Stockroom**:** the Stockroom DB contains all the molecules, hypotheses, laboratories (labs), and spreadsheets that have been saved to the shelf in your current Catalyst™ session. You can save all these objects to disk by saving the Stockroom DB (**Data** -> **save StockroomDB**)
- Data/Dispose from Workbench**:** to remove an object selected from a Catalyst™ workbench window (**Data** -> **Dispose from Stockroom**)

Note:

When you close a workbench clicking the x in the border of the window, it will appear a message: "*Do you really want to dispose of workbench ?* " You have two options: (a) *Perform*

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

Dispose, (b) *Cancel Dispose*. You have taking into account that *Perform Dispose* means that you want to close the workbench without saving the changes or the data.

5. Working with Catalyst™ (creating a laboratory to work)

First of all, you should create a folder (called *Laboratory* in Catalyst™) where you will work and where all your data will be stored. In order to do that, follow the next steps:

- Select **Data** in the menu -> Select **Create Lab** ... -> write a name without numbers nor the following characters ' " /*~`{}^|;:<>?&_' -> **CREATE**
- Open this laboratory (Lab) by double-clicking on its icon (or select it and select **Open** from the **Data** menu)
- One option that you should use when you have a Lab full of elements to tidy it is in **Lab** menu -> **Tidy Lab** (by Object Name, by Object Type or by Object Location)

You can create other Laboratories inside another one in order to organize your data.

6. Ligand preparation and conformational models generation

You have to generate conformational models for all compounds that will be used to build the pharmacophore (the so called *training set*), for compounds that will be used to validate it (the so called *test set*) and for the compounds that you want to predict their IC50 or EC50 activity data.

6.1. Preparing the molecules

The molecular structures of these compounds can be drawn using *View Compound Workbench* in Catalyst™ or importing molecules that have been drawn with another molecular editor.

6.1.1. Import the molecules

Previously, you have to draw all compounds and submitted them to energy minimization. This is possible by using, for instance, the program ChemDraw ultra v10.0 (http://www.cambridgesoft.com/) (see ChemDraw Protocol).

The structures of these compounds have to be saved in **mol** format in order to avoid losing information about molecular structures (*e.g.* double bonds, aromaticity, ...)

- To import the molecules into the lab that has been generated in the previous step you have to use the **Lab** menu and select **Data** -> **Import** -> select the directory where they are stored; then select the molecules (where you can select one by one or all of them by using shift or control) -> **IMPORT**

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

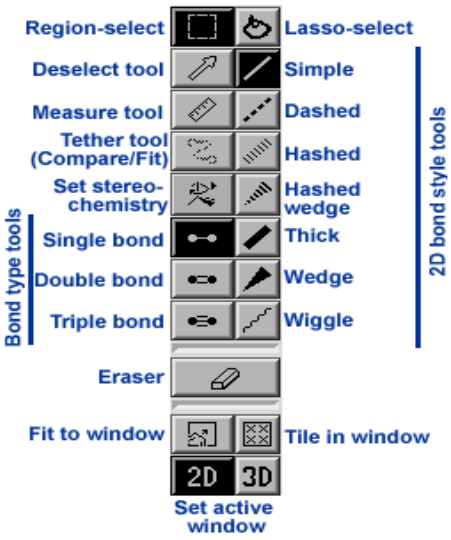
Note:

If you cannot found the directory where your molecules have been deposited you should check that the permissions in the files and directory that contain them are correct.

6.1.2. Building the molecules in Catalyst™

To build molecules directly with Catalyst™, you have to double click on the ***View Compound* Workbench** (in the *Catalyst Toolbar*) and follow the next steps:

- In the screen will appear a ***View Compound workbench*** window (on the right a 3D workspace and on the left a 2D workspace) and a condensed version of the periodic table. Then, for building your molecule you have to select one atom in the periodic table or pre-built chemical groups (*e.g.* aromatics, aliphatics, ...) –and add it by clicking in either workspace (2D or 3D). If you want to add any selected atom or chemical group to your molecule, you only need to make a click in the specific position where it has to go.
- You can also use the tools from the *Toolbox* (on the left of the workspaces) to add a bond (either single, double or triple), to deselect something, to eraser part of the molecule, etc.



- When you have already drawn the molecule in the 2D or 3D workspace, you have to generate a standard 3D structure [by correcting bond lengths and angles, by staggering chain geometry (where possible) and by avoiding significant van der Waals overlaps]. Therefore, the 3D molecule have to tidy up and make it easier to visualize: select **Tools -> Generate Standard 3D**
- To tidy up the 2D molecule: select **Tools -> 2D Beautify**

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

- The structures drown are not necessarily those with the lowest energy. Therefore, if you want to minimize the 3D structure: select **Tools -> 3D Minimize.**
- To save the molecule to the shelf: select **Data -> Save To Lab As** (this command saves the compound only to the current workbench for a temporary use).
  (for more information see Catalyst4.11 Tutorials Chapter 3)

6.2. Generating conformational models interactively

Now, for each molecule, you have to generate a representative set of low-energy conformers (the so called conformational model). Catalyst™ can generate them by following the next steps:

- Select the molecules whose conformers have to be generated by double-clicking in the selected one (or if you want to select more than one compound, then use the shift key or the menu **Edit -> Select all**)
- Drag and drop the selected molecules on the ***View Compound workbench*** button from the ***Stockroom*** (*Toolbar*) (or if it has already opened, then directly drag and drop the compounds on its shelf).
- Select **Tools -> Generate Conformational Model ->** write the Maximum Number of conformers (*i.e.* 255); select Best Quality; write the Energy Range minimum in Kcal/mol (*i.e.* 15 or 20)**-> GENERATE**
- After that, a message will appears ("*Setup for Batch Generate Completed. Batch Process Scheduled or Running, based on specified Start Time. Acknowledged*") and you have to accept it clicking on the **ACKNOWLEDGE** to start the run.

VERY IMPORTANT NOTE 3:

1) Mrs Katalin Nadassy (personal communication), recommend us 255 for the maximum number of conformations if these compounds are then used for hypothesis generation.

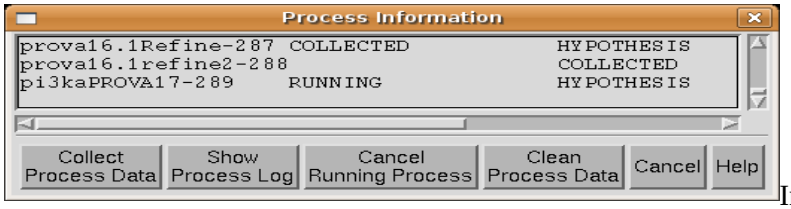
2) About the rang of energy it can be 15 Kcal/mol or 20 kcal/mol

3) Catalyst™ provides two types of conformational search, BEST and FAST. The BEST method is intended to build more precise conformational models of molecules for hypothesis generation. The representative conformers generated by Catalyst™ are not necessary at local minima on the potential surface but are distributed widely over the space. Catalyst™ focuses on the coverage of all possible bioactive conformations of a compound compared with methods that represent conformational space as a collection (clusters) of local minima [1].

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

– To follow the process: select **Data** (preferable in the stockroom of the laboratory where you are working)**-> Process Information**

A window appears showing the history of the process. The process can be in the following steps: QUEUED – REGISTERED - RUNNING - DONE – COLLECTED or DIED:



Information about the options:

- Collect Process Data:it retrieves the conformers and automatically saves them with the compound for which they were generated.
- Show Process Log: it shows the history of the process.
- Cancel running Process:it stops the running Process.
- Clean process Data: it cleans up the process data and removes all files that are created by the background process.
- Cancel: it closes the ***Process Information*** control panel
- **Help**

– When the process is DONE -> select the name of the compound or the name of the process -> click **Collect Process Data** (to bring the conformers that have been generated).

– To see the conformers that has been generated for one compound, select the compound in the shelf or open the compound again and: select **Tools** (in the open ***View Compound workbench***) **-> Show Conformational Model ->** click on the **U** or **E** and the 3D conformer appears. (E = Edit conformer; U = Unregistered conformer).

Even if you don't register the conformers, when you take your compounds as input to generated the hypothesis all of them will be considered.

Therefore, you can close this new window and close the ***View Compound workbench***. The conformations will be related with the corresponding molecule.

– To know the number of conformers generated you can check the ***Status Area*** to see the number of conformational model for each compound when you put the mouse on the name of the one compound.

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

6.3. Save StockroomDB

To save all the objects on the disk:

- Close all workbench windows (you cannot have active workbenches)
- Select **Data** (in the Stockroom) **-> Save Stockroom DB**
- (if you want to export all these compounds, hypothesis or spreadsheets, in another directory: select all you want to export and **Data -> EXPORT)**

7. Training and Test set selection

This step is very important to know if you have enough information to perform a hypothesis in Catalyst™. So, you should have enough compounds to generate these sets:

> Training set: set of compounds that have been assayed for particular activity which Catalyst™ will use to generate a hypothesis that represents the activity of these compounds.
>
> Test set: set of compounds that have not been used to generate the hypothesis. Both Training and Test set have been assayed for particular activity.

When you finish to read the following information about the ideal training set (step 7.1), you have to generate an excel spreadsheet shuch as is specified in the step 7.2 to know if you really have enough compounds to generate a hypothesis with Catalyst™.

7.1. Characteristics for the training set

	Characteristics for an ideal training set
1	18-25 molecules structurally diverse (15 minimum to assure statistical power)
2	the activity have to span 4-5 orders of magnitude, and each order of magnitude have to be represented by 3 or more compounds
3	they have to cover the activity ranges equally and try to choose compounds that each pair of them teaches something to HypoGen™
4	all members of the training set must possess the same binding mode
5	the selected compounds should not provide any redundant information, so compounds with similar activity must be structurally distinct
6	you should have a mixture of active and inactive compounds *****

***Active and inactive compounds:**

It is also very important to examine the compounds in terms of how they will be identified (active, inactive, etc.) during the constructive and substractive phases by HypoGen™ :

1) During the ***constructive phase*** the programme identifies the active compounds using the following formula: **MA*Unc(MA) – A/Unc(A) > 0.0**, where MA is the activity of the

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

most active compounds and the Unc(MA) is the uncertainty associated with that compound. A is the activity of the compound in question and Unc(A) is the uncertainly of that compound.

2) In the ***substractive phase*** the inactive compounds are identified as those where the activity is 3.5 orders of magnitude greater than the most active compound using the following equation: **log (A) – log (MA) > 3.5** where A is the activity of the compound in question, MA is the activity of the most active compound. It is also possible to reduce that value to include more inactive compounds by setting the following .Catalyst parameter: **GenerateHypo.inactive.spread=3.5**

VERY IMPORTANT NOTE 4:

Many adjustable Catalyst™ parameters can be changed by using the .Catalyst file, which is read when the Catalyst™ interface or a background job is started. Catalyst™ looks for the .Catalyst file first in the run directory and then in your home directory. Only one file is used per session. Since the .Catalyst file is read only at startup, if you want to change a parameter you must do so before starting Catalyst™. Also, you should be aware that Catalyst™ rewrites the .Catalyst file whenever the Stockroom is saved.

7.2. Building the Excel spreadsheet

This spreadsheet, which has been made with OpenOffice or Microsoft Excel, helps you to analyse your data and choose the optimum training set. For this reason, a template spreadsheet has been built (see Table1) and it includes the equations to calculate the active and inactive compounds of the training set :

- Fill the template Excel spreadsheet with your data according the intructions below the table
- Save the new spreadsheet
- You should check the spreadsheet information and compare it with the characteristics of an ideal training set:
    - check the number of compounds,
    - check the number of active and inactive compounds,
    - check the order of magnitude,
    - delete the redundant compounds (structural information), ...

According with the information of the table you have to decide if your data is adequate.

The remaining compounds, which have not used as a training set, will be kept for the test set (the evaluate the predictivity after the generation of a hypothesis).

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

8. Generating a hypothesis

In the first section of this tutorial there is a short explanation about the HypoGen™ and HypoRefine™ modules. The first module that should be executed is HypoGen™.

8.1. HypoGen™ module

8.1.1. Entering the training set molecules

You have to build a spreadsheet in ***Generate Hypothesis workbench*** that it contains all molecules of the training set (the compounds that you have selected which are in the definitive excel spreadsheet that you have analyzed in the last step) and their correspoding activity values:

- Select the molecules of the training set and drag them to the ***Generate Hypothesis workbench*** **->** the molecules will be in the shelf of the workbench, you have to drag all of them from this shelf into the report area.
- If you check the list of compounds and you want to remove one compound of the spreadsheet -> select the row number and **Edit -> Clear Select Report Rows**

8.1.2. Adding the activity data

Before adding the activity data, you should know the data that correspond to each column in the report area of a spreadsheet:

- **Name**:it contains the name of the compounds and they are entered automatically by Catalyst™ when compounds are dragged into the spreadsheet.
- **Activ**: the input value is IC50 (the inhibitor concentration that reduces the enzyme activity to 50%) or EC50 (the concentration of an agonist, which produces 50% of the maximum possible response for that agonist) among others. To introduce decimal numbers you have to use a dot.
- **Uncer**: it represents the ratio range of uncertainty in the activity value (3.0 by default).
- **Color**: display colour of molecule
- **Estimate**: it contains the output values, which represent the estimated activity of each compound based on the hypothesis. This column is reserved for Catalyst™ to write output data.
- **Error**: is contains the output value whcih represent the ratio of the actual activity to the activity estimated by the hypothesis. This column is reserved for Catalyst™ to write output data.
- **MolWt**: it contains the calculated molecular weight by Catalyst™

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

- **Principal**: this column is used in Hypothesis generation with the ***Common Features Only*** or ***With Excluded Volume (HypoRefine)***
- **MaxOmit Feat (Common Features Only)**

To introduce the activity data and the uncertainty associated to:

- In the **Activ** column -> click in the **Activ** of one compound and then in the **Edit** entry box above the report, enter the value -> press **ENTER** to register the value
- In the **Unc** column -> click in the Unc of one compound and then in the Edit entry box above the report, enter the value 3 that is the value by default.
- To save spreadsheet -> select **Data -> Save Report To Lab As Spreadsheet**
- If you want to generate the hypothesis, you do not have to close this spreadsheet yet. In this case, you have the option to iconify it while you carry out the next step.

