Chemoisosterism and its impact on drug polypharmacology

Xavier Jalencas i Giménez

TESI DOCTORAL UPF / ANY 2013

DIRECTOR DE LA TESI:

Dr. Jordi Mestres

CEXS Department

The research in this Thesis has been carried out at the Chemogenomics Laboratory (CGL), within the Research Programme on Biomedical Informatics (GRIB) at the Parc de Recerca Biomèdica de Barcelona (PRBB).

The research presented in this Thesis has been supported by Ministerio de Ciencia e Innovación Project BIO2005-041171, BIO2008-02329, BIO2011-26669 and PTA2009-1865P.

Printing funded by the Fundació IMIM's program "Convocatòria d'ajuts 2013 per a la finalització de tesis doctorals de la Fundació IMIM."

Al Quim, la Mar i la Montse.

Agraïments

Voldria donar les gràcies en primer lloc a tots aquells qui han contribuït a fer possible aquesta tesi, en especial al meu director de tesi, Jordi Mestres, per donar-me la oportunitat de desenvolupar la recerca que aquí es presenta i per tot l'esforç invertit en ella. Gràcies també per tot el recolzament, suport i paciència al llarg de tots aquests anys. Menció especial mereixen tots els companys amb qui he compartit laboratori al llarg de tots aquest anys: Bet, Rut, Montse, Ricard, Barbara, Jan, Ana, Fabio, Ferran, Miguel Ángel, Praveena, Núria, Ieda, Irene, David, Nikita, Ingo, Alfons, Albert, Andreas, Maricarmen, Joaquim, Sergio i Viktoria. A aquesta llista cal afegir-hi també tots els companys que han passat pel GRIB i han fet l'estada més fàcil i agradable: Alícia, Òscar, Carina, Alfons, Judith, Miguel Ángel, Martina, Ángel, Marta, Laura, Cristian i tots aquells que mereixerien ser igualment esmentats i que de ben segur m'estic deixant. També agrair als companys de la UPF els innombrables dinars i sobretaules que hem compartit al llargs de tots aquest anys de carrera i doctorat: Anna, Gerard, Jaume, Javi, Manolo, Mireya i Oriol. Finalment, no puc deixar de donar les gràcies a tota la família, en especial la Montse, pel suport rebut durant aquesta etapa.

A tots vosaltres, gràcies de tot cor.

Table of contents

útils per a la construcció de llibreries de fragments moleculars dirigides a una proteïna diana en particular. Partint de la premissa que entorns de proteïna similars molt probablement interaccionaran amb fragments moleculars similars, aquesta Tesi presenta un nou mètode per a identificar entorns de proteïna similars, utilitzat per predir noves relacions quimioisostèriques. S'aporten també alguns exemples de potencials aplicacions del quimioisosterisme en la disciplina del descobriment de fàrmacs. Un anàlisis de les implicacions que té el quimioisosterisme en la polifarmacologia ens duu a la hipòtesis de que els nivells de polifarmacologia observats en la majoria de fàrmacs no són res més que una signatura de l'explotació del quimioisosterisme al llarg de l'evolució.

Abstract

In medicinal chemistry, two chemical fragments are considered bioisosteric if they bind to the same protein environment. Accordingly, looking at the same players from an opposite perspective, two protein environments can be considered chemoisosteric if they interact with the same chemical fragment. In this respect, this Thesis introduces the term chemoisosterism, which represents a new concept in drug discovery. Currently available crystal structures for protein-ligand complexes constitute a basis for the identification of chemoisosteric protein environments, of great utility for the construction of focused fragment chemical libraries. Under the premise that similar protein environments will probably bind to similar fragments, a novel approach to assess protein environment similarities is introduced and used to predict new chemoisosteric relationships. Examples of the potential applicability of chemoisosterism in fragment-based drug discovery are provided. The implications of chemoisosterism for drug polypharmacology are explored, leading to the speculation that the levels of polypharmacology observed in current drugs may just be a latent signature of the exploitation of chemoisosterism during evolution.

Resum

En química mèdica, dos fragments moleculars són considerats bioisostèrics si s'uneixen al mateix entorn de proteïna. Canviant la perspectiva sobre el mateix esdeveniment, dos entorns de proteïna poden ésser considerats quimioisostèrics si interaccionen amb el mateix fragment molecular. Aquesta Tesi introdueix el terme quimioisosterisme, un nou concepte en química farmacèutica. Les estructures actualment disponibles de complexos de proteïna i lligand constitueixen una font d'entorns de proteïna quimioisostèrics potencialment

Preface

Drug design is a long and costly process very prone to failure, albeit of utter importance to provide new therapies to improve the health of a an increasingly aging population that is subject to new threats in the form of chronic or degenerative maladies associated to age, such as cancer, cardiovascular problems and diabetes. New diseases like Acquired Immunodeficiency Syndrome (AIDS), avian influenza or Severe Acute Respiratory Syndrome (SARS) also need to be addressed, as well of other endemic re-emerging threats such as malaria and tuberculosis. The massive amount of information involved in the process of drug discovery makes it an extremely complex activity of information management and interpretation. It is here where computational tools have been proved very useful to assist in decision making in any of its steps, pursuing the aim of reducing the required costs in both resources and time to bring a new drug to the market as well as providing new predictions that may ultimately lead to new advances in therapeutics.

The increasing generation of pharmacological data for small molecules and its public availability has boosted research and development in the area of computational tools for in silico pharmacology. In particular, virtual ligand screening has become a valuable tool to predict the likelihood of a chemical compound of having affinity for a certain target. In the last years several such ligand-based approaches for in silico target profiling have led to the successful identification of new targets for old drugs. However one of the inherent limitations of the use of ligand-based information, as a result of its incompleteness and bias, is the limited hopping ability, in phylogenetic terms, of the new targets predicted.

Therefore, in order to achieve much degree of hopping one might to go beyond ligand-based methods. A possibility in this regard is to incorporate all available knowledge on protein structure publicly available in the Protein Data

Bank. Some research has been done in that direction, leading to what is usually referred as inverse docking. However the relatively large number of false positives identified from such approaches, alongside with their computational costs that make profiling millions of small molecules currently unaffordable; reveals the need for new and faster structure–based approached to in silico target profiling.

This is precisely the aim of this Thesis; to contribute to fill in this gap by developing and exploring a structure-based approach focused on protein binding sites. The implementation of a method to describe and compare binding sites has been addressed, as well as the evaluation of several of its possible applications in the field of drug discovery. The Thesis has been divided in 5 parts. Initially a general overview to structure-based drug discovery is provided, along with a perspective on protein-ligand binding characteristics. Special emphasis is put in bioisosteric replacements and fragment-based drug discovery, as a special focus will be put in them in subsequent sections and discussions. After this introductory section, the main objectives of the Thesis will be listed, followed by the main achieved results, including a manuscript and the three publications that have resulted from this Thesis. Finally, a discussion and the main conclusions derived from this Thesis will be outlined. A bibliographic section containing the list of cited references will conclude the document.

List of publications

From this Thesis

Articles:

- Jalencas, X.; Mestres, J. Identification of similar binding sites to detect distant polypharmacology. Mol. Inf. Submitted. minf.201300082.
  Journal Impact Factor: 2.39.

- Jalencas, X.; Mestres, J. Chemoisosterism in the Proteome. J. Chem. Inf. Model. **2013**, 53, 279–292.
  Journal Impact Factor: 4.675.

- Jalencas, X.; Mestres, J. On the origins of drug polypharmacology. Med. Chem. Comm. **2013**, 4, 80-87.
  Journal Impact Factor: 2.8; Citations: 5.

Oral communications:

- Jalencas, X.; Mestres, J. A knowledge-based approach to assessing the target promiscuity of chemical fragments. Oral communication presented at 9th International Conference on Chemical Structures. 2011 Jun 5-9. Noordwijkerhout. Netherlands.

Poster communications:

- Jalencas, X.; Mestres, J. Structural hopping between protein cavities. Poster presented at VIII Jornadas de Bioinformática. 2008 Feb 13-15. Valencia. Spain .

- Jalencas, X.; Mestres, J. Indexing cavities in protein structures. Poster presented at 21st International Symposium on Medicinal Chemistry. 2010 Sep 5-9. Brussels. Belgium.

From other collaborations

Articles:

- Antolín, A. A.; Jalencas, X.; Yélamos, J.; Mestres, J. Identification of Pim Kinases as Novel Targets for PJ34 with Confounding Effects in PARP Biology. ACS Chem. Biol. **2012**, 7, 1962–1967.
  Journal Impact Factor: 6.446; Citations: 3.

Part I: Introduction

Drug discovery and development, from initial target identification, to optimization of drug candidates and final regulatory approval, is a long and costly process very susceptible to failure. The costs of discovering and developing an average drug, a process that typically lasts around 12 years, were estimated to lie between $800 million[1] and $1.2 billion[2] for the 1989-2001 time period, a figure that has been growing ever since to a recent estimation of $1.8 billion in 2010.[3] Besides its cost, ultimate success is never guaranteed and the process suffers from a very high attrition rate. Although success rates significantly differ between therapeutic areas (from around 20% in cardiovascular to 5% in oncology), an average success rate of 11% was observed during the 1991-2000 decade.[4]

Additionally, over the past years there have been major progresses in technological and scientific fields which have been incorporated to the drug discovery and development pipelines. Examples of such advances include combinational chemistry[5], which increased the rate at which drug-like molecules could be synthesized; high-throughput screening (HTS)[6], which greatly reduced the cost of testing compound libraries against protein targets; DNA sequencing, extremely faster and cheaper than it was some years ago[7], which has allowed the sequencing of the complete genome of multiple species; molecular modelling; metabolomics and systems biology. At the same time advances in scientific knowledge provide new drug targets or insight on disease mechanisms.

Nevertheless, despite those advances and new technologies, initially expected to positively impact drug discovery and development, the cost of drug discovery and development has been uninterruptedly growing (Figure 1).[8] The massive amount of information involved in the process makes appropriate the use of computational methods to provide tools that help in decision making in many of its steps improving in this way its efficiency and reducing its cost both in time and resources. Although cost and failure-rates are much higher at

clinical phases of drug development[4] and the impact of computer-aided drug design is mainly in pre-clinical stages such as target validation, hit discovery or hit to lead optimization, its purpose lies in providing better drug candidates that will reduce costs and improve success rates in subsequent clinical steps.

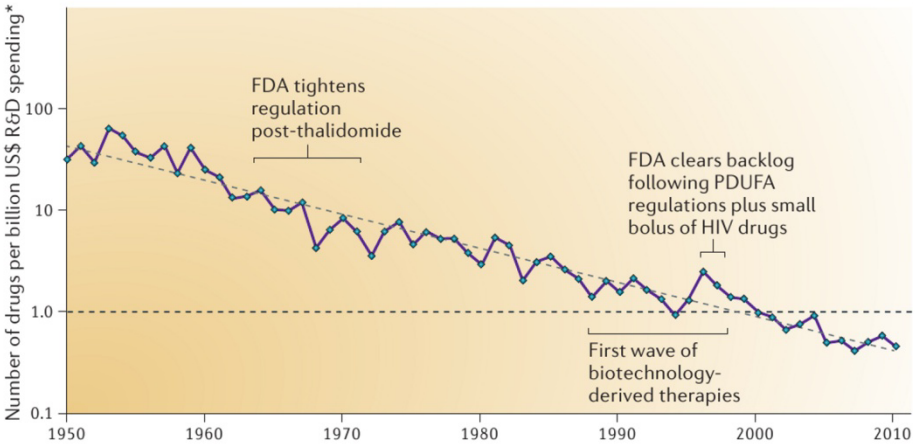


Figure 1: The number of new drugs approved by the US Food and Drug Administration (FDA) per billion US dollars (inflation-adjusted) spent on Discovery and Development (R&D). Extracted from Scannell *et al.*.[8]

In particular, the increasing generation of pharmacological data for small molecules and their availability in the public domain[9–11] has boosted research and development in the area of computational tools for in silico pharmacology.[12,13] This amount of information is efficiently exploited by ligand-based approaches to in silico target profiling which have emerged in the last years and that have successfully led to the identification of new targets for old drugs.[14–20] Based on the premise that similar molecules will probably exhibit similar properties and hence bind to similar targets, their usual workflow consists on predicting the targets and affinities of a particular molecule based

on known targets and affinities of similar molecules.[21,22] Their main difference resides on the molecular descriptors and metrics used to derive a similarity, which can be highly diverse: two or three-dimensional descriptors encoding different features or substructures are fed to a variety of classification schemes and similarity metrics.[23,24] However one limitation of the use of ligand-based information for predicting new targets is the limited novelty, in phylogenetic terms, of the new targets predicted. It can only explore the chemical space around molecules with known targets, which is known to be biased towards some target families such as GPCRs.

To overcome those limitations and attain a higher degree of hopping, as well as increment the usable information for other protein families (e.g. enzymes), the utilization of complementary methods are needed. Structure-based approaches are suitable candidates to assume this role, as they exploit a different body of information that ligand-based methods do, therefore its results are likely to be mostly different from those achieved by ligand-based approaches (Figure 2).

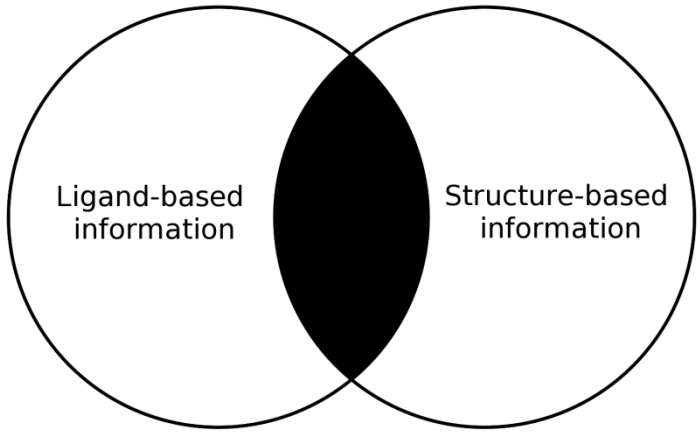


Figure 2: Venn diagram illustrating complementarity of ligand-based and structure-based approaches.

I.1 Chemogenomics

It is precisely to integrate all available chemical and biological information that chemogenomics has emerged as a new paradigm in drug discovery. Drug discovery has been in the last decades a multidisciplinary endeavour to optimize a compound's potency, selectivity and pharmacokinetics primarily towards a single target. Following a one drug-one target premise, its main objective has been to develop potent and selective small molecules against a particular target. The sequencing of the human genome[25] has allowed to estimate the existence of about 3,000 druggable targets[26,27], being only a small fraction of them investigated by pharmaceutical industry.[28] In fact, the average central nervous system drug has, for example, affinity on over 20 receptors.[29] The relative affinity of a given drug for all those receptors was never optimized per se but was "a given" after the molecule was optimized for one single receptor. Pergolide, an anti-parkinsonian drug, is usually referred as a "dopamine agonist", even though it has affinity for over 20 receptors. One of them, the serotonin receptor subtype 5-HT_{2B}, has been identified as the ultimate responsible of its cardiac valvulopathy[30] safety risk that forced its withdrawal from the market. With only a small portion of both chemical and target spaces been explored chemogenomics multi-target strategies emerge as an especially attractive approach to conceptually shift from the rather limited number of druggable targets to the millions of target combinations of potential therapeutic relevance.

The term chemogenomics was defined in 2001 as the "discovery and description of all possible drugs to all possible drug targets".[31] It consists on organizing drug discovery by protein families[32] in order to maximize the efficiency of biology and chemistry resources by the obtainment and accumulation of reusable knowledge across a target family.[33] As chemogenomics dwells at the interface between chemistry and biology,

computational tools integrating bioinformatics and chemoinformatics are required to extract and integrate reliable information. In short, chemogenomics aims at completing a two-dimensional matrix of targets and compounds (as columns and rows) with values of binding.[34] As this matrix is far from complete, predictive chemogenomics tools aim to fill in its existing gaps by anticipating new compound-protein relationship. A fragment of this matrix related to cardiovascular diseases is shown in Figure 3.[35]

Integrative chemogenomics tools need to tackle three main tasks, namely: annotate and classify data, generate and integrate knowledge and rational and systematic design.[33] Classification schemes and ontologies, both on the biological and chemical side are essential to establish relationships between proteins and ligands in an unambiguous manner. Several resources exist providing annotated gene and protein sequences[36–39] and strcutures[40–42], while several tools exist to univocally identify chemical compounds.[43,44] Once classification and annotation schemes are ready, analysis of data allow to establish links between proteins and ligands. Structural data on proteins and protein-ligand complexes, structural data of ligands and ligand activity, are useful to predict new protein-ligand interactions, leading to structure-based and ligand-based chemogenomics approaches. This is done under the assumption that similar ligands will likely bind the same target and that similar targets will likely interact with the same ligand. How similarities those similarities are obtained will characterize a particular chemogenomics approach. For example, Figure 3 illustrates the application of a chemogenomics approach based on cross-pharmacology to complete the cardiovascular target space and infer potential cardiovascular off-targets.[35]

.

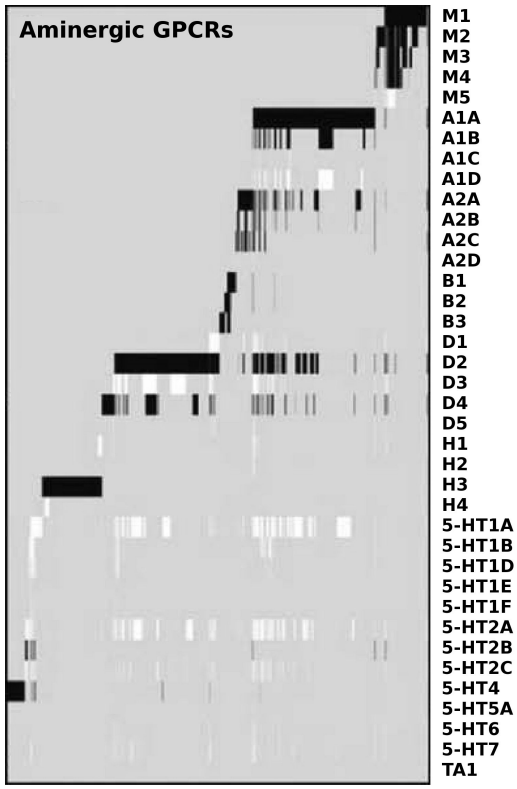


Figure 3: Ligand–protein interaction map between scaffolds from molecules annotated to cardiovascular targets (in columns) and the cardiovascular-relevant aminergic GPCRs (in rows). Black marks are extracted from literature mining while white marks are identified from cross-pharmacology relationships. Figure adapted from Cases and Mestres.[35]