8.1.3. Checking the functions mapping

Catalyst™ will generate hypotheses using the chemical functions more important in your molecules. However, you have to indicate previously which functions there are in your molecules. Therefore, before executing Hypogen™, you should analyse several active molecules to know the chemical functions that they contain. You can do it in a ***View Hypothesis workbench*** using the chemical functions that Catalyst™ have already defined that are: *HB acceptor, HB acceptor lipid, HB donor, Hydrophobic, Hydrophobic aliphatic, Hydrophobic aromatic, Ring aromatic, Pos Ionizable, Neg charge, Neg Ionizable and Pos Charge*.

To know which functions contain your compounds:

- Open a ***View Hypothesis workbench*** -> drag one molecule into the workbench report area
- In the ***Feature Dictionary*** panel that will appear: select **Functions Only** -> Select a function (*e.g.* HB acceptor) -> select **Tools** (in the ***View Hypothesis workbench***) -> **Show Function Mapping**
- The status area in the lower left corner of the workbench identifies which function is being displayed and the total number of this function found in the molecule.
- A control panel appear on the lower right corner of workbench with which you can step through all the positions of the one function in the molecule. If the control panel no appear it means that the function that has been chosen is not in the molecule.
- You should analyse which functions are important functions for your set of compounds.

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

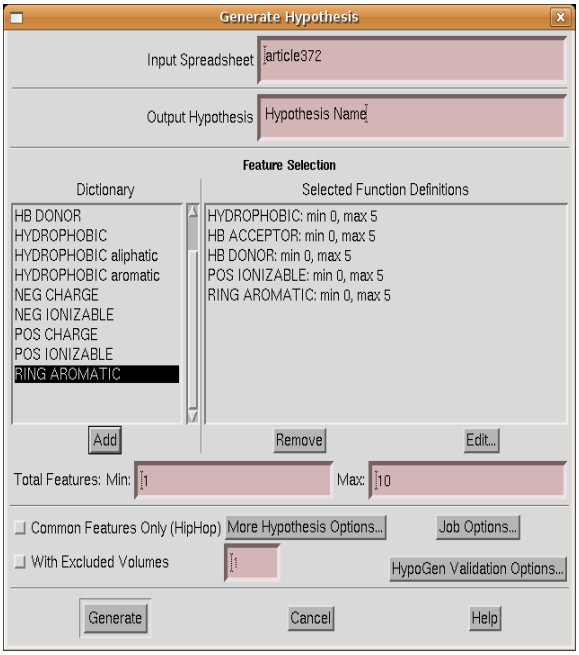
(for additional information see Catalyst4.10 Tutorials Chapter 7, on page 5, paragraph 4-5)

- With the same molecule you should check all functions of the ***Feature Dictionary*** panel and repeat this process with a representative set of your compounds to know the most important features.
- To analyze another compound you can open another ***View Hypothesis workbench*** or use the same cleaning the screen: select **Edit** -> **Clear Display**
- At the end, when you know the chemical functions that your active and inactive molecules contain, you can close the ***View Hypothesis workbench***. Since you don't have to save anything, you can close the workbenche clicking in the x in the top right corner and accept **Preform dipose.**

Usually, at the beginning you should choose the basic functions: ***HB acceptor, the HB donor, the Hydrophobic and the Ring aromatic***. However, it depends on your molecules. On the other hand, the maximum number of functions that HypoGen or HypoRefine will allow is 5.

8.1.4. Setting up to run hypothesis generation

- If you have closed the ***Generate Hypothesis workbench*** that contained the training set with the activity values (which you had generated in s*tep 8.1.1*) you have to open it again.
- Select **Tools -> Generate Hypothesis**
- Then, the ***Generate Hypothesis*** control panel will open:



UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

- Check the name in the **Input Spreadsheet.**
- Write the hypothesis' name in **Output Hypothesis.**
- In the **Dictionary** box there are all possible features. You have to select the functions that you consider more importants (according to the decision that you have done in the previous step). So, you have to select one by one the functions from the **Dictionary** box **->** and click in **add** , to add the functions in the **Selected Function Definitions** box.
- If you are wrong adding the functions, you can delete it selecting the function from **Selected Function Definitions** box -> **Remove**
- Select a function from **Selected Function Definitions ->** click in **Edit** -> in the ***Feature Editor*** panel you can change the minimum and maximum number of instance that you want for each function. You can force HypoGen or HypoRefine to search for one hpothesis with specific requirements.

To run HypoGen™ for the first time you should use parameters by default, so:

- **Total Features:** the default values are Min 1 and Max 10. It means that Catalyst™ is forced to search for one-feature hypothesis.
- **Common Features Only (HipHop):** this button responds to the additional options that are shown in the ***More Hypothesis Options*** control panel. It is not selected by default.
- **More Hypothesis Options:**
  - **VariableWeight:** This parameter is used to select the variable weight mode of HypoGen™. The default is 0 or standard mode. If you set this value to 1 you will use the variable weight mode. In this mode, HypoGen™ will allow the individual feature weights to vary during the optimization.
  - **VariableTolerance:** This parameter is used to select the variable tolerance mode of HypoGen™. The default is 0 or standard mode. If you set this value to 1 you will use the variable tolerance mode. In this mode, HypoGen™ will allow the individual feature tolerance to vary during the optimization.
  - **Spacing:** This parameter lets you specify the minimum distance between actual feature locations in molecules in the training set used for identifying candidate hypotheses.
  - **Weight Variation:** This parameter controls how large a range of function weights the hypothesis generator will explore during the hypothesis generation.

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

- **MinPoint:** This parameter controls the minimum number of location constraints required for any hypothesis.

– **With Excluded Volumes (HypoRefine™):** This option with a HypoGen™ run don't have to be selected. This option is selected to perform a HypoRefine™ run.

– **Job options:** Control panel also allows you to select a specific start time, set job priority, set up your job in the queue, and select a directory.
(For more information about these parameters, see Catalyst4.11 Tutorials, Chapter 7)

– **Validation Options:** This option have to be selected after generating the pharmacophore candidates in order to validate them.

– After checking the parameters, you can start hypothesis generation process clicking -> **Generate**

– Then, a message appears: "*Setup for Batch Generate Completed. Batch Process Scheduled or Running, based on specified Start Time. Acknowledged*". You have to click on the **Acknowledge** in order to accept the starting of the run.

8.2. HypoGen™ results

To analyse the **HypoGen™** results, you can generate a spreadsheet to write different parameters that will help you to choose the best hypothesis. The information that you should fill in is the following:

column1	*column2*	*column3*	*column4*	*column5*	*column6*	*column7*	*column8*	*column9*
Receptor name	**Catalyst TM spreadsheet**	**Hypothesis code**	**Chemical Functions**	**Specific parameters**	**Number of compounds**	**Number of Active compounds**	**Inumber of Inactive compounds**	**Uncertainty**

column10	*column11*	*column12*	*column13*	*column14*	*column15*	*column16*	*column17*	*column18*
Entropy (s)	**correlation (r)**	**RMS**	**Total Cost**	**Fixed cost**	**Null Cost**	**Difference: null - total**	**Difference: null - fixed**	**Analysis**

The next step will explain the HypoGen™ outputs and the meaning of the parameters to analyse the hypotheses.

8.2.1 Output

HypoGen™ creates a directory on the disk where the results of the process are stored (this directory is inside the operating directory where you started Catalyst™, in step 3.1).

In the output directory you can found the following files:

- **catHypo.cmd** (UNIX command executed by the program to run catHypo),
- **.chm** (hypothesis files, CatHypo returns the 10 "best" hypotheses),
- **.cpd** (compounds multiple-conformer files),

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

- **.debug** (equivalent to .full in CatHypo),
- **dict.ch** (dictionary with the functions defined in the Generate Hypothesis panel),
- **feat.lst** (list of the functions used in the hypothesis generation process),
- **.full** (full log file) **This file is very important to check at the beginning to know if the run is viable/feasible. You have to open this file and look for the entropy value. This value is in the following paragraph called "*Entropy of hypothesis space*":**

Summary of feature definition hit statistics:

HBA hits/lead: mean= 8.43 stddev= 3.55
HBD hits/lead: mean= 2.36 stddev= 2.61
HYDROPHOBIC hits/lead: mean= 3.00 stddev= 0.69
PosIonizable hits/lead: mean= 0.00 stddev= 0.00
RING AROMATIC hits/lead: mean= 5.90 stddev= 0.44
Entropy of hypothesis space: **17.3159**

VERY IMPORTANT NOTE 4

Mrs Katalin Nadassy (personal communication), defines the ***Config*** value as the ***Entropy*** of hypothesis Space. And that this value is calculated early in the run and is in the .full file near the value for Fixed cost. The number is the exponent to the base 2 of the number of models Catalyst™ will attempt to optimize during the run.

The most important is that:

- **If this number < 18** : a thorough analysis of all models will be carried out (it means that the parameters have been well definited).
- **If this value > 18** for the data set: it means that not all of the data will be considered when optimizing the hypotheses. In theses cases it is best to **re-evaluate the training set used for the hypothesis generation**.

  You should do the following tasks:

  - check the training set, remove some of the structures
  - check to make sure that the conformations for the training set compounds are right and they sufficient sample the conformation space of the compounds
  - check the active and inactive compounds and modify the number of inactive compounds (*i.e.* reduce the GenerateHypo.inactive.spread parameter from the default 3.5 to something lower, maybe 2.5 or 3.0 in order to include more inactive compounds in the run
  - try to reduce the configuration value by imposing restrictions in terms of the number and/or types of features that are included in the hypothesis. It is an effective strategy if you know or deduce the kind of functions included in the molecules and the number of them. Then, you can try to improve the

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

configuration cost modifying the minimum number of one, two o all functions from 0 to 1 and re-run HypoGen™. After that, you can try to modify the **Total Features option** (where the default values are Min 1 and Max 10). It means that Catalyst™ is forced to search for more feature in the hypothesis.

Both parameters should be changed progressively so starting with 1. At the end of each HypoGen™ run you should analyse if the changes are improving the configuration cost.

- **getresults.jrnl** (a journal file used by Catalyst™ during import),
- **.hypos** (ASCII file describing the hypotheses, coordinates),
- **leads.spst** (spreadsheet of molecules and activities/uncertainties. Each molecule is listed in this spreadsheet),
- **.log** (Log file, results are printed to this file),
- **.stderrout** (journal file checking that the job runs smoothly, it can be used for debugging)

– You can follow the status of hypothesis generation: select **Data -> Process Information**

– when the process is DONE **->** select the name of the process (hypothesis' name)-> click **Collect Process Data** to bring the hypotheses that have been generated.

– The hypothesis generation process returned 10 hypotheses, which contain some of the features that you have chosen. The hypotheses are scored according to their cost analysis and stored in the stockroom of the laboratory where you have collected the results. The ten hypotheses will appear in it represented with icons that are bulbs. So, each bulb is a hypothesis.

– You can analyse the pharmacophore doing a double click on each hypothesis icon. A ***view hypothesis workbench*** will be opened showing the distribution of the functions that define the pharmacophore.

The next step explain you how to interpret the results, the meaning of the different parameters used to accept or reject one hypothesis generation run or one specific hypothesis.

VERY IMPORTANT NOTE 5

In some situations can occur that only 9 hypotheses are obtained. It happens when the NULL hypothesis is exported as one of the top ten. So, it means that there aren't many hypothesis above that cost to choose from. In the .log file, if the first hypothesis reported is the null hypothesis this indicates that the null hypothesis had the lowest total cost of all the hypotheses returned which indicates that the hypotheses do not predict activities well. If the cost of the

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

NULL hypothesis is in the top 10 then it is exported. Thus, the missing hypothesis is the NULL hypothesis. This could mean that you need to change your starting conditions if it is possible.

When HypoGen™ returns only 9 the best option must be to re-run HypoGen™ with changing the initial conditions and re-run HypoGen™:

a) try to change the starting compounds of the training set (check the number of active and inactive compounds, the order of magnitude in the activity data, etc.)

b) change the starting features according the characteristics of your molecules

8.2.2. Results analysis

During an automated hypothesis generation run, Catalyst™ considers and discards many thousand of models. HypoGen™ distinguishes between alternatives hypothesis by applying the cost analysis. The overall assumption is based on Occam razor; that is, that between otherwise equivalent alternatives, the simplest model is the best.

Catalyst™ uses bits for language, so the program assigns costs to hypotheses in terms of the number of bits required to describe them fully. The program calculates the cost parameters summing three factors (weight cost, an error cost and a configuration cost).

Factors	***Definition***
weight cost	a value that increase in a Gaussian form as the feature weight in a model deviates from an idealized value of 2.0
error cost	a value that increase as the rms difference between estimated and measured activities for the training set. The standard deviation of this parameter is given by the uncertainty parameter. **It has the greatest effect in establishing hypothesis cost among these three terms.**
configuration cost	a fixed cost that depends on the complexity of the hypothesis space being optimized. It is equal to the entropy of the hypothesis space. This parameter is constant among all the hypotheses.

8.2.2.1. Cost Parameters

In order to evaluate a HypoGen™ or HypoRefine™ run, you have to analize three cost parameters:

1) The total cost of each returned hypothesis (in one HypoGen™run, ten total costs)
2) The total cost of the fixed hypothesis (one fixed cost for each HypoGen™run)
3) The cost of the null hypothesis (one null cost for each HypoGen™run)

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

In the ***.log file*** there is the values of theses cost parameters and others such as rms value and correlation (r). You have to fill the spreadsheet with these values and compare with the following table to know whether the hypothesis candidate presents a true correlation or not:

Δ cost parameters		**Evaluation**
Cost of the null hypothesis - Cost of any returned hypothesis	> 60 bits	There is an excellent chance that the model represents a true correlation
Cost of the null hypothesis - Cost of the fixed hypothesis	> 70 bits	It has a high chance of representing a true correaltion in the data.
	40-60 bits	It has a 75-90% chance of representing a true correaltion in the data
	< 40 bits	the likelihood of the hypothesis representing a true correlation in the data rapidly drops below 50%. It may be difficult to find a model that can be shown to be predictive.