I.2 Protein-ligand interactions

By looking in detail into the interface between the biological and chemical space protein-molecule interactions are found. Interactions between proteins and small molecules (hereby named ligands) play an important role in many cellular functions such as enzyme catalysis or signal transduction. Alteration of those interactions is often implicated in disease and can be taken advantage of to deliberately modulate altered functions. Detailed knowledge on molecular recognition between proteins and ligands provides useful information in many fields, including drug discovery, as most drugs exert their function through a binding event with a protein. From a thermodynamics point of view, the affinity from a ligand to a protein can be described as a Gibbs free energy (ΔG), which is linked to the experimental binding constant (K_B). The Gibbs free energy can be decomposed in its enthalpic and entropic components, being enthalpy related to the internal energy of both the protein and the ligand and entropy to the degree of disorder of the system. For a binding event to occur, a desolvation of the ligand and the binding site is followed by a certain conformational change of both, after which interactions are formed between ligand and protein.[45] The main forces driving molecular recognition include electrostatic forces and the hydrophobic effect. Simple electrostatic forces include ion-ion and ion-dipole interactions and Van der Waals forces (attractive between permanent or induced dipoles and repulsive between electron densities defining the molecular volume).[46,47] Combining those forces hydrogen bond interactions appear. Those occur when an electronegative atom (named acceptor) and a hydrogen atom attached to a second electronegative atom (donor). Hydrogen bonds also have the remarkable feature of being directional; contributing is such way to the specificity of intermolecular interactions. Combining Van der Waals, hydrophobic and electrostatic forces aromatic interactions involving π systems exist. This kind of interaction combines the

strength of the hydrophobic effect with the selectivity of electrostatic interactions. The electrostatic component is attributed to quadrupole moments in aromatic rings, where a greater electron density exists on the faces of an aromatic ring relative to their edges. This favours certain geometries such as face to face (stack), edge to face (T-shaped), parallel displaced (offset stacked), and cation-π .[48]

A major component of the forces that stabilizes biological structures is the hydrophobic effect, which is defined by the IUPAC as a "tendency of hydrocarbons (or lipophilic hydrocarbon-like groups) to form intermolecular aggregates in an aqueous medium". It is believed to be mainly entropically driven; as hydrocarbon molecules are not solvated due to their incapacity to form hydrogen bonds with solvent water molecules, those waters become more ordered around the hydrocarbon molecule than in bulk water. This leads in a higher degree of order in the system (loss of entropy), that can be counterbalanced by aggregation of hydrocarbon structures, reverting on an entropic gain as their contact surface to solvent is reduced. An example of interaction pattern in a sample complex structure from the PDB is shown in Figure 4 showing hydrophobic contacts, hydrogen bonds and aromatic stacking.

A precise analysis of the balance of forces governing any ligand-receptor specific interaction (using docking or molecular dynamics; empirical scoring or energy functions) still remains a challenge due to the uncertainty of entropic contributions to the binding , like hydrophobic effects as well as Van der Waals interactions and the role of water.[47] Scoring functions use empirical formulas and approximations assuming that a binding affinity can be defined as a sum of independent terms, disregarding cooperativity and non-additivity of molecular interactions.[49] Furthermore, their computational cost is still too demanding to afford high-throughput calculations.

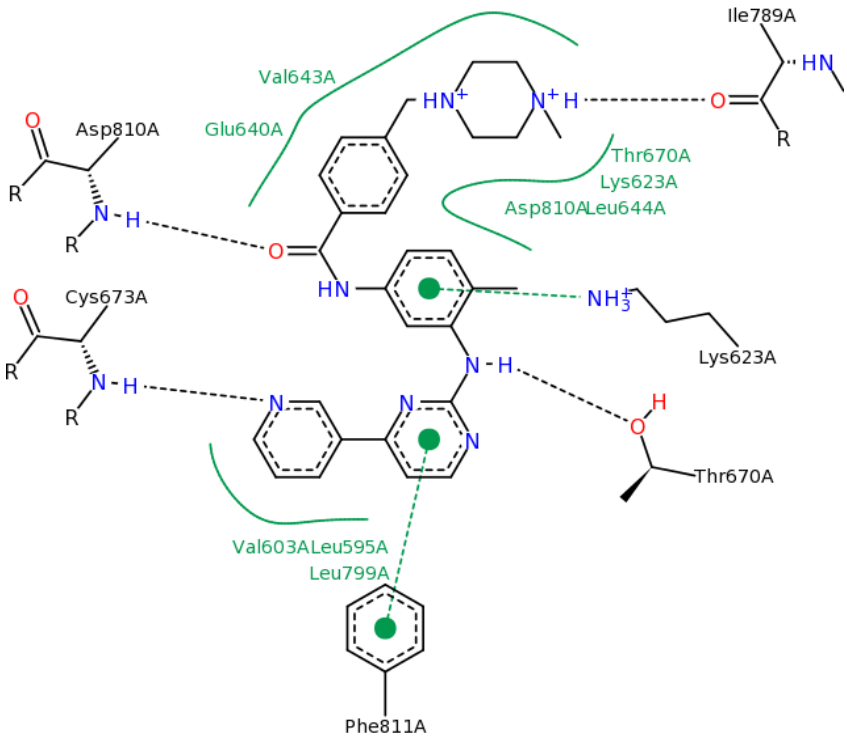


Figure 4: Diagram of predicted interactions patterns for a v-kit Hardy-Zuckerman 4 feline sarcoma viral oncogene homologue in complex with imatinib (PDB code: 1T46). Hydrogen bonds are represented as black dashed lines; π interactions as green dashed lines with dots marking the interacting π systems. Hydrophobic contacts are represented by green residue labels and green splines along the contacting ligand region. Figure extracted from Stierand and Rarey.[50] The interactions are predicted by the PoseView software.

Such limitations enable the development of a wide range of different approximations aimed to simplify the problem. The molecular recognition was initially described by the lock and key image, where geometrical and physicochemical complementarity is often required between the protein and the ligand. In a more realistic model accounting for protein flexibility; the induced-

fit model, the protein undergoes a certain conformational change to be able to bind the ligands.[51] An alternative mechanism named selected-fit consists on the ligand selecting and stabilizing a complementary protein conformation among different protein conformations in equilibrium.[52] Those models allows to reduce binding predictions to molecular similarity and target complementarity.[53]

With an increasing number of protein structures available in the Protein Data Bank[54], there is a sufficient corpus of data on protein-ligand complexes to allow extraction of knowledge that can be further applied in molecular recognition research and structure-based drug design. An effort has been made during the last decades to collect all those information in publicly available databases of protein-ligand interactions along with related useful information such as ligand similarities, interaction patterns. IsoStar, Relibase, CREDO and many others are examples of such databases (see Table 1 for details).[55–59] Such amount of data allows the existence of approaches that estimate the probability of a certain interaction based on how often is observed in x-ray structures. SuperStar is one of such approaches, that uses used the experimental knowledge stored in IsoStar to predict drug-targets interactions by obtaining the propensity of different probes at different positions around the template protein binding site.[55,60]

Notwithstanding, being all protein-interaction data derived from protein structures, most of them resolved by x-ray crystallography, several issues arise. Crystallization might produce new intermolecular interactions that may affect the structure of the complex and crystallization conditions, which may not be the same as biological ones, can also alter protein-ligand recognition event. Considering such caveats, the Protein Data Bank is an excellent source of structural information on protein-ligand biding. Exploitation of such structural information to enhance drug design constitutes what is known as structure-based drug discovery.

Table 1. Available protein-ligand interaction databases

Name	*Description*	*Site*
AffinDB[61]	Affinity database for protein-ligand complexes	http://pc1664.pharmazie.uni-marburg.de/affinity/
BindingMOAD[62]	Subset of the PDB containing every high-quality example of ligand-protein binding.	http://bindingmoad.org/
CREDO[59]	A Structural Interactomics Database For Drug Discovery	http://marid.bioc.cam.ac.uk/credo
Het-PDB Navi[63]	Database for protein-small molecule interactions.	http://daisy.nagahama-i-bio.ac.jp/golab/hetpdbnavi.html
IsoStar[55]	A knowledge-based library of intermolecular interactions	www.ccdc.cam.ac.uk/Solutions/CSDSystem/Pages/IsoStar.aspx
LigandExpo[64]	Chemical and structural information about small molecules in the PDB.	http://ligand-expo.rcsb.org/
LigBase[65]	Ligand binding proteins aligned to structural templates.	http://modbase.compbio.ucsf.edu/ligbase/
PDBbind-CN[66]	Experimentally measured binding affinity data for complexes deposited in the PDB	www.pdbbind.org.cn/
Relibase[56]	Searching, storing and analysing 3D structures of protein-ligand complexes.	www.ccdc.cam.ac.uk/Solutions/FreeSoftware/Pages/Relibase.aspx
sc-PDB[67]	Druggable Binding Sites from the Protein Data Bank	http://bioinfo-pharma.u-strasbg.fr/scPDB/

I.3 Structure-based drug discovery

As it has already been mentioned, in order to achieve a higher degree of phylogenetic hopping and explore a wider region of the chemical space one ought to go beyond ligand-based methods. A possibility is to incorporate the vast amount of protein structural data available publicly in the Protein Data Bank[54] (the major repository for biomolecule structures, Figure 5). Computational tools that make use of protein structural information are generally referred as structure-based methods. In this case, its applicability is limited to structural knowledge, which is again biased, but this time towards enzymes.[68] This contributes to make both structure and ligand-based methods complementary in nature.

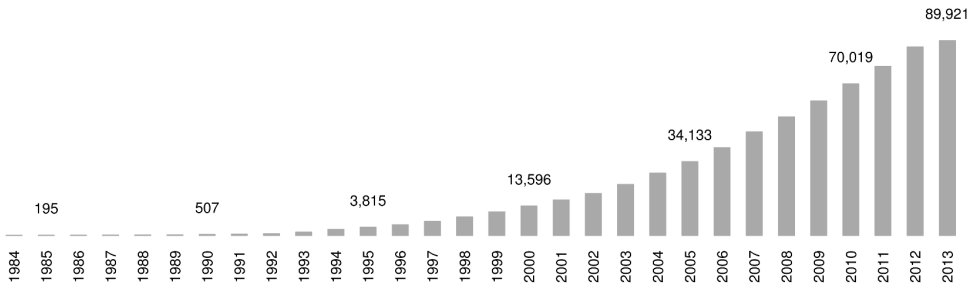


Figure 5: Number of searchable structures available in the Protein Data Bank for the last three decades. An exponential growth through time is observed resulting in a current number of almost 90,000 structures. Data extracted from www.rcsb.org on April 2013.

Structural knowledge of a particular protein provides insights into the molecular basis of its biological function, as well as information on its structural features governing its interactions with ligands, valuable data that can be incorporated in a drug discovery project. One of the most common structure-

based methods to probe the interaction of small molecules in multiple protein structures is often referred to as inverse docking, where multiple small molecules are docked to a receptor in an attempt to discover putative ligands.[69] Docking tools aim to predict the binding site location for a ligand and its conformation when binding to a particular protein by computational methods. It roughly consists on exploring the conformational space of the ligand in the protein binding site and ranking these conformations by means of a scoring function that allows finding the best ligand pose. Such scoring schemes usually rely on a combination of geometric, energetic and empirical functions, which differ from one docking tool to another.[70]

As highlighted above, accurately predicting a binding free energy is still challenging due to the variety and complexity of forces that drive protein-ligand binding events. In addition, the computational cost of such calculations is currently unaffordable if aimed to profile millions of small molecules in thousands of protein targets. New and faster structure-based approaches to in silico target profiling are thus needed to fill in this gap and complement existing ligand- and structure-based methods in order to spread the chemical space that can be explored. This is the ultimate aim of this Thesis: The development of a fast and novel structure-based approach to predict the binding of small molecules to multiple protein structures.

I.4 Fragment-based drug discovery

Fragment-based drug discovery (FBDD) is an approach mainly used for finding lead compounds in the drug discovery pipeline that is becoming popular and widespread among pharmaceutical companies. It is based on identifying small low affinity molecules, or chemical fragments, that are subsequently optimized to lead molecules of higher affinity, either by growing or linking them (Figure 6).[71–73] The strengths of FBDD lie mainly in the fact that the fragment space of chemical diversity can be sampled more effectively than the much bigger drug-like chemical space (even though the explored fragment space is still small). Using NMR or X-ray techniques it is possible to discern weak binders from a set of fragments with significant higher hit-rates than high throughput screening (HTS) can yield when detecting active compounds. This is not a surprising performance, as it is consistent with an already described trend of less complex molecules to exhibit a higher promiscuity.[74,75] As molecular complexity increases, chances of a mismatch between the ligand and the receptor that can disrupt the interaction also increase. The hit rate gain provided by FBDD acquires special relevance when difficult targets, like protein-protein sites, are involved.

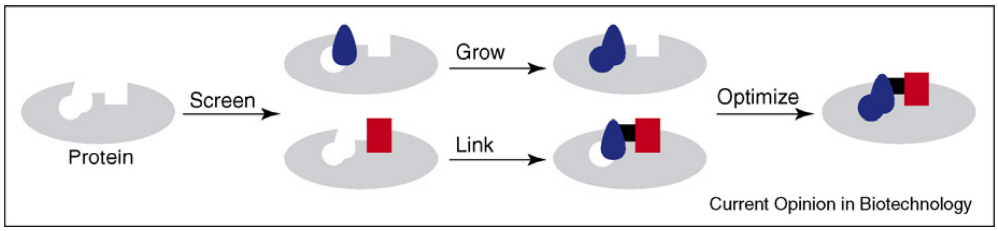


Figure 6: Schema illustrating the basis of fragment-based drug discovery. The blue and red shapes represent fragments that bind to the target protein. These can be linked or expanded to produce high-affinity ligands. Figure extracted from Erlanson.[71]

Although chemical fragments tend to have weak potencies (generally K_d between 100μM and 10mM), they usually exhibit similar binding efficiencies as larger molecules. Ligand efficiency (or ligand efficiency), the binding energy of a molecule divided by the number of its heavy atoms, has been gaining usage since it was shown that high molecular weight is not indispensable for high binding affinity.[76] A "Rule of three" has been proposed to select suitable fragments (≤3 hydrogen bond acceptors, ≤3 hydrogen bond donors and Clog ≤3).[77] Due to their reduced complexity and high binding efficiency fragments are usually suitable starting points for hit to lead optimisation, as they allow for more freedom to property optimization.

Successful examples of fragment-based lead discovery are starting to be common.[78] One marketed drug for metastatic melanoma, vemurafenib, was developed by Plexxikon in 2005 using this approach and approved in 2011. A compound developed by Merck against Alzheimer's disease entered phase II/III trials in late 2012. A fragment approach allowed in this case the targeting of β-secretase (BACE1), a target that was traditionally considered almost undruggable. Several other examples of drug candidates discovered by such means are now in phase II.

Besides its inherent advantages, fragment approaches have several drawbacks: the low affinity and size of fragments makes them more difficult to identify, forcing the screenings to detect binding instead of inhibition. The number of interaction sites on protein surfaces able to accommodate low-weight compounds such as solvents might result in false positives. On the other side, screening fragments by X-ray diffraction or NMR has the advantage that the binding pose is determined at the same time a hit is found, although the fact that the cavity might be significantly larger than the fragment volume can result in incorrect binding modes.

Application of computational tools such as docking to fragments still remains challenging. Most scoring functions have been developed to reproduce energies of drug-like compounds having much higher affinity than fragments. Since the development of a reliable scoring scheme for drug-like molecules is still not entirely resolved, scoring and ranking fragments forming fewer interactions can be even more problematic. Even though, docking under pharmacophoric constraints or post-docking processing with interaction fingerprints can be used to prioritize relevant poses for low-molecular weight fragments.[79]

Given the complexity of the drug discovery process, combinations of fragment-based approaches with other existing methods for lead discovery seems to be the best option, so all its advantages are exploited while its drawbacks minimized. The results of a fragment-based screening, coupled with X-ray crystallography provide an invaluable source of information for identifying bioisosteric fragments.[80]

I.5 Privileged substructures

The low specificity of chemical fragments discussed in the last section is consistent with the concept of privileged scaffolds or substructures. The term "privileged scaffold" was first used by Evans *et al.* [81] referring to a molecular framework that is able to bind to a diverse set of receptors. Although the term initially made reference to the benzodiazepine core, many other privileged scaffolds have been identified since then. Simultaneously, the definition of privileged scaffold has been loosened to a scaffold that is frequent in bioactive molecules. These privileged scaffolds were suggested to provide affinity to the target, while selectivity would be introduced by variations on the decoration of the scaffold with different chemical groups. A comprehensive list of such scaffolds was assembled by Welsch et al.[82], being the fragment-like structures of this list shown in Figure 7 (extracted from Barelier and Krimm[83]). Noteworthy, most of them are formed by rigid and aromatic ring systems, well suited for binding hydrophobic pockets in the target protein.

Privileged structures may guide the design of chemical libraries, as enrichment in molecules containing such scaffolds will likely produce higher hit rates as well as throwing hits with enhanced drug-like properties.[84] Discovering novel privileged scaffolds in not an easy task. Hajduk and co-workers used NMR derived binding data of 11 targets to identify molecular motifs preferred for protein binding.[85] Most of the structures identified were already considered as privileged, suggesting that a significant amount of privileged scaffolds is already known. Nevertheless, the fraction of explored chemical space is so small, that it is highly probable that there is plenty of space for advances in discovery of privileged scaffolds. In this direction, natural products are a source of chemical structures that differ from those usually found in chemical libraries, offering a chance to the discovery of new privileged scaffolds.