- You have to analyze the parameters. They may be not the most adequate parameters. So, in order to improve the results you can try to apply these changes:
  1) In case that the difference between the Cost of the null hypothesis and the Cost of the first returned hypothesis is less than 40 bits you should change the training set, analyze the molecules and the parameters for the ideal training set. You can analyze which are the compounds that are wrong tested and change it for others to try to improve the hypothesis.
  2) you can use the variables weight and/or variable tolerances options. So, you can rerun the HypoGen™ changing these parameters in **More Hypothesis Options**:

     a) The ***VariableWeight*** *parameter* is used to select the variable weight mode of HypoGen™. The default is 0 or standard mode. If you set this value to 1 you will use the variable weight mode. In this mode, HypoGen™ will allow the individual feature weights to vary during the optimization.

     b)The ***VariableTolerance*** *parameter* is used to select the variable tolerance mode of HypoGen™. The default is 0 or standard mode. Set this value to 1 if you wish to use the variable tolerance mode. In this mode, HypoGen™ allows the individual feature tolerances to vary during optimization. Please note the when using variable tolerances compensates for deficiencies in the conformational coverage of unusually flexible compounds and tightens the tolerances when large tolerance values are not needed. The resulting pharmacophore models produce improved search queries.

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

However please also note that although it is possible to vary the weights and tolerances simultaneously, this adds a greater number of degrees of freedom (shown by the greatly increased configuration cost) and should be done with caution.

You could test to see if you only vary the weight and not the tolerances and then re-run the calculation and only vary the tolerances but not the weight to examine how these affect the results of the calculations. The new hypotheses obtained with changing these parameters will be only considered if they improve the resulting cost parameters.

3) Another parameter you could experiment with is to see if you could reduce the spacing from 300 picometer to something lower, maybe 200 picometer. This would allow more features to be included in the models. So, you can rerun the HypoGen™ run changing these parameters in **More Hypothesis Options.**

Although the cost parameters are important it is also important the regression parameters and the predictive power of your hypotheses (which will be evaluated with the remaining set of compounds).

8.2.2.2. Regression Parameters

The activity of the compounds of the training set is estimated using the regression parameters (you can found these values in ***.log file*** in the same directory than the ***.full file***):

Regression parameters	*Definition*
RMS	the rms factor represents the deviation of log (estimated activities) from the log (measured activities) normalized by the log (uncertainties). This parameter indicates the quality of "prediction" for the training set. This parameter indicates how well Catalyst™ was able to correlate actual with estimated activities for a given hypothesis. **It does not actually reveal anything about the Predictivity of a hypothesis**
Correlation	linear regression derived from the geometric fit index where 1.0 would correspond to a perfect correlation

You can do a quantitative estimation to know how each molecule in the training set fits to the hypothesis:

- You should open the ***Generate Hypothesis*** workbench that has been generated in previous steps (which contain the activity values of the compounds of the training set) double-clicking its icon.
- Drag one or all hypotheses generated, from their stockroom or laboratory to the shelf of the opened ***Generate Hypothesis*** workbench.
- Select simultaneously (a) one hypothesis workbench and (b) the icon of the spreadsheet, on the shelf of ***Generate Hypothesis*** workbench (use the computer key SHIFT)

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

- Select **Tools -> Score Hypothesis**

  Shortly, a 2D graph showing the regression line will appear and then, the estimate activiy values with its error are written in the corresponding column and row in ***Generate Hypothesis*** workbench.
- In this spreadsheet you can tidy the compounds according the estimated activity values: select **Lab -> Tidy Lab** (you have different potions: by Object name, by Object Type or Object Location.)
- To save the ***Generate Hypothesis*** workbench or spreadsheet with the estimated and error data: select **Data -> Save Report To Lab As Spreadsheet**.

8.3. HypoRefine™ module

HypoRefine™ module contains a new algorithm to resolve HypoGen's difficulties in generating hypotheses that correlate well when the steric properties of the data set play a large contribution to the activities. HypoRefine™ have to be applied after a HypoGen™ run in order to compare the results.

- To execute the HypoRefine™ module you have to enter again the training set molecules (the same training set that you used with the previous HypoGen™ run), add the activity data and add the uncertainty value associated to each compound (the same that you specified for HypoGen™ run) on a ***Generate Hypothesis*** workbench. So, you can use the same ***Generate Hypothesis*** workbench that you had built to perform the HypoGen™ run.
- Now, you have to chose the active and inactive molecules in order to define the location of the excluded volumes. You can chose active or inactive according the activities values of these compounds. So, in the ***Principal*** column you have to introduce the number **2** to treat one compound as Active, the number **1** when you want that the program treat one compound as Inactive and **O** or **empty** when the compound have to be ignored in excluded volume addition.

  Although this column for standard HypoGen™ is ignored, for HypoRefine™, these values can be used to determine which molecules are used when placing excluded volumes.
- Select **Tools -> Generate Hypothesis**
- The Generate Hypothesis control panel will appear, then:
  - Check that the name of the input spreadsheet is right in **Input Spreadsheet.**
  - Write a name for the output hypotheses in **Output Hypothesis.**
  - Select in the Dictionary box the same functions that you had chosen in the previous HypoGen™ run-> **add**

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

- If you are wrong adding the functions, you can delete it selecting the function from ***Selected Function Definitions*** box -> **Remove**
- Change the minimum and maximum values of each function if you changed it in the previous HypoGen™ run (you have to execut the HyporeRefine™in the same conditions than HypoGen™).

  The conditions (kind of function, minimum and maximum number of functions, etc. used for the HypoGen™ run have to be the same for HypoRefine™ run if you want to compare the results of both procedures later.
- **With Excluded Volumes (HypoRefine)** option has to be selected to perform a HypoRefine™ run and you can specify the number of excluded volumes to include during the optimization phase. A low value for this number is preferable. If this value is too high, then you are unnecessarily expanding the "search space" the algorithm has to cover, thus increasing the number will not necessarily improve results. You should start with 1, the value by default.
- After to check the parameters you can start hypothesis generation process clicking -> **Generate**

  (for more information see Catalyst4.11 Tutorials Chapter 17)

  A message appears: "*Setup for Batch Generate Completed. Batch Process Scheduled or Running, based on specified Start Time. Acknowledged*". You have to click on the **Acknowledge** to accept the starting of the run.

  The HypRefine™ run will generate the same files and parameters than HypoGen™. You can analyse them following the same steps that you followed for standard HypoGen™. The only difference can be the addition of one or more excluded volumes in the hypotheses.

9. HypoGen™ or HypoRefine™ validation

9.1 Test set validation

For one hand the pharmacophore model have to show a true correlation (cost parameters). However, the **predictivity** of a hypothesis must be determined by estimating the activities of compounds outside of the training set (*i.e.* the test set).

There are two ways to estimate the activity for the test set:

Option (a)

- Open a new ***Generate Hypothesis*** workbench or spreadsheet (double-click in **Generate Hypothesis** button in the Stockroom)

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

- Select the compounds of the test set -> Drag them to the shelf of the new spreadsheet.
- Drag one or all hypotheses from your laboratory o stockroom to the shelf of the new spreadsheet.
- Select one hypothesis and the icon of the spreadsheet simultaneosly.
- **Tools -> Score Hypothesis**
- You can tidy the compounds according the estimated activity values. **Select Lab -> Tidy Lab** (by Object name, by Object Type or Object Location.)
- To save spreadsheet -> select **Data -> Save Report To Lab As Spreadsheet**

Option (b)

- Select and drag one or all hypotheses to the ***View Hypothesis workbench*** button.
- When the workbench is opened -> drag the compounds of test set to the workbench shelf
- **Edit -> Clear Display**
- Select one compound and one hypothesis : **Tools -> Estimate Activity**

  The compound should appear in the workspace with the hypothesis functions. The estimated activity value appears in the ***Estimate*** text box of the ***Compare/Fit*** control panel. Above this box, you will see the calculated Fit number, and below it the relative energy of the conformer used in the fit.

  Note: Remember that the '**Fast Fit** ' is used to calculate the geometric fit during the optimization phase of HypoGen™ run. In comparing a compound and a hypothesis, the quality of the mapping is indicated by the fit value. A higher fit value represents a better fit, a perfect mapping of features would result in a value of fit equivalent to the sum of the weights of the features in the hypothesis.
- You can set the options under the Compare/Fit panel:

  (a) 'Find Best' option to improve the fit or (b) if you are using as a input file the crystal structure/single conformation the Best fit using the 'This Conf' option.

Catalyst™ computes, for each molecule in the training set or test set the difference between its experimental and estimated activity and call it *error value*. The *error value* is computed as the ratio of the experimental IC50 relative to the estimated IC50 (or the inverse if the later one is greater than the former one). If the all compounds of the test set have an Error values less than 10, it means that the activity prediction of these compounds falls between 10-fold greater and 1/10 of the actual activity. An error with a negative sign indicates that the experimental IC50 is higher than the corresponding estimation.

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

Note:

Mrs Katalin Nadassy (personal communication), recommend to examine the models returned to see if the pharmacophore:

- is able to distinguish between stereo isomers
- are the training set's order of activities accurately predicted
- is able to predict correctly the affinity of test molecules
- activity range of training and test compounds must be estimated to within 1 log unit, but with chemical relevance
- alignment of ligands onto model must be chemical meaningful
- are important features identified
- coverage of conformational space important
- are the selected features reasonable
- do the most active molecules fit in a reasonable manner

9.2. catScramble™ utility

Catalyst™ validates statistically a pharmacophore by performing a randomization with a utility called catScramble™ that is based on Fisher's randomization test. In order to perform it, catScramble™ does a random reassign of the activity values among the training set compounds. Each random reassignment generates a new spreadsheet where the active molecules can become inactive or have an intermediate level of activity and the originally inactive molecules can also become active. The number of these new spreadsheets will depend on the level of statistical significance that is wanted to achieve. To achieve a 95% of confidence level, 19 random spreadsheets (or 19 HypoGen™ runs) have to be generated [whereas for a 98 or 99% confidence level, 49 or 99 random spreadsheets (or HypoGen™ runs) have to be generated, respectively]

Then, a HypoGen™ process is performed with each randomized spreadsheet with identical experimental conditions (*i.e.* features, parameters) that were used in the original HypoGen™ run. After the new hypotheses generation, the statistical significance of the original HypoGen™ run was calculated with the expression: $significance = (1-(1+n)/N)*100$, where: (a) **n** is the total number of new hypotheses having a total cost lower than the hypotheses obtained by the original HypoGen™ run; and (b) **N** is the total number of HypoGen™ runs (initial + random runs). Therefore, if the randomized data set results in the generation of a large number of pharmacophores with similar or better cost values, RMS and correlation, then the hypotheses generated by the original HypoGen™ run are considered to be generated by chance and, therefore, they are no reliable (*i.e.* they have a low significance).

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

Therefore, it checks whether there is a strong correlation between the chemical structures and the biological activity or not.

- To perform the randomization test, the procedure is the same that when you perform a HypoGen™ run or HypoRefine™ run.
- So, to validate the HypoGen™run you have to repeated the steps you had followed previously: (1) Open the ***Generate Hypothesis workbench*** or spreadsheet; (2) select **Tools -> Generate Hypothesis** and chose the same conditions (restrictions, functions, etc.) used for HypoGen™.
- Now, select **Validation Options**: (a) you have to chose the Confidence level you want achieve to validate the paharmacophore (90%, 95% or 99%); and (b) you should not chose the **Clean upon finish** option because delete the spreadsheets generated during the randomization process. If you want analyse the information about these spreadsheet it is better don't delete it so don't chose this option.
- After to check the parameters you can start hypothesis generation process clicking -> **Generate**
- When the run is finished you can validate the HyporeRefine™ run repeting the same procedure that you made for validate HypoGen™. Remember that you have to choose the same parameters in order to run the Randomization Test in the same conditions than the HypoRefine™.

10. Activity prediction

The use of a hypothesis to estimate the activity of other compounds with similar receptor binding behaviour is a powerful concept. For estimate the activity of a group of compounds that can interact with the target you have to generate conformational models for each molecule. There are two ways to obtain the conformers : (a) with Catalyst ™, or using docking conformations.

10.1 Generating conformations with Catalyst™

You have to generate a conformational model for each molecule that consists of a representative set of conformers taken from the rang of energetically reasonable conformations of the molecule. Catalyst™ can generate this conformational model that represents the flexibility of a molecule following the same way that you had followed to generate the conformational models for the training o test set:

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

- Import the compounds: select **Data -> Import** ... -> select the compounds' names -> **Import**
- Select the compounds: (a) double-click in one compound or if you have more than one compound (b) select all compounds (use shift key) or (b) select all using the menu **Data -> Select all**
- Drag and drop them on the ***View Compound workbench*** button in the stockroom, if it has already opened drag and drop the compounds on the shelf of this workbench.
- Select **Tools -> Generate Conformational Model ->** write the Maximum Number of conformers (255); select Best Quality; write the Energy Range minimum in Kcal/mol (15 or 20)**-> Generate**

10.2 Using docking conformations

eHits® (electronic High Throughput Screening) software package provides energy optimised 3D coordinates of docked poses and conformations of ligand molecules in the active site of the receptor. The output of eHits is a sdf file (a multiple ligand file) containing all the active confomations of the ligand.

This sdf file is read by Catalyst™ as a multiple ligand file. So, if you want to use these conformations to test what is the most active conformation of the ligand and its activity value you have to import this sdf file into Catalyst™ :

- Select **Data -> Import** ... -> select the name.sdf file -> Import -> choose MDL Structure - data (SD) -> **Aknowledge**

  This sdf file will be visualized as one compound (one icon) with its multiple conformations, so you can tested the activity.