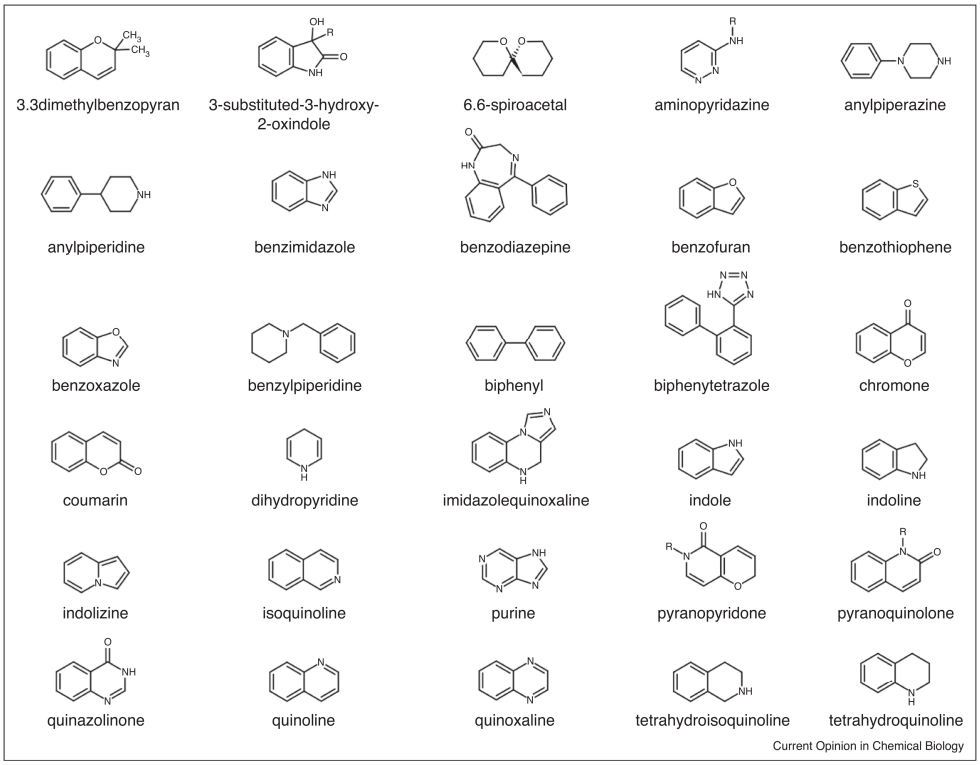


Figure 7: Fragment-like privileged scaffolds from drugs and natural products. Extracted from Barelier and Krimm.[83]

I.6 Bioisosterism

The concept of bioisosterism has long been used to describe functional groups that independently of their similarity can form similar intermolecular interactions, thus retaining the biological activity of the compound.[86] In medicinal chemistry, bioisosteric replacements are often employed to modulate compounds, especially in lead optimization process. Moreover, its application also include hopping from one lead structure of a competitor to another structure outside patent coverage and the modification of the structure of a compound with a suboptimal pharmacodynamic profile. A series of common bioisosteric replacements for kinase drug candidates from different companies are shown if Figure 8 as an illustrating example. [87]

A more generic concept of isosterism was first introduced by Langmuir in 1919 to define atoms and compounds with the same arrangement of electrons.[88] It was not until the early 1950s when the term bioisosterism started to be used as the concept of isosterism was applied to biological molecules.[89] Since then, a wealth of bioisosteric replacements has been described and successfully applied in drug discovery projects.[90,91] For example, bioisosteric pairs extracted from the literature are available from the BIOSTER database and several methods to automatically identify bioisosteres have been reported, both knowledge-based or *ab initio*.[92] Wagener and Lommerse[86] described a strategy for suggesting bioisosteric replacements based on fingerprints encoding topological pharmacophore information. Also, the IsoStar[55] database contains crystallographic and theoretical data on intermolecular non-bonded interactions that can be used to identify bioisosteric replacements. A method to identify potential target-specific bioisosteres analyzing sets of different ligands complexed with structures of a given protein has been recently described by Kennewell et al.[80] Information of bioisosteric equivalence has also been used in similarity-based virtual screening to improve its performance.[93] Opposite to

bioisosteric replacements, chemical substitutions that produce an activity cliff (structurally similar compounds with high potency differences, usually unexpected challenge SAR approaches)[94] can also be exploited for lead optimization.[95]

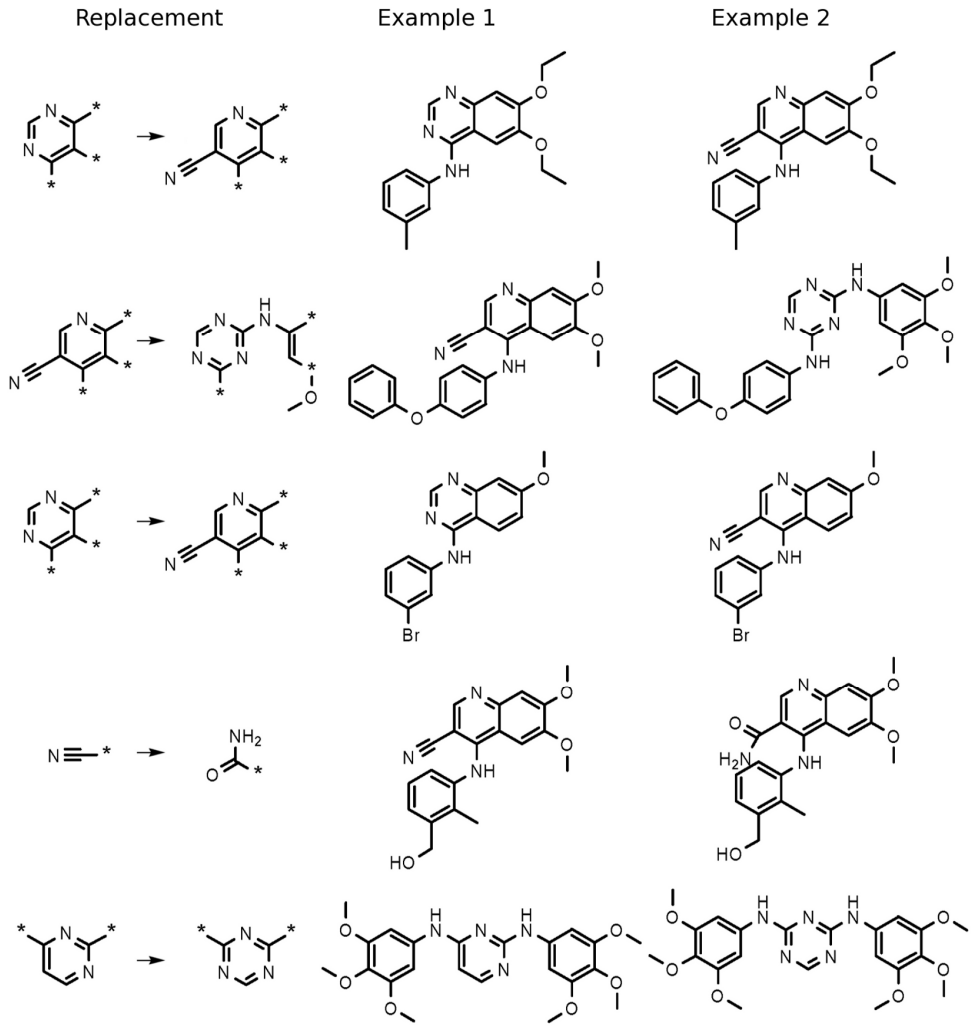


Figure 8: Most frequent replacements between pharmaceutical companies ranked by the total number of examples that connected the two companies. The chemical replacement, a typical compound from the first company and a typical compound from the second company are shown. The most common replacement observed (first row) connects 70 molecules from AstraZeneca (left) and Wyeth (right). Figure adapted from Southall and Ajay, only first 5 replacements are depicted.[87]

I.7 Chemoisosterism

As stated in the previous chapter, bioisosterism applies to different chemical groups being able to interact with the same protein environment. Moving the point of view from the protein environment to the chemical fragment, a counterpart concept to bioisosterism emerges to define all the protein environments that are compatible with the same chemical fragment. Stressing on the complementarity to bioisosterism, the term "chemoisosterism" was coined to define this concept.[96] Figure 9 shows three chemoisosteric protein environments that are compatible with a phenyl ring.

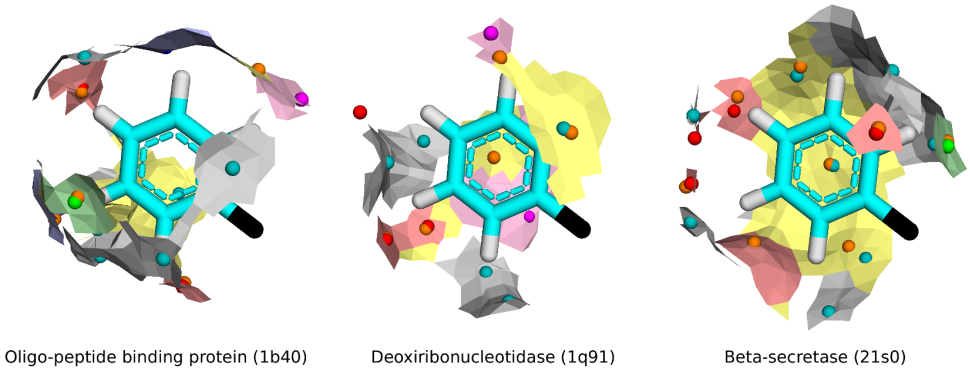


Figure 9: Three chemoisosteric protein environments compatible with a phenyl group. Protein surface close to the phenyl is shown and coloured according to its pharmacophoric properties. (Grey: hydrophobic, yellow: aromatic, red: hydrogen bond donor, blue: hydrogen bond acceptor, green: positively charged and magenta: negatively charged). Representative points for the surface are also shown.. Further details on the protein environment representation are provided in Chapter III.3.

It is remarkable from Figure 9 that the three different protein environments are quite different from each other, exhibiting diverse pharmacophoric properties. Although they are found in unrelated proteins, they all have in common that have been co-crystallized with the same chemical fragment. Accordingly, they are linked by a chemoisosteric relationship. As seen in section I.2, a common graphical representation that is used to describe interactions between protein environments and chemical fragments is commonly known as heat maps. In such representation, the values contained in a matrix are colour-coded; with chemical fragments as rows and protein environments as columns, a cell in i, j position is coloured if the ith chemical fragment is compatible with the jth protein environment. Interpreting a single column, all bioisosteric chemical fragments compatible with a particular protein environment are obtained. In the same way, all chemical fragments in a row are chemoisosteric, as they are compatible with the same chemical fragment (Figure 10).

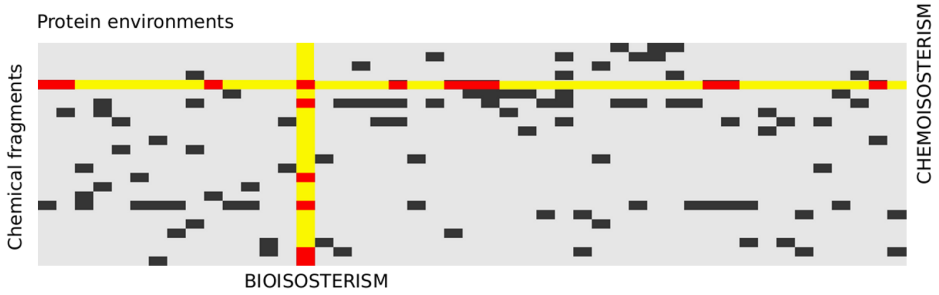


Figure 10: Sample heat map to illustrate both bioisosterism and chemoisosterism concepts. The latest corresponds to heat map rows, while the former to heat map columns.

According to the definition of chemoisosterism, two protein environments are chemoisosteric if compatible with the same chemical fragment. It is worth to note here that compatibility does not necessarily imply optimality. The fact

that a chemical fragment is compatible with a particular protein environment does not imply that it is able to bind there with high affinity (although it could well be), but that it able to be there under certain circumstances. One obvious source of chemoisosteric protein environments is the Protein Data Bank[54], where environments co-crystallized with a particular chemical fragment can be easily extracted. As it is often assumed that bioisosteric fragments do not need to be similar to interact with the same protein environments and that similar chemical fragments will likely bind to similar protein environments, we can also assume that similar protein environments will bind the same chemical fragments. This makes of binding site similarities another source for chemoisosteric protein environments. The definition and exploration of chemoisosterism through binding site similarity is one of the milestones of this Thesis and will be further extended in the following chapters, especially in Chapter III.3.

Part II: Objectives

The main objectives pursed by this Thesis can be summarised as follows:

i) To design and implement a novel structure-based methodology to extract, describe, store and compare protein binding sites based on their surface physicochemical properties.

ii) To apply the new methodology in *de novo* design and fragment-based drug discovery

iii) To explore the implication of chemoisosterism for drug polypharmacology

The achievement of the first has been achieved in Chapter III.1, where a new methodology for binding site description and comparison in described in detail. The expertise accumulated during this process has been materialized in the elaboration of a review on the topic in Chapter III.2. In Chapter III.3 the methodology is tailored by a fragment-based approach, resulting in the coinage of the term "chemoisosterism" and in some validation examples of its potential utility for drug discovery, achieving the second objective. Finally a review putting binding site similarities in the global context of drug polypharmacology takes advantage of all the knowledge gained during this Thesis in Chapter III.4.

Part III: Results

III.1: Family-wide pharmacophore signatures of protein binding sites.

Family-wide pharmacophore signatures of protein binding sites

Xavier Jalencas and Jordi Mestres

Introduction

The amount of protein structures available in public databases is expanding at an exponential rate. This implies a significant increase both in the number of structure entries for the same protein as well as of first entries for what had been structurally-orphan proteins. In spite of this, efforts to identify and further exploit the essential pharmacophore features exposed by protein cavities accessible for ligand and protein interactions are still limited.

This contribution introduces a methodology to detect, describe, store, and compare protein binding sites as a means to identify pharmacophore signatures of binding sites among all members of protein families for which structures have been determined and made publicly available. The program suite was developed in Python language (version 2.6, compatible with all 2.x versions). Most time-consuming algorithms were implemented in a C library to improve their efficiency. Libraries from BioPython[1] and Pybel[2] were also included into the pipeline when indicated. Specific PyMol[3] functions were devised to visualize all the results produced.

Binding site detection

As protein-ligand interactions occur in the binding site, the first step is to identify such binding sites in a protein. To this purpose the LIGSITE algorithm was implemented.[4] This algorithm, a modification of the original POCKET[5], relies on the fact that most ligand binding usually takes place in the largest cleft on the protein surface, especially in enzymes, or in internal cavities.[6] Using a purely geometric approach, the protein is embedded in a regular Cartesian grid, being all grid points occupied by the solvent accessible surface of the protein identified and leaving the rest of grid points labelled as solvent. Each of the solvent grid points is scanned in seven directions (x, y, z axis and the four cubic diagonals of the grid); if a clash with the protein surface is found when scanning in both sides along a particular direction, the score of the solvent grid point is increased by one. This way, each of the solvent grid points gets a score ranging from 0 to 7 (as seven directions are evaluated), which provides an indication of how buried the grid point it is within the protein surface (Figure 1a). Finally, all the grid points scoring above a predefined threshold are grouped, being each group considered a pocket in the protein surface. A threshold for the minimum number of grid points that define a pocket is needed to discard small pockets being too small to accommodate a single chemical fragment of a given size. Lowering the threshold to which a grid point is considered a pocket allows for detecting deeper, or even internal, cavities where crystallographic waters may often be present. After a thorough validation process, it was found that a threshold of 5 was appropriate in most of the cases. Figure 1b shows the selected grid points predicted to be part of the binding site in a thrombin structure (pdb code 1vzq). Grid points represented as spheres are coloured according to how buried they are inside the protein cavity. The co-crystallized ligand is also shown in this figure, although it was not considered for binding site detection. As can be observed, in this case the predicted

binding site fits very well with the volume occupied by the ligand. An improved version of this algorithm called LIGSITEcs, using Connolly surfaces instead of solvent accessible surfaces,[7] was also downloaded and tested.[8]

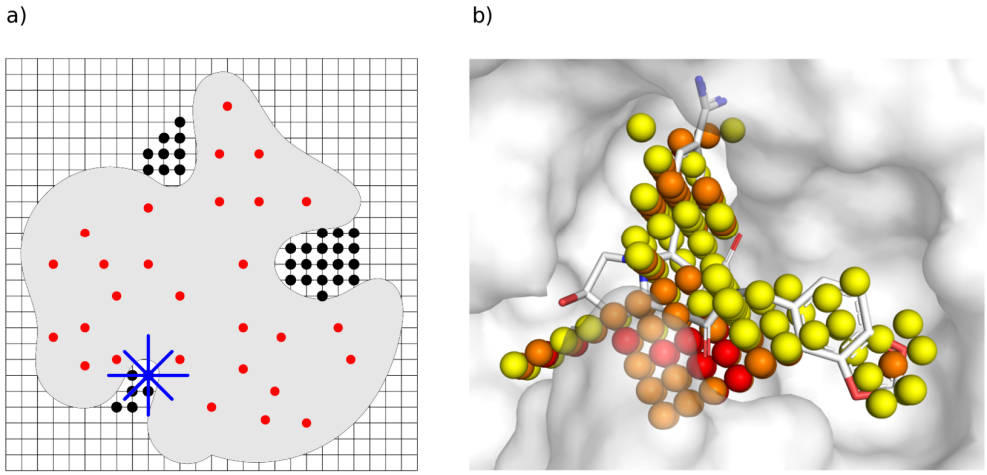


Figure 1. a) Schematic illustration of the LIGSITE algorithm. The grid points not in the protein volume (gray coloured) are scanned along 7 axes (blue) and those with higher score are considered part of a pocket (black dots). Figure adapted from Huang and Schroeder.[8] **b)** LIGSITE outcome for a thrombin structure (pdb 1vzq). Selected grid points are coloured according to their score (yellow, orange, and red for scores of 5, 6, and 7, respectively), that is related to its depth of burial within the protein cavity.

Besides binding site prediction, there are several situations where defining the volume of a binding site based on the geometric position of a ligand can be appropriate, being the most obvious of the cases when a ligand is co-crystallized with the protein of interest. This possibility allows also for using external ligands, placed in the binding site by other methods such as docking or structural alignments, or even using consensus ligands, defined by an ensemble of ligands co-crystallized in multiple structures of the same protein. An example of the binding site volume detected in a thrombin structure by each of the available implemented is shown in Figure 2.