However, if you want to visualize the conformations as different compounds in a spreadsheet to do easier the estimation of the activity for each conformation and compare them, you have to translate this sdf to mol2 file (another multiple ligand file). It is possible with SYBYL 7.1 package (Tripos Inc., St. Louis, USA):

- Open Sybyl basic (first of all, you have to reserve Sybyl basic package at **CESCA - Reserves SCF** on the webpage: http://reserves-scf.cesca.es:8085/ as you had done to execute Catalyst™
- Write the following command line in a terminal:

```
% trigo -shell sybyl 7.1
% sybyl7.1
```

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

- Select **File -> Convert MACCS File yo Spreadsheet ...** -> **OK** -> write the name of the compound **->** answer **NO** (Include Property Data?) **->** select **Edit** -> **Save as mol2 format**
- Import the compounds to Catalyst™ selecting mol2 format: select **Data -> Import** ... -> select the compounds' names -> **Import**

10.3 Predicting IC50 or EC50 values

After the generation of conformations, the process is the same that you used to calculated the activity values for the Test set. So, you have two options to calculate the estimated activity:

Option (a)

- Open a new spreadsheet, double-click in ***Generate Hypothesis*** button in the stockroom
- Select the compounds -> drag them to the shelf of a new spreadsheet.
- Drag one or all hypotheses to the shelf of the new spreadsheet.
- Select one hypothesis and the icon of spreadsheet.
- **Tools -> Score Hypothesis.** Catalyst™ won't calculate the 2D graph because in this case you don't know the real activity of these compounds.
- You can tidy the compounds according the estimated activity values: select **Tools -> Sort by Property ...**
- To save spreadsheet **->** select **Data -> Save Report To Lab As Spreadsheet**

Option (b)

- Select and drag the hypotheses to the ***View Hypothesis workbench*** button
- Drag the compounds to the workbench shelf
- **Edit -> Clear Display**
- Select one compound and one hypothesis : **Tools -> Estimate Activity**

  The compound should appear in the workspace superimposed on the hypothesis and a number for the estimated activity appears in the ***Estimate*** text box of the ***Compare/Fit*** control panel. Above this box, you will see the calculated Fit number, and below it the relative energy of the conformer used in the fit. Furthermore, the compound should appear in the workspace superimposed on the hypothesis
- you can set the options under the Compare/Fit panel:

  (a) 'Find Best' option to improve the fit or (b) if you are using as a input file the crystal structure/single conformation the Best fit using the 'This Conf' option.

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

11. Web Material

http://www.accelrys.com/doc/life/catalyst411/index.html

Katalin Nadassy (Support information) katalinn@accelrys.com

12. References

[1] T. Langer, R.D. Hoffmann, Pharmacophores and Pharmacophore Searches, in: W. WILEY-VCH Verlag GmbH & Co. KGaA, Germany (Ed.), vol. 32, 2006.

[2] D. Schuster, C. Laggner, T.M. Steindl, A. Palusczak, R.W. Hartmann, T. Langer, Pharmacophore modeling and in silico screening for new P450 19 (aromatase) inhibitors, Journal of Chemical Information and Modeling 46 (2006) 1301-1311.

[3] O.F. Guner, Pharmacophore perception, development, and use in drug 2000.

[4] Y. Kurogi, O.F. Guner, Pharmacophore modeling and three-dimensional database searching for drug design using catalyst, Current Medicinal Chemistry 8 (2001) 1035-1055.

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

Table 1.
This spreadsheet allow to know how HypoGen or Hyporefine identify the active or inactive compounds and which are used during the first phase (called constructive phase) or in the second phase (called subtractive phase). In order to know how many active compounds and inactive compounds there will be you can follow the instructions below and complete the table (the most active compound will labelled with **MA** and the remaining compounds with **A**):

column1	*column2*	*column3*	*column4*	*column5*	*column6*	*column7*	*column8*	*column9*
name of the compounds	**EC50 / IC50 (from the most active to the less active)**	**uncertainty**	**activity column2/Uncertainty (except for MA = column2*Uncertainty)**	**activity value of MA (column4) – activity column4 value**	**Identification of actives: column5>0.0**	**log(activity column2)**	**activity colum7– activity column7 of MA**	**inactive = column 8>3.5**
1		3,0000		**0.0000**	**ACTIVE**		**0.0000**	**ACTIVE**
2		3,0000						
3		3,0000						
4		3,0000						
5		3,0000						
6		3,0000						
7		3,0000						
8		3,0000						
9		3,0000						
10		3,0000						
11		3,0000						
12		3,0000						
13		3,0000						
14		3,0000						
15		3,0000						
16		3,0000						
17		3,0000						
18		3,0000						
19		3,0000						
20		3,0000						
21		3,0000			**INACTIVE**			**INACTIVE**

Column 1	Write the name of each compound.
Column2	Introduce the activity value and put in order this compounds from the most active (the lowest activity value) to the most inactive (the highest activity value). The activity values have to be IC50 value or EC50 values therefore the higher the IC50 or EC50 of a molecule is, the lower is its inhibitory power.
Column3	The uncertainty is used to calculate the activity range, based on the initial value. By default, the value suggested is 3 for each compound. This means that if you have a compound with an activity of '6', Catalyst will actually consider a range of activities for that compound corresponding to : 6*3=18 and 6/3=2
Column 4	The equation applied in this column for the MA (the most active compound) is : the activity value of MA (column2) * Uncertainty. For the remaining compounds the equation is: the corresponding activity value (column2)/Uncertainty.
Column 5	The equation applied in this column for all compounds is the same: activity value of MA (column4) – the corresponding activity value for each compound (column4). In fact, the first value of this column should be 0,000.
Column 6	Write the level **ACTIVE** when the value from the column5 is higher than 0.0. Write the level **INACTIVE** when the value from column5 is lower than 0.0. This level is useful to know which compounds are identified as active and used during the first phase of HypoGen (constructive phase).

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

Column 7	Perform the logarithm to each activity value from column2 [=LOG10()]
Column 8	Perform the difference between the corresponding activity value of each compound from the colum7 and the activity value of MA from the column7.
Column 9	Write the level **ACTIVE** when the value from the column8 is lower than 3,5. Write the level **INACTIVE** when the value from column8 is lower than 3.5. This level is useful to know which compounds are identified as inactive in the second phase of HypoGen (subtractive phase). In fact, the first one should be ACTIVE and the last one INACTIVE. **The value 3.5 is the default value used by HypoGen or HypoRefine during the run. It is very important to make sure that there are enough compounds in the training set identified as inactive. If not, it is also possible to change this value in order to include more inactive compounds by setting up a lower value for the *GenerateHypo.inactive.spread* parameter in the .Catalyst file generated by Catalyst™ [where this file is located both at the run and at the home directories and it is read each time a Catalyst™ job (either interactively or in background) is started].**

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

eHiTS® tutorial

Index

1. What is eHiTS®

1.1 A short introduction

eHiTS® (**e**lectronic **Hi**gh **T**hroughput **S**creening; http://www.simbiosys.ca/ehits/) is a software package that provides energy optimised 3D coordinates of docked poses and conformations of ligand molecules in the active site of the receptor. The binding energy of each pose is calculated and reported as a score. At the end of each run the ligand that produced the lowest binding energy (i.e. best score) is reported.

eHiTS® is being developed by SymBioSys Inc. in Toronto (Canada) and its main strengths are: (a) it is easy to use; (b) it performs very well (it is both quick and accurate); and (c) it has a lot of automated features that simplify the drug design workflow and provide innovative solutions to common docking problems (*e.g.* the protonation state of the ligand/receptor pair; an exhaustive search of the ligand poses; the speed-up of the calculations in VLS; automatic identification of probable binding sites, the capacity to tailor the scoring function to the characteristics of the receptor binding site, etc.).

On the other hand, eHiTS® has some limitations such as: (a) there is no way to take into account the receptor flexibility; (b) all ring systems in the ligands are considered as rigid and therefore, their conformations are not changed during docking (hence, for a complete conformational sampling it is necessary to use multiple ring conformers); and (c) no knowledge-based constraints can be imposed on the docking (*e.g.* a specific ligand atom cannot be forced to be in a specific location in the poses; certain interactions cannot be prevented from occurring; etc.).

1.2. About docking process

The exhaustive search of the ligand poses made by eHiTS® follows a five-step process (the so called *divide and conquer* approach): (1) the ligand that has to be docked is automatically divided into rigid (*i.e.* non rotable bonds) and connecting fragments; (2) each rigid fragment is independently docked into the receptor binding site to obtain the corresponding rigid fragment poses; (3) pose-sets are built with all possible combinations of rigid fragment poses (where each rigid fragment contributes with a single pose to each set) but only those sets from which the complete ligand structure can be rebuilt by adding the connecting fragments are kept for further use (*i.e.* the rest are discarded); (4) the rigid fragments of the remaining pose-sets are joined with the connecting fragments; and (5) the complete ligand poses are refined by a local energy minimization in the active site of the receptor that is driven by the self scoring function. This process has the advantage that rigid ligand fragments that by themselves have poor interaction scores with the receptor but, in contrast, are part of a pose set that scores very well are not discarded in the initial steps (because decisions on which poses are retained for further processing and optimization are made on the basis of an overall score of the full ligand, and not on partial structures). In fact, it is very important to keep fragment poses that do not get good scores, because even for high affinity ligands some fragments may be acting simply as linkers that have a minor contribution to the binding. Another advantage of the way in which eHiTS® searches for

ligand poses is that when two rigid fragment poses are connected, any dihedral angle in the connecting segment can be virtually analyzed (although the one with the lowest energy is the one that is finally selected). Therefore, the dihedral angle sampling of eHiTS® is, in practice, equivalent to a continuous sampling. This reflects what is found in the experimental complexes deposited in the PDB where: (a) many ligand fragments have no interactions with the protein, or even interactions that are clearly repulsive (obviously, in both cases, the energy loss due to these "bad" interactions must be compensated for by strong attractive interactions formed by other fragments in the same ligand); and (b) the coordinates of the ligands do not always correspond to their most stable conformations.

1.3. What is CheVi®

A graphic interface for eHiTS® has been recently released by SymBioSys Inc. [CheVi® (*i.e.* ***Che****mical* ***Vi****sualizer*); http://www.simbiosys.ca/chevi/index.html]. It makes it very easy to set up eHiTS® with full control over the run characteristics: (a) input selection; (b) output selection; (c) parameter selection; (d) database options; (e) parallel processing or distribution options; and (f) filtering options. Moreover, CheVi® can be used to make a straightforward analysis of the docking results from eHiTS® runs, with a particular focus on the intermolecular interactions between the ligand and the receptor. Finally, both eHiTS®and CheVi® are also easy to install.

2. The eHiTS®Input

The eHiTS® run needs as an input the molecular structure of one protein, the structure of one or more ligands and the coordinates of the clip (see the definition of the clip below).

2.1. Ligand and receptor files

The ligand and receptor file should be in one of supported 3D input format:

- MDL Molecular files (**mol**) - note: must be 3D coordinates;
- SD file format (**sd/sdf**) - note: must be 3D coordinates;
- Protein Data Bank files (**pdb**);
- Tripos Mol2 files (**mol2**) - note: must be 3D coordinates;
- Tagged Molecule Ascii (**tma**) - native eHiTS® format;
- Tagged Molecule Binary (**tmb**) - native eHiTS® format.

The input ligand should be in 3D coordinate system. You should build the ligands or do a 2D-> 3D conversion. You can use software such as ChemDraw. On the other hand, the ligand should be connected because eHiTS® cannot work with disconnected ligands therefore you have to use 3D formats which contain connected ligands (*e.g.* mol2, mol, ...) otherwise perform the connection. You can use DeepView to save the ligand coordinates with connections between atoms.

ISBN:978-84-691-1553-4 /DL. T.151-2008

The DeepView-Swiss-PdbViewer (http://www.expasy.org/spdbv/) is an application that provides a user friendly interface allowing to analyze several proteins at the same time:

- Select **File** menu -> **Open PDB file…** or **Open MOL file …** -> select the name of the molecule
- Select **Select** menu -> **All**
- Select **File** menu -> **Save** -> **Save Selected Residues … ->** write the name of the molecule

2.2. The Clip file

The clip file is a file that is used to reduce the input receptor to contain only parts relevant to its active binding pocket. This clip file defines a clipping box that is used to reduce the size of the receptor. To ensure that the correct active site is selected for docking, the user may to split the ligand and receptor prior to giving them to eHiTS®. Therefore, you have to use a receptor/protein (without ligand) and a clip parameter defining the active binding pocket or pocked to study of the receptor. The coordinates of the ligand, which has been crystallized with the protein, is a clip file usually used if the structure of the protein is crystallized with the corresponding ligand.

The receptor file must be clipped (it means that it don't have to contain the ligand coordinates), unless this is used together with the clip parameter command. Therefore, you can use the protein and ligand in different files or the protein with the ligand in the same file. In the second case, you have to use the option clip when you execute eHiTS® by the terminal. The script has been generated to easier the execution of eHiTS®. So, if you use the script, you have to use the first option, you have to prepare a protein without ligand and a ligand that will be the clip around which the docking will be performed.

3. The eHiTS® ligand output

The ligand output of eHiTS® can be a ".sdf" or a ".tma" file for each ligand tested, both format files are multiple ligand files. The "tma" format is a eHiTS® native file format called Tagged Molecular Ascii (TMA) file format, that it uses internally to represent molecules used during docking. Therefore, it means that these files contain all the conformations docked into the protein. In order to obtain both formats (sdf and tma files) you have to run eHiTS® twice specifying sdf or tma for the output file. (you can not generate both files in one eHiTS® run).

4. How to execute eHiTS®

4.1. Installation and License (eHiTS®package)

eHiTS® 5.1 is offered free for academic institutions. The installation of eHiTS® require download the .bin file and execute it in the terminal writing the following command:

```
$ ./eHiTS_version_platform.bin <INSTALL_DIRECTORY_PATH>
```

(where <INSTALL_DIRECTORY_PATH> is the path of the directory where you want to install the program).

eHiTS®package is not a graphic interface and you will have to execute it using the command line by the terminal.

5. Working with eHiTS®

5.1. Download protein structures

Firstly, you have to download the coordinates of the protein/receptor to study. These coordinates can be obtained (a) from the PDB database if the protein has been crystallized, (b) from a dataset of models like ModBase (models of proteins that are built using crystallized templates); or (c) models that you can build using different homology modeling programs.

(a) To download crystallized proteins from the PDB you have to write in the terminal the following command substituting in the following command line the XXXX by the PDB code of your protein:

```
$ wget http://www.pdb.org/pdb/file/XXXX.pdb.gz
```

and decompress the pdb file:

```
$ gunzip XXXX.pdb.gz
```

(b) In ModBase there are deposited three-dimensional protein models calculated by comparative modeling (http://modbase.compbio.ucsf.edu/modbase-cgi/index.cgi). You can find a model using the sequence, the PDB code or other characteristics of your protein.

(c) They are different software that permit you to generate a model using homology structures.

5.2. How to run eHiTS®

5.2.1. Script (automatically)

A eHiTS® run can be performed automatically following a script called “script_ehits”. This script is prepared to automatized the docking of a list of ligands with different receptors.

To execute this script you have to type in the terminal the following command line (write in the directory where there is the script file):

```
$ script_ehits
```

The script will ask you about names of the structures (ligands, proteins and clips), full paths, and file formats.

So the script will processes all input information provided by user asking for:

1st absolute path/ of eHiTS® executable
2nd the results file's format: sdf or tma (file's extension)
about ligand
3rd absolute path/ of the file which contain the ligands' coordinates
4th absolute path/ of the directory which contain the file with the ligands' names
5th the file name which contain the ligands' names
6th the ligand file format(sdf, mol2, pdb, ...)(file's extension)
about receptor
7th absolute path/ of the file which contain the receptors' coordinates
8th absolute path/ of the directory which contain the file with the receptors' names
9th the file name which contain the receptors' names
10th the receptor file format (pdb, mol2,...)(file's extension)
about clip
11th absolute path/ of the directory which contain the file with the clips' names
12th the file name which contain the clip names
13th absolute path/ of file which contain the clip's coordinates
14th the clip file format (sdf, mol2, pdb, ...,) (file's extension)
errors
15th absolute path/ of the directory which will contain a file with the execution errors
16th the file name which will contain the execution errors produced
prints
17th the file name which will contain the prints produced
18th absolute path/ of the directory which will contain the file with the prints and directories named with the receptor code which will contain sdf files. If you have chosen the option to generate sdf files previously.

- At the end of this questionnaire, while the process is running, you can not close the terminal. When the process will be finished it will warn you about it printing a screen message in the terminal.

5.2.2. Not automatically

To work without script_ehits you can use the command line arguments in the terminal. In the terminal you have to write ehits.sh followed for the specific options, for example:

```
$ ehits.sh -complex [path/pdb_complex_file] -ligand [path/ligand_name] -out
[result_file.sdf] -toprank 10 –clean
```

or

```
$ ehits.sh -ligand [path/ligand_name] -receptor [path/receptor_name]  -clip
[path/clip_molecule] -out [result_file.sdf] -toprank 10 –clean
```

In order to understand the command line arguments, you should read the description of them in next step.

5.2.3. Basic command line argument description

eHiTS® option	Description
-ligand [path/ligand_name]	It specifies the ligand that you want to screen against the a given receptor. The type of the input file is identified by the file's extension. A set of ligands can be specified but we detected some problems. When you have to use more than one ligand the best option is run eHits® with the script.
-receptor [path/receptor_ name]	The receptor file must be clipped, unless this is used together with the -clip parameter. Alternatively one can use the -complex option.
-complex [path/complex_ name]	The parameter specifies a PDB ligand-protein co-crystallized complex. The program will separate the receptor and ligand from the co-crystallized complex. The extracted receptor is clipped with the ligand and is then used for the docking. It is worth checking the generated "_ligand" file in the pre-processing directory to verify if the program has found the correct ligand in the complex. (To extract the ligand file using -complex option you have to follow command syntax (a), but you must have a separate ligand file or not specify the ligand)
-clip [path/clip_molecule]	This parameter specifies a file name of a molecular structure, which is used to reduce the input receptor to contain only parts relevant to its active binding pocket. The surrounding box of the given molecular structure is used as a clipping box to reduce the size of the receptor. You can pass the co-crystallized ligand file OR a file containing a set of receptor residue atoms around the active site. **Very important:** the molecule should be saved in the **SAME COORDINATE SYSTEM** as the receptor. (1)
-out [path/result_file.sdf]	It tells the script where to put the output results. The default is: `~/ehits_work/results/protein/ligand/ehits_best.sdf` Please note that the file must have a .sdf or .sd file extension. To obtain result_file.tma, which is used to compute the interactions with the receptor using CheVi®, you don't have to include the option -out. Furthermore, the .tma file format can't be convert to .sdf file format. Therefore, you have to run two docking process, one specifying .sdf output and another without specify output format because for default it generate .tma files.
-toprank N	When using this option you can save only the given "N" number of top ranking solutions into the output SDF file if the -rms option was also used, then the first solution in the best.sdf will be the one with the lowest rms deviation from the X-ray structure.
-clean	Clears the SQL database for the DockTable and removes any pre-processed data at the start of the calculation. This option is useful when rigid docking parameters have been changed and old results should be removed (it has to call like the old directory).
-margin N	It specifies the clip box margin value as "N" Angstroms (the default is 7). The clip box , as defined in eHiTS® is the area "clipped" out of the receptor, into which the ligands are docked.
-allowflat	It force eHiTS® to accept flat molecules, it is sometimes the case where databases may contain flat 3D molecules. eHiTS®is a 3D docking tool, and as such it will reject 2D representations of molecules during a normal docking screen.
-rms [path/Xray_ligand]	Calculate the RMS deviation of each generated pose from the atom coordinates in the given file. Note, that the program assumes that the atoms occur in the same order in the X-ray file as they do in the input ligand file (passed with the -ligand argument). If this is not the case, the reported RMS values will be incorrect! So if the input ligand has been altered, e.g. by running energy minimization or other modelling on it, then make sure that the order of the atoms is preserved in the file with respect to the X-ray file.

(1) You may wish to specify the binding site in one of the following ways:

- The protein is co-crystallized with the ligand. In this case the complex is split and the cavity around the ligand is used as the binding site (use the script with "-complex" parameter)
- The protein is already clipped around the binding site. In this case it is simply given as the receptor (use the script with the "-receptor" parameter)

- The protein is not clipped, but the user has a separate file (either ligand, or some residue atoms around the active site) to clip it with (in this case use the script with "-receptor" and "-clip" parameters) .

To select some residue atoms around the active site, you can use RasMol (http://www.openrasmol.org/). RasMol is a program for molecular graphics visualisation that will allow yoy to select residues around the ligand using the following command line:

```
RasMol> restrict within(10.0, name_ligand)
RasMol> save pdb name_protein.pdb
```

6. Results

6.1. The results file name structure

eHiTS® generates two kind of files as a output:

a) **ligand_name.tma**

Tagged Molecule Ascii ,TMA, is the native eHiTS format which permit to visualize the results with CheVi® and calculate the interaction between receptor and ligand.

b) ligand_file_name.sdf

File that contain multiple conformations in this case of the same molecule, each conformation shows ligand_name.sdf, pose number, coordinates and specific parameters (*e.g.* eHiTS-Score, Term-metal, Term-H_bond, Term-Lipophil, Term-pi_stack, Term-other, vdWaals, Term-solvent, Term-steric, Term-family, Term-depth, Term-strain, Term-lig_int, Term-entropy, Pose)

6.2. The structure of eHiTS® working directory

Under the **home** directory eHiTS®creates a directory called **ehits_work,** where all the output files will be saved. Inside the ehits_work directory there are other directories where there are the inputs, the outputs, etc. The scheme of the content of these directories is:

/ehits_work contains:

/preprocess

This directory stores the pre-processed results, which will be reused for later docking runs.

*/preprocess/**ligands***

This directory contains the ligand files, converted from the input file format into the tma file format. Before processing a new ligand file the script examines this directory. If the pre-processed results for the currently tested ligand is found there, they will be reused for the docking and the ligand pre-processing step will be skipped.

NOTE: this also means, that if you have a ligand file first docked with the eHiTS®, then you modified the ligand file and would like to re-dock this ligand, you have to remove the pre-processed ligand data from the *ehits_work/preprocess/ligands/...* directory.

ISBN:978-84-691-1553-4 /DL. T.151-2008

/preprocess***/receptors***

The directory stores the pre-processed results for receptors and co-crystallized complexes.

- separated ligand and receptor files (for co-crystallized complexes only);
- tma receptors files;
- Steric grid files (sga), describing the spatial features of the receptor;
- Feature graph file (fga), representing the extracted chemical features of the receptor.

/results

The "results" directory is the directory storing docking results.

/results*/r***eceptor_file_name***

Every processed receptor will have its own separate directory which contains a "scores.txt" file which reports the chemical parameter values, the total number of solutions, the scores for each solution and the best overall score from the docking. The score value represents the binding energy of the ligand to the receptor and is in units of logKd(Ki) (also known as: pKi).

*/results/***ligand_file_name**

Every ligand screened against the receptor has its own directory, storing all its docked poses. The resulting poses are saved in the, Tagged Molecule Ascii (TMA) and SD file formats (defined before).

/logs

7. Training option

It is one of the methods available to train the eHiTS® scoring function with known data. Validation training uses co-crystallized structures (PDB complexes) as well as an optional set of decoy or inactive ligands to train the scoring function to better predict binding modes for a family of receptors.

7.1. Text file: complexes and ligand names

A text file have to be build with the path of complexes and ligands conform to the naming convention which contain the PDB file corresponding:

(a) It can be with –complex option, then the text file will be as following:

-complex /home/myname/mypath/PDB_complex.pdb

-complex /home/myname/mypath/PDB_ complex.pdb

-complex /home/myname/mypath/PDB_ complex.pdb

-complex /home/myname/mypath/PDB_ complex.pdb

-complex /home/myname/mypath/PDB_ complex.pdb

or

(b) It can be –receptor and -ligand options then the text file will be:

ISBN:978-84-691-1553-4 /DL: T.151-2008

-receptor /home/myname/mypath/PDB_protein.pdb -ligand /home/myname/mypath/PDB_ligand.pdb
-receptor /home/myname/mypath/PDB_protein.pdb -ligand /home/myname/mypath/PDB_ligand.pdb
-receptor /home/myname/mypath/PDB_protein.pdb -ligand /home/myname/mypath/PDB_ligand.pdb
-receptor /home/myname/mypath/PDB_protein.pdb -ligand /home/myname/mypath/PDB_ligand.pdb
-receptor /home/myname/mypath/PDB_protein.pdb -ligand /home/myname/mypath/PDB_ligand.pdb

or

(c) when a job has been ran previously and processed data exist already.

-code PDB_code
-code PDB_code
-code PDB_code
-code PDB_code
-code PDB_code

7.2. How to execute training option

To run the training utility you have to type in the terminal the following command line:

```
$ train.sh [text_file_name]
```

The docking processes have been performed with values by default. However, you can use optional arguments which are not used by default such as **-PoseMatch, -DockOptim, -individual, -enrich**

8. Enrichment option

It takes a set of active ligands and inactive ligands and trains the scoring function on a single receptor, with the goal of ranking actives higher than inactives.

eHiTS® allows the user to specify the accuracy/speed trade-off by indicating the accuracy level (*i.e.* accuracy N, where N can be accuracy 1 to 6)

The accuracy option will not explain in detail in this tutorial because it has not been applied since the solutions of eHiTS® docking will be combined with the construction of a pharmacophore model. Therefore, in this tutorial has been considered speed over accuracy during the docking process. The accuracy is achieved fitting docking solutions with a specific pharmacophore to know which solution can explain the ligand activity.

9. How to visualize the results (CheVi®)

CheVi® (Chemical Visualizer) , a 3D molecular viewer primarily developed for the eHiTS® docking tool, gives the user ability to examine eHiTS® docking results, with corresponding eHiTS® scores. It allows to see the interactions between the ligand and receptor and save them in a text file.

ISBN:978-84-691-1553-4 /DL. T.151-2008

9.1. Installation and License

The installation of CheVi® require download the .bin file and execute it in the terminal writing the following command:

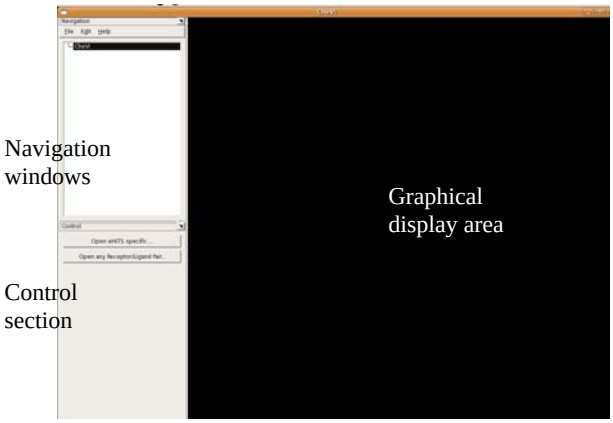
```
$ ./CheVi_version_platform.bin <INSTALL_DIRECTORY_PATH>
```

(where `<INSTALL_DIRECTORY_PATH>` is the path of the directory where you want to install the program).

9.2. Working with CheVi®

To execute CheVi® you have to type in the terminal the following command:

```
$ chevi
```



- Select **Edit** -> and **Open** to visualize the structure of your ligand or protein.
- You can display single and multiple molecules files if you are opening a multiple ligand file (*.sdf, *.mol2,*.tma)
- If you open a tma file, you will can display ligand-receptor interactions and save it in a text file. This text file can be use to analyze the more important kind of interactions and the atoms involved in them.

The tma file is usually saved in /home/ehits_work/results/receptor_name/ligand_name/ehits.tma.

When you open a filename of ehits_best.tma, ehits.tma or solutions.sdf, the entire docking job tree will be created fro display. A solution parent node, with receptor, ligand, and the eHiTS®solutions child nodes. Expanding each of these child nodes will display that particular molecule for you.