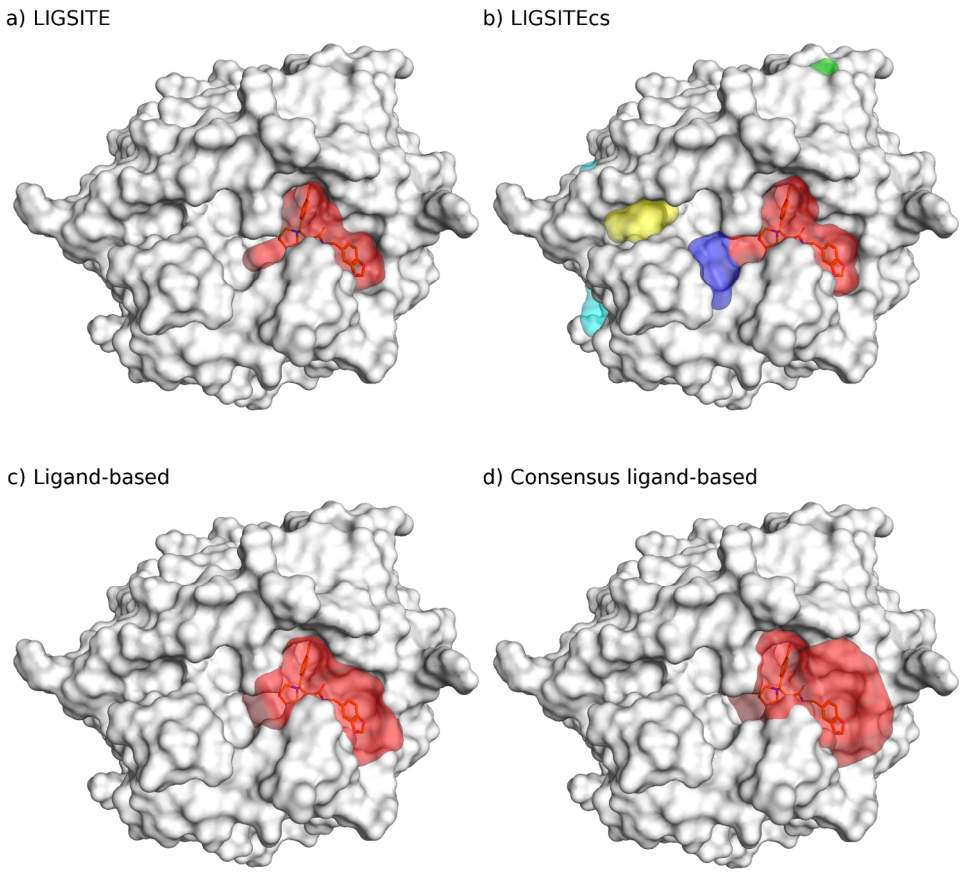


Figure 2: Predicted pocket volume shown in the surface of a thrombin protein obtained by different methods (pdb 1vzq).

Once the binding site location is detected and its extent characterized, the next step is to define its corresponding surface. To this aim, a triangulated surface of the protein is obtained using SMART.[9] At this stage, two different strategies can be followed. The simplest one consists on selecting all the vertices whose distance to any of the grid points that define the binding site volume is below a pre-defined threshold that is set to 1Å, which is equivalent to a distance of 2.5 Å from the surface to a hypothetical ligand atom. This leads to the selection of only the closest surface to the binding site volume, which does

not need to be complete, but is usually patched (Figure 3a). In most cases, in order to determine the volume of the cavity, it is more convenient to select all the surface vertices that belong to the binding site, despite the possible imperfections linked to this action. To this purpose, the distance threshold can be increased to roughly select all surface vertices, irrespectively of selecting surface vertices clearly out of the binding site. Then, all selected vertices neighbouring an unselected one are identified as surface borders and grouped by neighbourhood, in such a way that each continuous border encloses a defined region of the protein surface. At the same time, a single surface vertex is assumed to lie outside the binding site (the one with the largest distance to the binding site volume). Finally each of the borders is iteratively filled with neighbouring vertices until the region is complete or the external point is reached, in which case the section is discarded. This procedure leads to the selection of a continuous surface that can potentially cover a slightly wider surface than the original detection of the binding site. Figure 3b shows the selected vertices for a cavity in a structure of thrombin (pdb 1vzq), highlighting in red two borders involving vertices that were ultimately selected. One of them (the smaller) is completed while the biggest one is not, as it would lead to the selection of the surface vertex that is assumed to lie outside the binding site, and thus constituting the mouth of binding site.

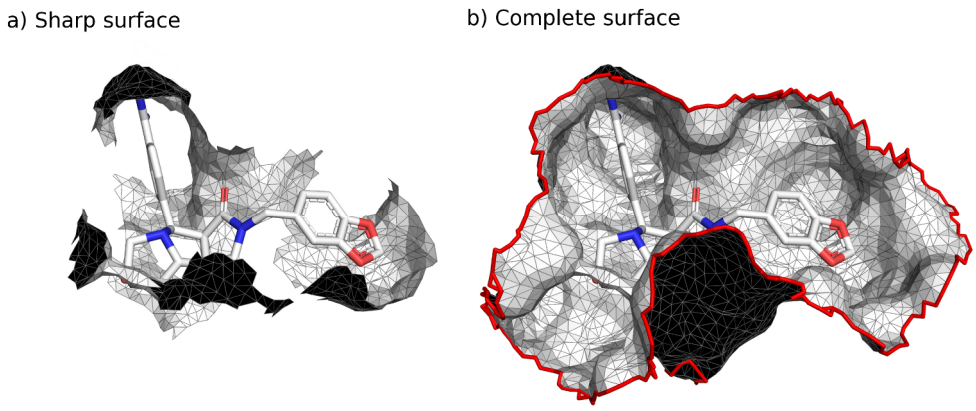


Figure 3: Surface selection modes, complete and sharp in thrombin structure (pdb 1vzq)

Pharmacophore description of binding surfaces

Given the large number of surface vertices that compose each binding site, a reduced description of its surface is required that is able, nonetheless, to capture the essential shape, size, and chemical properties of the protein cavity. For this purpose, a similar approach to Cavbase[10,11] and SiteEngines[12,13] was adopted. Four pharmacophoric features relevant to protein-ligand binding are used, namely, hydrophobic (H), aromatic (R), hydrogen-bond acceptor (A) and hydrogen-bond donor (D). Optionally, features such as positively charged (P) and negatively charged (N) can be also included. Figures 4 and 5 illustrate the entire process. In figure 4a, a thrombin structure (pdb 1vzq) is shown along with the defined binding site surface. The first step consists on assigning those features to the protein atoms underlying the surface of the cavity. This is done using a predefined table that assigns features to atoms in each of the 20 standard amino acids (Appendix B). These features are subsequently transferred from the protein atom to their corresponding surface vertices. The labelled

surface vertices are grouped into surface patches, being each patch defined as a set of connected vertices assigned to the same pharmacophoric property. In this step, some rules are incorporated to capture the directionality of the main forces driving protein-ligand binding. For hydrophobic surface patches, all connected hydrophobic surface vertices assigned to the same heavy atom are grouped into a single surface patch. Likewise, all aromatic surface vertices assigned to the same aromatic group of atoms generate an aromatic surface patch but, in this case, only the surface vertices for which the angle between the normal vector of the aromatic group and the vector from the vertex to its corresponding atom is below 60º were considered. This type of filtering ensures that only the surface vertices at both sides of an aromatic group generate aromatic patches. A similar filter is applied to describe the directionality of hydrogen bonding. Only surface vertices in the direction of a theoretical hydrogen bond interaction (allowing for a deviation of 30º) are considered when defining the surface regions assigned to hydrogen-bond acceptors and donors. Figure 4b shows a patched surface coloured according to their features (gray: hydrophobic, orange: aromatic, red: hydrogen bond acceptor and blue: hydrogen bond donor). Each of the surface patches is then condensed into a sole surface feature point (a pharmacophore centroid), generating a surface feature point of the same pharmacophoric type at the position of the surface vertex closest to the centre of mass from all vertices defining a given surface patch (Figure 5a).