- You can display molecule surface

ISBN:978-84-691-1553-4 /DL. T.151-2008

10. References

[1] Z. Zsoldos, D. Reid, A. Simon, S.B. Sadjad, A.P. Johnson, eHiTS: a new fast, exhaustive flexible ligand docking system., J Mol Graph Model 26 (2007) 198-212.

[2] Z. Zsoldos, D. Reid, A. Simon, B.S. Sadjad, A.P. Johnson, eHiTS: an innovative approach to the docking and scoring function problems., Curr Protein Pept Sci 7 (2006) 421-435.

PHASE™ tutorial

Index

ISBN:978-84-691-1553-4 /DL: T.151-2008

1. What is Maestro

Maestro is the graphical user interface for all of Schrödinger's products (i.e. LigPrep™, Macromodel®, Phase™, etc.). It contains tools for building, displaying, and manipulating chemical structures and associated data; and for setting up, monitoring, and visualizing the results of calculations on these structures. Maestro allows top execute a Schrödinger's modules, for example a Phase™ process using a specific panel, so all Phase™ jobs can be started from Maestro. However, you can run Phase™ or another module from the command line. In this tutorial the Schrödinger's modules have been executed from Maestro. The command line procedure can be found in the Phase™ 2.5 User Manual.

2. Executing Maestro

2.1 Executing Maestro remotely on CESCA

To connect with CESCA in order to execute the program, you have to open a terminal and connect to the server where Maestro is installed (*i.e.* encantat.cesca.es) by using the ssh protocol with the -X option. In order to do that, type the following command:

```
$ ssh -X username@encantat.cesca.es
```

(where username needs to be a registered user for encantat.cesca.es)

where pujadas@encantat.cesca.es is the computer's IP where Phase™ has been installed in CESCA

2.2 Executing Maestro locally

Maestro can be installed and executed with either shell. You will need a license to use Phase™ or another module. For license or technical problems you should contact with Ingrid Bàrcena i Roig from CESCA (Centre de Supercomputació de Catalunya; e-mail: ingrid@cesca.es; Dept. Assistència Tècnica Tel: +34 93 205 6464, Fax: +34 93 205 6979, Edifici Nexus, Gran Capità, 2-4, 08034; Barcelona). For scientific doubts or questions about the algorithm, you should contact with help@schrodinger.com

3. Starting Maestro

3.1 Creating the operating directory

Firstly, you should create a working directory to run Maestro (except if you want to work with data obtained previously):

```
$ mkdir directory_name
```

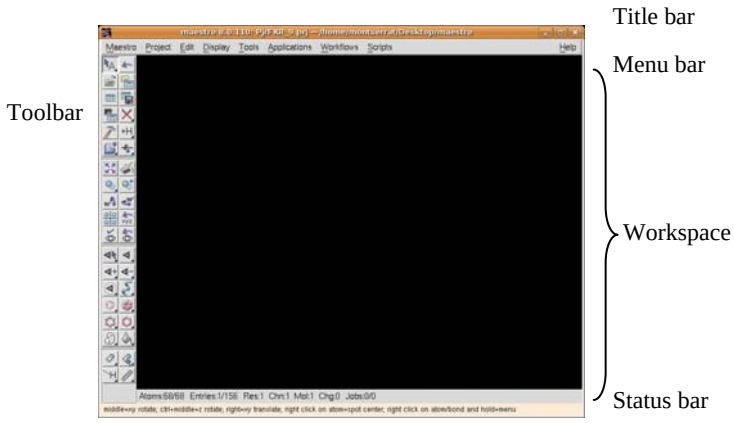
3.2 Starting Maestro

Starting Maestro by typing the next command in the terminal:

```
% maestro
```

ISBN:978-84-691-1553-4 /DL. T.151-2008

This is the main window of Maestro:



4. Working with Maestro

4.1 Starting a Project and saving the runs

In the Menu Bar, there is the Project Panel that allow to open a project. When you open a project it will report all that you will do. So, you will be able to recuperate a complete process. The project will save all actions that you perform, so it is not necessary save during the process.

- Create a new project: select ***Project*** in the Menu bar -> **new**
- Open the project in the ***Project*** in the Menu bar -> **Open**

The project that you have created is empty, so it will record all that you will do.

5. What is Phase™

Phase™ is a product of Schrödinger that is integrated into Maestro which contain tools to generate pharmacophores and 3D-QSAR models when activity data is known. Phase™ can generate structure alignments, activity prediction, and 3D database searching.

Phase™ searches the common pharmacophore hypotheses that explain the characteristics of 3D chemical structures which are propose to be critical for binding. Each hypothesis is accompanied by a set of aligned conformations that suggest the relative manner in which the molecules are likely to bind. A given hypothesis may be combined with known activity data to create a 3D-QSAR model that identifies overall aspects of molecular structure that govern activity.

6. How to execute Phase™

In this tutorial Phase™ will be executed from Maestro. Another way to execute Phase™ is using the command lines, in this tutorial it is not shown, for that if you want to execute Phase™ using command lines see the Phase™ 2.5 User Manual.

ISBN:978-84-691-1553-4 /DL: T.151-2008

6.1. Starting Phase™

Phase™ is divided in the following workflow:

1) Building a pharmacophore model and an optional QSAR model

2) Preparing a 3D database that includes pharmacophore information

3) Building or editing pharmacophore hypotheses

4)Searching the database for matches to a pharmacophore hypothesis

These workflows can be executed separately or using one panel. In this tutorial we will only use one panel that contains all modules required to built the pharmacophore and QSAR models:

- In the Menu bar select ***Applications*** -> **Phase** ->**Development Pharmacophore Model ...**
- The ***Development Pharmacophore Model*** panel will appear.

  At bottom of this panel five buttons labelled as ***Prepare Ligands***, ***Create Sites***, ***Find Common Pharmacophores***, ***Score Hypothesis*** and ***Build QSAR Model*** which are the steps that you have to follow to generate a 3D-QSAR model.

  Since one step depends on the information generated in the previous step, the next step are inactive while they could not be executed. So, each step depends on the before one, however, you can always move back. Moreover, at the end of the QSAR generation process you can click in the step that you desire.

It is important to mention that in one project you will perform two or more different calculations using for example the same compounds. So, in this cases, you can save each process as a separately run and recuperate it when you need it. However, when you are changing the conditions of a run the program ask you to save this modification as a new run.

- To save a run select **File** in toolbar of ***Develop Pharamacophore Model***: **Save as ...** -> write the name of the new run or simply change the number of the run -> **OK**

6.2 Ligand preparation

Phase™ requires 3D structures for the ligands that can be generated in Phase™ (it has not been considered in this tutorial) or they can be generated and further minimized with ChemDraw Ultra™ v10.0 (CambridgeSoft Corporation, Cambridge, MA, USA; http://www.cambridgesoft.com/software/details/?ds=2&dsv=9) (see ChemDraw Tutorial).

6.2.1 Import molecules

The formats of the 3D structures of the ligands that you can introduce in Maestro are: mol2, pdb, sd, among others. The mol2 format obtained from ChemDraw is not appropriated to use in Phase™ because the name of the compounds is the same for all compounds. Moreover, since Phase™ has problems to read the PDB format you should use the sd format.

ISBN:978-84-691-1553-4 /DL. T.151-2008

IMPORTANT NOTE 1:

Since the file format that contain all information about double bounds, etc., is the mol format, you should convert the .mol structures to sd. You can do it with a converter like Babel converted.

Babel converter allow you to convert your mol format to sd format.

- First, you have to install Babel converted.
- Starting Babel by typing the next command in the terminal:

  `$ babel –imol name.mol –osd name.sd`

  To convert a lot of compounds you should use the following perl script:

```
#!/usr/bin/perl

#Fem un ls del directori on estem i el guardem en un fitxer de text

system ("ls /home/montserrat/Desktop/MOL >/home/montserrat/Desktop/format/ls.txt");

#Obrim aquest fixter i el llegim linia a linia

print "prova\n";


open (IN, "</home/montserrat/Desktop/format/ls.txt")||die "Cannot open IN";

while (<IN>){

print "prova\n";

#treiem el salt de linia

        chomp;

#assignem el nom del fitxer a la variable fitxer

        $fitxer=$_;

#comprovem que el fitxer sigui un .mol

        if ($fitxer =~ /\.mol$/){

#dividim el nom del fitxer pel punt de manera que poguem accedir al nom sense
l'extensió

                @dades=split (/\./, $fitxer);

#emmagatzem aquest nom a la variable $nom

                $nom=$dades[0];

#li diem al sistema quina comanda ha de llençar, posant els noms dels fitxers gracies
a la variable $nom

                system ("babel -imol /home/montserrat/Desktop/MOL/$nom.mol -osd
/home/montserrat/Desktop/format/$nom.sd");

        }

}

close IN;
```

  You have to substitute the path and the name of the corresponding directories and type the next command in the terminal:

  `$ perl script_name.pl`

ISBN:978-84-691-1553-4 /DL: T.151-2008

- Once you have the compounds with the adequate format, you should introduce them in the ***Develop Pharmacophore Model*** panel: select **From File** (in the step ***Prepare Ligand***) -> select the format (sd) -> select **All Files** -> **Add**
- Add the activities (*i.e.*pIC50 or pEC50) values in the corresponding column. The activity must be a positive quantity that increases with increasing activity. Since the activity must be positive quantity if your activity values are expressed in nM you should use –log10(EC50 or IC50 value) + 9 formula to calculate pEC50 or pIC50.
- After that, click **Clean structures** in dialog box using the parameters by default: retain specified chiralities (vary other chiral centers); maxium number of stereoisomers: 32; retain original states; ionize at target pH: 7.0) -> **START** -> a green box dialog will appear showing the different options for execute the run -> click **START**.
- Save the run selecting **File** -> **save run** -> name_run

6.2.2 Conformational models generation

In the step Prepare Ligand you can generate the conformations for each compound. The module that generate the conformations is called MacroModel, which is already implemented in Phase™:

- Click **Generate conformers...** option -> and the parameters by default (current conformers: Discard; Sampling: Rapid; maxium numbers of conformers: 1000 per structure; Ligand torsion search; Vary amide bond conformation; Preprocess, minimization steps:100; Postprocess, minium steps: 50; OPLS_2005; Distance-dependent dielectric; maximum relative energy/difference: 10Kcal/mol; and Distance cutoff for redundant conformers: 2.00Å -> **START** -> in the green box dialog click **START**.

6.3 Creating sites

In this step, Phase™ identifies the chemical features (that are also called as pharmacophore features) in the ligand structures. The search is done using patterns of chemical structures and when one features is located in a specific place in one ligand conformation it is called pharmacophore site. All pharmacophore sites of each function analyzed in each ligand conformation is recorded.

It is important to be sure that the features are located correctly, the features are enough or new features that have to be added or redefined because specific functions in the structure of our compound set are important to consider the interaction with the receptor.

You have to choose the active and inactive ligands manually or using the threshold option:

ISBN:978-84-691-1553-4 /DL. T.151-2008

- Select **Activity Thresholds ....** option –> write a threshold for active ligands and a threshold for inactive ligands.

NOTE:

It might be a good idea to structure your activity thresholds more, ideally you want the actives to be strongest binders say in the nM range. For example if you have 5 or more compounds with actives greater than 8 then set the active threshold at 8 not 7. If there are fewer than 5 compounds with pEC50 > 8, I would use 7.5. An inactive threshold of 5 might be more appropriate but this is not used unless you are scoring with respect to inactive compounds and using the top n hypotheses to build QSAR models on. You could also require a larger proportion of the actives to match the pharmacophore to say at least 80%. This may not work with an active threshold of 7.0 because there could be a lot of compounds in that activity range that don't satisfy the same 5-point pharmacophore. But of the active threshold is raised it should be possible to increase the fraction of actives that match the pharmacophore.

6.4 Finding the common pharmacophore

The search of common pharmacophore is performed among the set of high-affinity (active) ligands that you have to choose in the previous step. The *partial match* is not considered in this process, so a k-point pharmacophore must be matched on all k-sites by a minimum required number of active. The common pharmacophores are selected using a tree-based partitioning technique (see in the NOTE below).

To perform this process you can modify the following parameters:

- You should choose the *number of sites* in the pharmacophore (**Number of sites**). Start with high number and decrease it if no common pharmacophores are found. By default, Phase™ looks for common 5-point pharmacophores, that is, pharmacophores containing 5 sites. The number of sites can be set to any value between 3 and 7 inclusive. If the number of sites is too large, you might not find any common pharmacophores, but if the number of sites is too small, the common pharmacophore might not contain all required features, and therefore might not discriminate between active and inactive very well
- You can select a lower value *limit on the number of ligands that must match a pharmacophore* before it can be considered to be a hypothesis (**Must match at least**):
  By default, Phase™ looks for pharmacophores that are common to all active ligands. However, Phase™ allows you to relax this criterion so that a common pharmacophore need only match a subset of the active ligands. This is required when the set of active compounds are highly diverse. So, in general a common pharamcophore must match a minimum required number of active.

ISBN:978-84-691-1553-4 /DL: T.151-2008

- Filter the *number of each kind of features in the pahramacophore* (**Feature frequencies**). You can specify the minimum and maximum limits for one feature when you want common pharmacophores to contain at least one but no more than specific number one feature.

  The feature combinations in the Variant list (variant is a combination of features that defined a pharmacophore) are determined entirely by the Minimum and Maximum limits in the Features frequencies table.
- You can click the dialog box Options, and choose the parameters by default (distances, box size, etc.).
- You have to ensure that all variants in **Variant list** are selected and click Find, and then click Start in the green box. During the common pharmacophore search, those pharmacophores that contain identical sets of features with very similar spatial arrangements are grouped together. Therefore, the common pharmacophore is the pharmacophore of the one group which contain one pharmacophore from each ligand
- When the process is finished different variants are incorporated in the *Results box*. You can *select the variants of filter list* to continue o select all list.

NOTE:

Common pharmacophores are perceived using a tree-based partitioning technique that groups together similar pharmacophores according to their intersite distances (the distances between pairs of sites in the pharmacophore). All pharmacophores of a given variant (AAAHR) are enumerated and partitioned into successively smaller high-dimensional boxes according to their intersite distances. Pharmacophores that are clustered into the same box are considered to be equivalent and therefore common to the ligands from which they arise. Boxes that contain pharmacophores from the minimum required number of ligands are said to *survive* the partitioning process. Each *surviving box* contains a set of common pharmacophores, one of which is ultimately single out as a hypothesis.

- When all or specific variants have been selected, you can continue with the next step, the scoring step where one pharmacophore from each box will be selected as a potential hypothesis.

6.5 Scoring the hypotheses

In this step you apply a scoring function that identifies the best candidate hypothesis from each *suvirving box*, and provides an overall ranking of all the hypotheses. So, common pharamacophores are examined and a scoring procedure is applied to identify the

ISBN:978-84-691-1553-4 /DL. T.151-2008

pharmacophore from each box that yields the best alignment of the active ligands. The scoring procedure provides a ranking of the different hypotheses, allowing you to make rational choice about which hypotheses are most appropriate for further investigation.

6.5.1 Scoring Active

The first task is to align the active compounds to the hypotheses and calculate the score for the active.

- Click **Score Actives… option** -> It opens the *Score Active dialog* box which contain the Alignment Scores (vector and site filtering) and Survival Score Weighting Factors (Survival score formula).

Terms	Definition	Default values
Vector and site filtering		
Keep those with RMSD below	Threshold for RMS deviation of the intersite distances of any contributing ligand from those of the reference ligand.	1.2 Å
Keep those with vector scores above *threshold*	Threshold for the variation in the alignment of vectors between any contributing ligand and the reference ligand. The maximum is 1.0, which corresponds to perfect alignment. The minimum is -1.0, which would keep all hypotheses, regardeless of vector alignment.	0.5
Keep the top	Limit on the percentage of hypotheses to keep, in order of combined alignmnt score.	10%
Keep at least and the most	Lower and upper limits for the number of hypotheses to keep	10-50
Use feature matching tolerances *	In addition to using the RMSD to filter out hypotheses, you can set matching tolerances on individual features. Features are considered to match if the site points are within the specified tolerance. This feature is useful if the RMSD matching is satisfied, but one or more features does not match well enough. The tolerances for each feature type are listed in the table below, and can be edited. All tolerances are applied: if you want to disable matching tolerances for a particular feature type, set the tolerance to a large value.	No check
Survival score formula		
Vector score	This score measures how well the vectors for acceptors, donors, and aromatic rings are aligned in the structures that contribute to this hypothesis, when the structures themselves are aligned to the pharmacophore.	1.0
Site score	This score measures how closely the site points are superimposed in an alignment to the pharmacophore of the structures that contribute to this hypothesis, based on the RMS deviation of the site points of a ligand from those of the reference ligand.	1.0
Volume score	It is the average of the individual volume scores. The individual volume score is the overlap of the volume of an aligned ligand with that of the reference ligand, divided by the total volume occupied by the two ligands.	1.0
Selectivity score	Estimate of the rarity of the hypothesis, based on the World Drug Index. The selectivity is the negative logarithm of the fraction of molecules in the Index that match the hypothesis. A selectivity of 2 means that 1 in 100 molecules match. High selectivity means that the hypothesis is more likely to be unique to the actives.	0.0
Number of matches	Number of actives that match the hypothesis.	1.0
Reference ligand relative conformational energy	Relative energy of the reference ligand in kcal/mol. This is the energy of the referent conformation relative to the lowest-energy conformation.	- 0.0
Reference ligand activity	Activity of the reference ligand.	0.0

* If you are using feature matching tolerances in the scoring step and this can eliminate all the potential matches for some of the less active compounds and can cause them to be eliminated from consideration when a QSAR model is built. In this case, these tolerances could be turned off

ISBN:978-84-691-1553-4 /DL: T.151-2008

NOTE:
This is the step with more time requirements.

6.5.2 Scoring Inactive

If we are assuming that a inactive compounds is a compound that is not able to accomplish the pharmacophore because it do not fit in all features. Then, we will use the Scoring inactive process to penalize the hypothesis in which the inactive compound is able to match on all features

- Click **Score Inactives…**-> **START** (the weight for the inactive score is 1.000 by default) -> **START** in green dialog box.
- At the end of the job, in the table of the hypotheses, in the Survival-inactive column will appear the adjusted scores.

6.5.3 Rescore

In this step you can calculate score taking into account different parameters. You should find the more adequate parameters for your data.

- You can finish at this point or select hypotheses for the generation of QSAR models and continue to the next step. On the other hand, you can select all or some hypotheses and proceed to find matches to the hypotheses. You should choose the option more appropriate according your goals.
- If you want to generated a 3D-QSAR model you have to select the desirable hypotheses according the scoring values and click in the next step.

(In this step you can also add to the hypothesis volumes that should not be occupied by atoms in any active molecule, known as *excluded volumes*. If you don't have this information you should do it at the end or using automated process to add excluded volumes)

6.6 Building QSAR model

In this step you should build a QSAR model with the hypotheses selected in the previous step, using the activity data for ligands.

- Select all hypotheses that you have chosen in the previous step -> click **Options** button -> PLS (the maximum number of PLS factors is N/5, where N is the number of ligands); and check the Atom-based option in the dialog box.
- Put 50% in random training set -> and **Apply**.
- **Build Models…** -> Start.
- If you want to view the statistics results click QSAR results.

ISBN:978-84-691-1553-4 /DL. T.151-2008

In the Build QSAR Model step, you will build QSAR models for the hypotheses that have been selected in the Score Hypotheses step and using the activity data for all the available ligands. You can choose atom-based or pharmacophore-based models and select different training sets and test sets. You can also modified the grid spacing and other parameters. So, you should modify the parameters in order to find the better options for your compounds. Moreover, you can visualize the resulting models. So, you can use them to (a) visualize parts of the ligands (atoms or pharmacophores) that contribute positively or negatively to activity, and to (b) predict activities of matches to the hypotheses from a database.

- First, ensure that the View model toolbar button is selected (the third button below the File, Display and Step options).
- Select one ligand, for example the most active ligand. It will be visulaized in the Workspace.
- Click **QSAR visualization** -> and the ***QSAR Visualization Settings panel is displayed***. This panel has different options for displaying characteristics of the QSAR model.
- Select the **Workspace ligands** option -> you can also view the cubes associated with the ligand ; or select **QSAR model** -> you can view the union of the cubes occupied by all training set ligands)
- Move the positive and negative coefficient threshold sliders to an intermediate value.

Then, in the Workspace, you will see many blue cubes and a smaller number of red cubes. The blue cubes indicate regions that are favourable for activity and the red cubes indicate regions that are unfavourable.

When you have completed this step, you can export the hypotheses used to build the model to an external file for use with other projects

7. Creating a 3D Database

Once a pharmacophore model has been developed, you can use it to search a database, with the goal of identifying additional active molecules. Thus, a Phase™ database is a set of files that resides in a directory. The following steps explain how you can create a directory by yourself.

- Open a new terminal and type the following command:
  ```
  $ mkdir name_database
  ```
- Close Maestro if it has been opened and start Maestro in the directory that you have recently created it:
  ```
  $ cd name_database
  ```
- In the ***Manage 3D Database panel*** -> click **Add**
- Use the **browse** option to find the structures of your molecules.

ISBN:978-84-691-1553-4 /DL: T.151-2008

- When the ***Clean dialog*** box is opened you can select the option **Structures have already been cleaned (confirm and proceed)** if you are certain that you have good structures. Then, the job that is run simply performs some checks on the structures.
- If the database has been created from chemical files that contain fairly crude structures, it is important to clean the structures. Then, you should not select the option: **Structures have already been cleaned (confirm and proceed)**
- Select the options according with your structures and click on **Start... ->** in the green box click **START**.
- Since this database will be searched using hypothesis developed from a set of ligands that bind to distinct biological targets, you should provide reasonable coverage of conformational space to increase the chance of finding active conformations against any given target. So, select all compounds in the table -> click **Generate Conformers** -> select **Selected structures** and (a) the **rapid** option, (b) the **Maximum number of conformers** 100 per structure. The option Preprocess and Postprocess are recommended for very small databases or when you want to obatin high quality alignments for a set of ligands, for example to build a QSAR model (it is not the situation).
- Click start in the **Generate Conformers ->** in the green box click **START**.

8. Finding matches to a hypothesis

The pharmacophore or 3D-QSAR model developed can be used to identify new active compounds by searching in the 3D database created in the last step.

- If Maestro is running, change to the directory where the pharmacophore or 3D-QSAR model was created by writing in the ***Command text box*** :
  `cd /full_path/name_directory`
  or close Maestro and start it in the corresponding directory.
- In the ***Applications*** menu in the main window choose **Find Matches to Hypothesis** or select **Develop Pharmacophore** and click the option **Search for Matches** in the step of ***Score Hypothesis*** or ***Build QSAR Model.***
- Start using the parameters by default.
- Select 3D database in the option **Search in** -> browse and select the name of your database -> browse and search the name of a hypothesis (if it no appear)
- In the *Conformers tab* -> select **Use existing conformers**
- In the *Matching tab* -> select **Find new matches**,
- In the *Hit Treatment tab* -> select **Use QSAR model**
- Click **Start ->** write a name and click **Start** (the job should take little time because the conformers have been generated when you built the Database)

ISBN:978-84-691-1553-4 /DL. T.151-2008

- The compounds that match with the hypothesis or 3D-QSAR model are collected in the ***Project Table.*** When there isn't any compound in the ***Project Table***, it means that there isn't any solutions.
- Since most of the compounds are not able to simultaneously match in all sites, you can perform a *partial matching*. So, you have to choose to match fewer sites than the number of sites in the hypothesis. Change this number in the *Matching tab* and click **Start** again. Accordingly, in that situation, the resulting fitness score has been modified to penalize the hits that do not match all sites.
- The compounds that match with the hypothesis or 3D-QSAR model are collected in the ***Project Table.*** These solutions can be compounds that match in all sites or as a minimum match on the specified match number
- You can visualize the 3D-QSAR model or hypothesis in context of this compounds, selecting the compound in the ***Project Table*** and the hypothesis o 3D-QSAR model in the ***Develop Pharmacophore Model*** panel.

9. References

[1] S.L. Dixon, A.M. Smondyrev, E.H. Knoll, S.N. Rao, D.E. Shaw, R.A. Friesner, PHASE: a new engine for pharmacophore perception, 3D QSAR model development, and 3D database screening: 1. Methodology and preliminary results., J Comput Aided Mol Des 20 (2006) 647-671.

[2] S.L. Dixon, A.M. Smondyrev, S. Rao, PHASE: a novel approach to pharmacophore modeling and 3D database searching., Chem Biol Drug Des 67 (2006) 370-372.

[3] D.A. Evans, T.N. Doman, D.A. Thorner, MJ. Bodkin, 3D QSAR methods: Phase and Catalyst compared., J Chem Inf Model 47 (2007) 1248-1257.

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

Table . Summary of the centers that have the license for the computer program BDT (Blind-Docking Tester)

Center	Address	Scientist	e-mail
Bioinformatics Institute	Singapore	J. Thomas Leonard	thomasj@bii.a.star.edu.sg
St. Louis College of Farmacy	St. Louis, MO (USA)	John M. Beale	jbeale@stlcop.edu
Lab. LCMBA, Faculty of Science of Nice Sophie-Antipoles	Nice (France)	Charlier Landry	landry.charlier@unice.fr
Stanford University	Stanford, CA (USA)	Adam Lesser	adam.lesser@stanford.edu
University of Cincinnati.	Cincinnati, OH (USA)	Yizong Cheng	cheng@uc.edu
Laboratoire de Biotechnologies et Pharmacologie Appliquée (LBPA)	Cachan (France)	Pascal Rigolet	pascal.rigolet@lbpa.ens-cachan.fr
Auckland Cancer Society Research Center, University of Auckland	Auckland (New Zealand)	Raphael Frederick	r.frederick@auckland.ac.nz
Lab Info- LNCC/MCT	Brazil	Jorge H. F.	jorgehf@lncc.br
Quilmes National University	Buenos Aires (Argentina)	Pablo Lorenzano Menna	plmenna@unq.edu.ar
Universidad de Sao Paolo	Brazil	Vanessa Adriana Jarina, Renato Ferreira de Freitas	renatoff@usp.br
Universidad de Cordoba	Cordoba (Argentina)	Mario Alfredo Quevedo	alfredoq@mail.fcq.unc.edu.ar
Kasetsart Universiy	Bangkok (Thailand)	Kiattawee Choowongkomon	fsciktc@ku.ac.th
Dep of Biochemistry, Kangwon National University	Chunchon (South Korea)	Sanghwa Han	hansh@kangwon.ac.kr
Hong Kong Baptist University	Hong Kong	Chu Kwun Pok	06459793@hkbu.edu.hk
Shenyang Pharmaceutical University	Shenyang (China)	Jian Wang	zhaodun@syphu.edu.cn
Bryn Mawr College	Bryn Maur, PA (USA)	Judith La Londe	jlalonde@brynmawr.edu
Shriners Hospital for Children	Montreal, QC (Canada)	John S. Mort	jmort@shriners.mcgill.ca
The Scripps Research Institute (Division of Biochemistry)	La Jolla, CA (USA)	Eric F. Johnson	johnson@scripps.edu
University of Montreal	Montreal QC (Canada)	Jinjiang Fan	Jinjiang.fan@sympatico.ca
University of California	Livermore, CA (USA)	John H. Lee	lee19@lllnl.gov
Marquette University	Milwaukee, WI (USA)	Aurora Costache	aurora.costache@mw.edu
Washington University in St. Louis	Saint Louis (USA)	Garland Marshall	garland@pcg.wustl.edu
Foundation for Chemistry	Upper Artington (USA)	Stephen P. Molnar	s.molnar@sbcglobal.kef
University of Atlanta	Atlanta (Georgia)	Pahk Theopchatri	pthepch@emory.edu
University of California	Los Angeles, CA (USA)	Michael Sawaya	sawaya@mbi.ucla.edu
Melton Institute	Pittsburgh, PA (USA)	Gabriela Mustata	mustata@andrew.cmu.edu
UCSD	San Diego, CA (USA)	Rober Konecny	rok@ucsd.edu
University of Louisiana at Monroe	Louisiana (USA)	Seetharama D. Satyanarayanajois	jois@ulm.edu
Università Al Modena e Reggio Emilia	Modena (Italy)	Giulio Rastelli	rastelli.giulio@unimore.it
Biotec TU-Dresden	Dresden (Denmark)	Joan Teyra i Canaleta	
The Ohio State University	Columbus, OH (USA)	Rohit Tiwari	tiwari@pharmacy.ohio-state.edu
Institute of Life Science University of Hyderabab	Hyderabad (India)	Sachchidanand	sachchidanand@ilsresearch.org
KollegieHaldean	Lyngby (Denmark)	Liang Yang	ly@biocentrum.dtu.dk
University of North Carolina at Chapel Hill (UNC-CH) School of Pharmacy	North Carolina (USA)	Jni Hua Hsieh	jnihua-hsieh@unc.edu
Universidad Nacional de Tucuman	Tucuman (Argentina)	Rosana Chehin	rosana@fbqf.unt.edu.ar

ISBN:978-84-691-1553-4 /DL: T.151-2008

Center	Address	Scientist	e-mail
University of Oxford, Pharmacology Department	Oxford (UK)	Raman Parkesh	raman.parkesh@pharm.ox.ac.uk
University of Alabama at Birmingham	Birmingham AL (USA)	John Streiff	jstreiff@uab.edu
	Athens, Georgia (USA)	Austin B. Yongye	ayongye@chem.uga.edu
University of Rome "La Sapienza"	Roma (Italia)	Rino Ragno	rino.ragno@uniroma1.it
University College London	London (UK)	Timothy Boyle	t.boyle@ucl.ac.uk
Escola Superior Agrària de Braganca	Braganca (Portugal)	Rui Miguel Vaz de Abreu	
Institut d'Investigacions Químiques i Ambientals de Barcelona (CSIC)	Barcelona (Spain)	Jordi Bujons	jbvqob@iiqab.csic.es
CNR Instituto Chimica Biomoleculare	Italy	Mauro Marchetti	mauro.marchetti@icb.cnr.it
Dept. of Orthodontics Craniofacial Genetice Medical Center, University of Regensburg	Regensburg (Germany)	Uwe Baumert	Uwe.Baumert@klinik.uni-regensburg.de
Université de Nancy, School of Medicine	Nancy (France)	Nadir Mrabet	nadir.mrabet@medicine.uhp-nancy.fr
Liverpool University Chemistry Department	Liverpool (UK)	Neil Berry	ngberry@liv.ac.uk
Institut de Chimie des Substances Naturelles	Gif Sur Yvette (France)	B. Iorga	iorga@icsr.cnrs-gif.fr
Departamento Scienze Chimiche Farmaceutiche e Farmacologiche	Novara (Italy)	Alberto Massarotti	alberto.massarotti@gmail.com
University College London	London (UK)	Charles Allerston	ucbccka@ucl.ac.uk
Labo Medicinale Chemie	Leuven (Belgium)	Tong Li	tong.li@raga.kuleuven.be
Health Protection Agency	London (UK)	Steve Platt	steven.plat@hpa.org.uk
Dipartimento Farmaco Chimico Tecnologico	Siena (Italy)	Maurizio Botta	botta@unisi.it
The Fundation for Biomedical Research of the Academy of Athens	Athens (Grècia)	Vassilis Atlamazoglou	vatlam@bioacademy.gr
Dip. Chimica e Tecnologia del Farmaco Faculty of Pharmacy, Univeristy of Perugia	Perugia (Italy)	Antonio Macchiarulo	antonio@chimfarm.unipe.it
IBBMC Université Paris-Sud	Orsay (France)	David Perahia	david.perahia@ibbmc.u-psud.fr
Institute of Experimental Phisics	Kosice (Slovakia)	Tibor Kozar	tibor@saske.sk
Department of Bioinformatics	Kattankulathur-tamilnadur (India)	M. Vijaya	hodbisrm_04@yahoo.com
CEMB, University of the Punjab	Lahore (Pakistan)	Khalid Masood	khalid@cemb.edu.pk
Bioinformatics Institute	Singapore	J. Thomas Leonard	thomasj@bii.a.star.edu.sg
St. Louis College of Farmacy	St. Louis, MO (USA)	Jhon M. Beale	jbeale@stlcop.edu
Lab. LCMBA, Faculty of Science of Nice Sophie-Antipoles	Nice (France)	Charlier Landry	landry.charlier@unice.fr
Stanford University	Stanford, CA (USA)	Adam Lesser	adam.lesser@stanford.edu
University of Cincinnati. Dept of CS	Cincinnati, OH (USA)	Yizong Cheng	cheng@uc.edu
Laboratoire de Biotechnologies et Pharmacologie Appliquée (LBPA)	Cachan (France)	Pascal Rigolet	pascal.rigolet@lbpa.ens-cachan.fr
Auckland Cancer Society Research Center, University of Auckland	Auckland (New Zealand)	Raphael Frederick	r.frederick@auckland.ac.nz

Center	Address	Scientist	e-mail
Universidad de Cordoba	Cordoba (Argentina)	Mario Alfredo Quevedo	alfredoq@mail.fcq.unc.edu.ar
Guangzhou Institute of Biomedicine and Health	Guangzhou (China)	Jinsong Liu	liu_jinsong@gibh.ac.cn
Telhori Academic College – upper Galilee	Israel	Soliman Khatib	solimankh@migal.org.il
	Korea	Chun Kwangwoo	zeuschem@yonsei.ac.kr
Institute of genomics & Integrative Biology Delhi	New Delhi (India)	Souvik Maiti	souvik@igib.res.in
Chem. Dept. Taipei	Taipei (Taiwan)	Ying-Chieh Sun	sun@ntnu.edu.tw
National Institute of Immunology	Delhi (India)	Pankaj Khurana	pankaj@nii.res.in
School of Chemical Enginiering and Technology, Tianjin Univeristy	Tianjin (China)	Zhangxing Wang	tju@hotmail.com
ICGEB	New Delhi (India)	Amit Sharma	
Bioinformatics Institute	Singapore	Chandra Varma	
Injo University	Busan (Korea)	Hyun Joo	tachok@gurum.inje.ac.kr
TU , Department of Technical Chemistry	Braunschweig (Germany)	Avne Homann	a.homann@tu-bs.de
University of Stanford	Standford, CA, (USA)	Qingping Xu	qxu@slac.stanford.edu
	Gaithersburg, MD	Mark Semenuk	marksem@starpower.net
The Howard Florey Institute	Melbourne Univeristy	Brett Cromer	brett.cromer@florey.edu.au
University of Cagliari	Italy	Shailendra Asthana, Prof. Paolo La Colla	shailendraasthana@microbiologia.ca.it
Materials Physics and Applications - Materials Chemistry	Los Alamos, New Mexico (USA)	Brian L. Scott	bscott@lanl.gov
Pacific Northwest Research Institute	Seattle (USA)	Dr. Andrey P. Babenko	ababenko@pnri.org
Department of Biological Science, National University of Singapore	Singapore	Zhu Guili	u0402064@nus.edu.sg
Bioinformatics Institute	St.Matrix Singapore	Dr.Amor A.San Juan	amorasj@bii.a-star.edu.sg
State Key Laboratory of Virology, Wuhan university,	Hubei Province, P.R.China	Hui Cai	chncaihui@hotmail.com
TJL School of Pharmacy University of the Pacific	Stockton, CA (USA)	Wade A. Russu	wrussu@pacific.edu

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

ISBN:978-84-691-1553-4 /DL: T.151-2008

CONTRACT TO ESTABLISH THE RIGHT TO USE THE COMPUTER PROGRAM

BLIND-DOCKING TESTER (BDT):

A SOFTWARE PACKAGE FOR AUTOMATIC BLIND DOCKING ANALYSIS

BETWEEN

The Rovira i Virgili University (hereafter, URV), holder of tax identification number Q-93.50.003-A, based in Tarragona at C/ Escorxador, s/n, represented by the rector of the University, Dr. Lluís Arola i Ferrer, holder of national identity card number 39.642.182-A.

And ....., having an address at ……..…..(hereafter, RECIPIENT)
and ……..(hereafter, RECIPIENT SCIENTIST).

This document is also signed by Dr. Gerard Pujadas i Anguiano and Mrs. Montserrat Vaqué i Marquès, lecturers in the Biochemistry and Biotechnology Department of the School or Faculty of Chemistry of the URV.

The representatives of each party recognise their mutual legal capacity to draw up this document.

ANTECEDENTS

1. Dr. Gerard Pujadas i Anguiano and Mrs. Montserrat Vaqué i Marquès are the authors of the Blind-docking tester (BDT): a software package for automatic blind docking analysis computer program (hereafter, BDT Software).

2. The URV holds the exploitation rights of the above computer program (both the source program and the target program) developed by Dr. Gerard Pujadas i Anguiano and Mrs. Montserrat Vaqué i Marquès of the Biochemistry and Biotechnology Department of the URV in accordance with article 97.4 of the Law on intellectual property and article 133 of the Statute of the URV, which say: [***art. 97.4 of the Law on the intellectual property: When a paid member of staff in fulfilment of his/her job responsibilities or following the instructions of his/her manager, creates a computer program,*** the ***exploitation rights of the above computer program (both the source program and the target*** program) ***belong to the manager exclusively except otherwise agreed***] [***art. 133 of the Statute of the URV: The inventions or innovations created by a member of staff as a result of his/her activity during his/her period of employment are official property of the university***], and in accordance with the attached document where Mrs. Montserrat Vaqué i Marquès transferred her exploitation rights to URV.

3. The RECIPIENT is interested in acquiring from the URV the right to use the executable version of the BDT Software.

Therefore, both parties agree the following:

CLAUSES

1. Aim of the contract

The aim of this contract is for the URV, to allow the RECIPIENT to use the computer program BDT Software. According to article 99 of the Law on intellectual property, and in order to satisfy the needs of the user, this permission to use the program is understood to be neither exclusive nor transferable.

2. Grant of rights

ISBN:978-84-691-1553-4 /DL: T.151-2008

This contract authorizes Recipient on a nonexclusive basis to use 1 copy of the BDT Software. Recipient may retain one additional copy of the BDT Software for archival purposes. Recipient agrees to use the BDT Software for internal non-commercial research purposes only, and shall not distribute or transfer the BDT Software to anyone not under the Recipient Scientist's direct supervision or beyond the Recipient Scientist's laboratory.

3. Delivery

URV shall deliver to Recipient a master copy of the BDT Software licensed hereunder in binaries files form, suitable for execution, in electronic files only.

4. Length of contract

This permission will come into force when this contract is signed and will be granted for a period of 12 months.

The contract will be automatically renewed for 12-month periods unless one of the parties informs the other in writing at least thirty days before the end of the contract in force that they do not wish to renew it.

5. Modifications

Recipient may, from time to time, request that URV incorporate certain features, enhancements or modifications into the BDT Software. URV may, in its sole discretion, undertake to incorporate such changes, which shall be the sole property of URV, and distribute the BDT Software so modified to all or any of URV licensees. Any modifications or derivative works based on the BDT Software are considered part of the BDT Software and ownership thereof is retained by URV.

6. Copies and records

Recipient agrees to maintain appropriate records of the number and location of all copies of the BDT Software.

7. Responsibility for use of the program

The RECIPIENT is responsible for selecting the program that provides it with the desired results, installing it, operating it and using it efficiently.

This program is used at the full risk of the RECIPIENT. Under no circumstances will the authors of the program be responsible to the RECIPIENT or to third parties for any type of damages, including loss of profits or loss of savings, arising from use of the program.

8. Confidentiality of information

Both parties agree not to divulge under any circumstances the scientific or technical information that belongs to the other party and that they have had access to while carrying out the project in this contract, unless this information is in the public domain. Both parties also agree to always mention the authors of the work.

9. Modification of the contract

The parties can, by mutual agreement, cancel or modify this document at any time.

10. Cancellation of the contract

ISBN:978-84-691-1553-4 /DL: T.151-2008

The permission of use governed by this contract may be interrupted by mutual agreement of the contracting parties if they consider that the work is finished before the stipulated time, or for any other reason.

11. Rescission of the contract

If, for reasons that can be attributed to either of the two parties this contract is not fulfilled, it will be automatically rescinded.

12. Jurisdiction

The RECIPIENT and the URV undertake to resolve amicably any disagreement that may arise from this contract.

In case of conflict deriving from this contract, both parties agree to submit to the institutional arbitration of the ***Tribunal Arbitral de Tarragona*** (Arbitration Court of Tarragona) of the ***Associació per a l'Arbitratge a Tarragona*** (Arbitration Association of Tarragona), which will be responsible for designating an arbiter or arbiters and for administrating the arbitration, and to accept the arbitration decision. Decisions will be taken and formalities will be carried out in the city of Tarragona.

As proof of conformity, we sign this document in triplicate in the place and on the date specified below.

e-mail address of the RECIPIENT: _____

RECIPIENT location and date: _____

For the **Rovira i Virgili University**	**For** **(full name of the RECIPIENT).....**
Dr. Lluís Arola i Ferrer Rector	Mr/Ms ....
Mrs. Montserrat Vaqué i Marquès Person responsible for software	Dr. Gerard Pujadas i Anguiano Person responsible for software

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO STUDIES OF THE EFFECT OF PHENOLIC COMPOUNDS FROM GRAPE SEED EXTRACTS ON THE ACTIVITY OF PHOSPHOINOSITIDE 3-KINASE (PI3K) AND THE FARNESOID X RECEPTOR (FXR)
Montserrrat Vaqué Marqués
ISBN:978-84-691-1553-4 /DL: T.151-2008