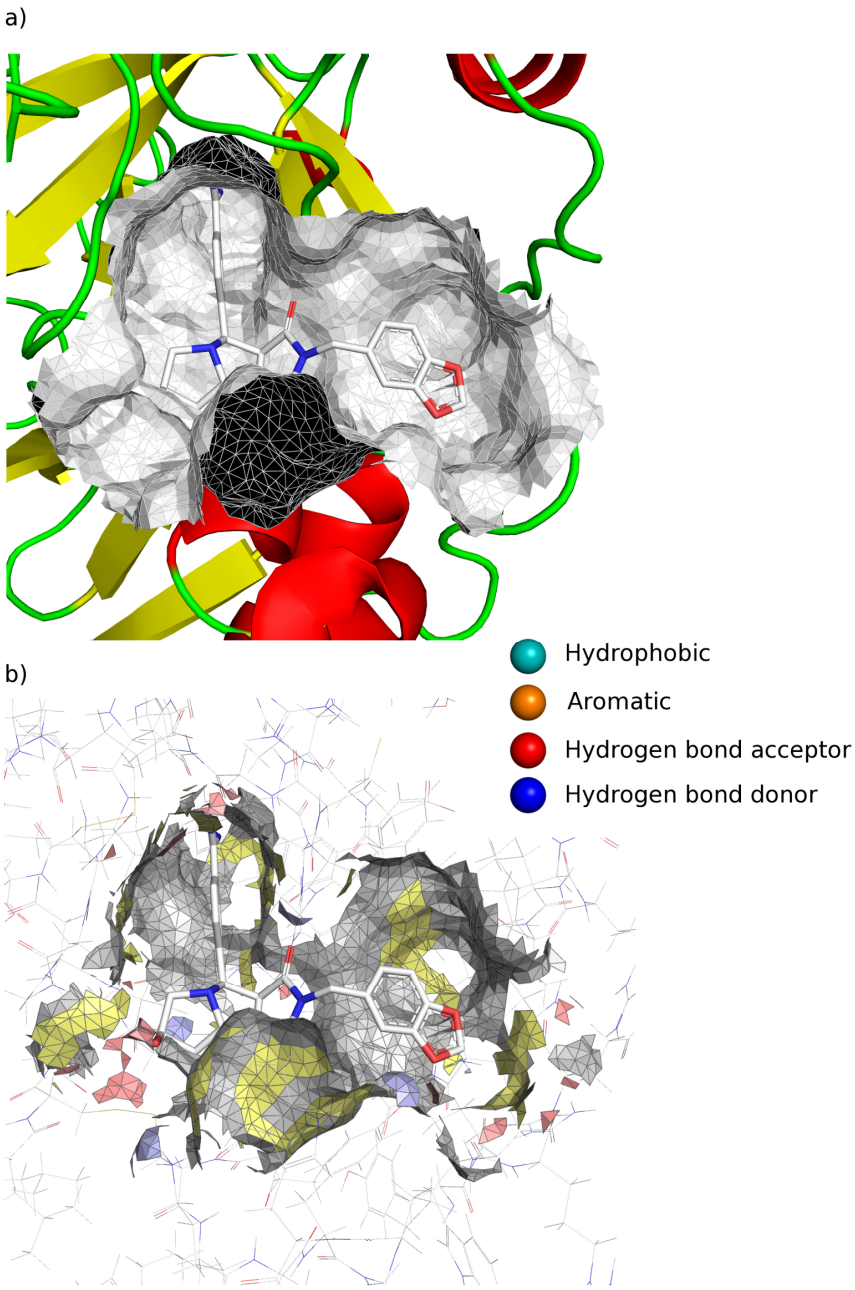


Figure 4: Binding site description of a thrombin structure (pdb 1vzq): **a)** Definition of the binding site surface and **b)** is decomposition into surface patches of pharmacophoric features

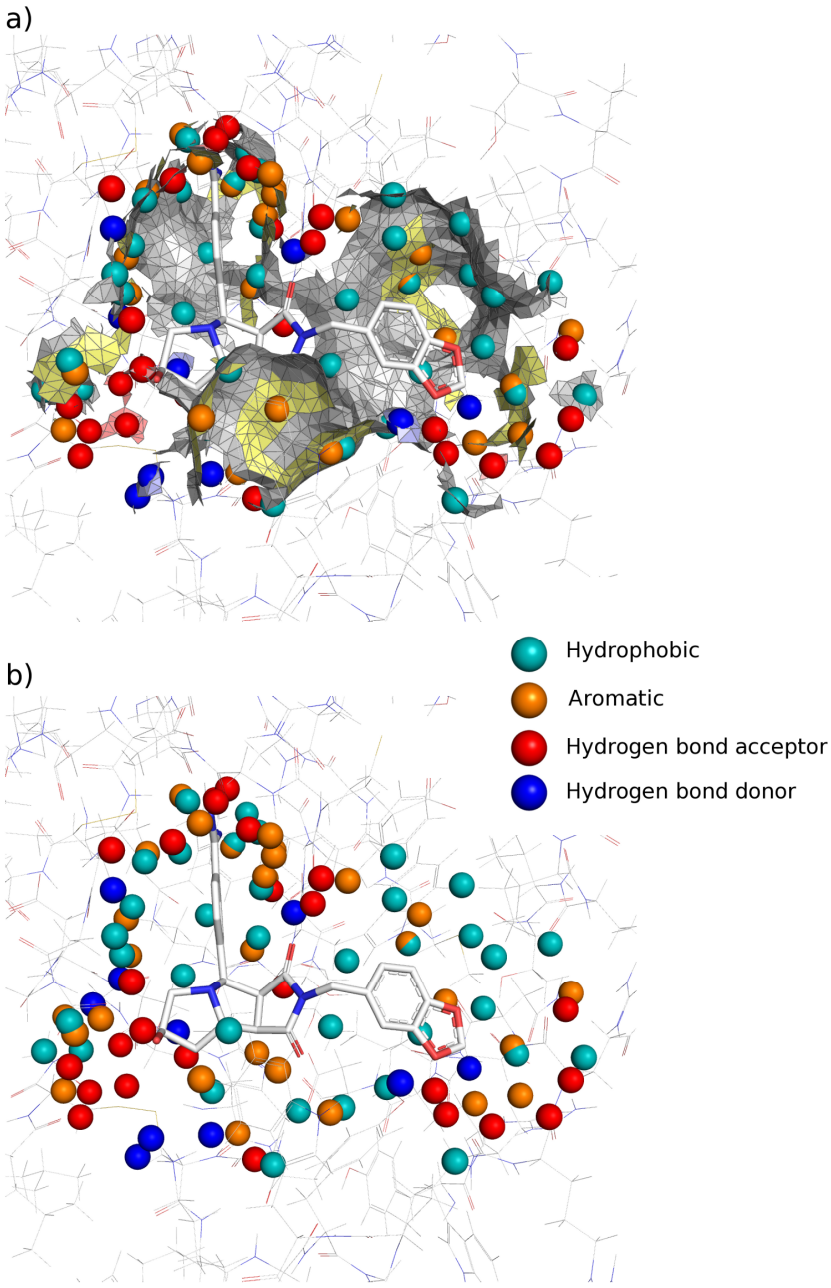


Figure 5: Binding site description: **a)** Generation of pharmacophore centroids representing surface patches and **b)** storage of their feature type and position as a simplified description of the binding site.

The generated surface pharmacophore centroids (their features and geometric position) constitute a simplified representation of the binding site that aims at retaining only its key features relevant for ligand binding. Those ensembles of points are stored to be subsequently used for binding site comparison. Figure 6 shows in detail some of the feature surface points generated by some particular amino acids of the protein. In Figure 6a, the hydrophobic points for the side chains of an isoleucine residue are depicted. In Figure 6b, a tyrosine produces hydrophobic and aromatic points above and below the phenyl ring plane and hydrogen-bond acceptor and donor points for its hydroxyl group. Figure 6c shows the pharmacophoric centroids that result from the surface features exposed by a glutamic acid. The carboxylate group is represented by hydrogen bond acceptor points as well for aromatic and hydrophobic ones. Some additional hydrophobic points are provided by the aliphatic part of the side chain. In Figure 6d a lysine is shown with its corresponding hydrogen-bond donor feature points. Finally, Figure 6e shows the surface feature points that are generated by a peptidic bond close to the protein surface.

Due to the system used to assign hydrogen-bond acceptor and donor feature points, it is relevant to note that the protonation state, as well the actual position of the hydrogen atoms, will determine their existence and position. As X-ray crystallography is unable in most cases to provide enough resolution to assign hydrogen positions, the GROMACS[14] suite is used to assign those protonation states.

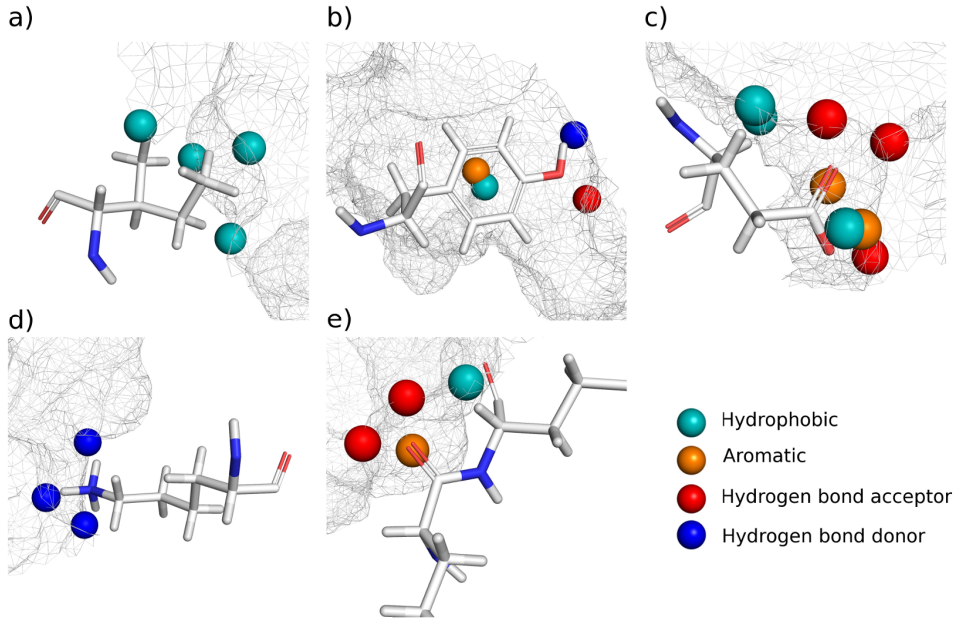


Figure 6: Details of the pharmacophore centroids generated by some protein residues on the surface of the binding site of a thrombin structure (pdb 1vzq).

Binding site comparison

Once the binding sites are encoded in a set of labelled pharmacophore centroids in space, one cap perform pair-wise comparisons using a technique commonly referred to as clique detection.[15–17] Basically, it consists on considering binding sites as graphs and apply a maximum common subgraph isomorphism solving algorithm to find an optimal match between a pair of binding site graphs. A maximum common subgraph isomorphism problem for two graphs G_1 and G_2 consists on finding the largest subgraph of G_1 that is isomorphic to a subgraph of G_2. One possible solution to this problem, and the one that is selected in this work, is to build a modular product graph, in which the largest clique is taken as the representative solution to the problem.

Figure 7 illustrates the matching process, which starts by the representation of the two binding sites to be compared as graphs (Figure 7a). Each surface feature point is assigned to a node that is coloured according to its pharmacophoric property. The nodes are completely connected with edges labelled with their distances in Armstrongs. To build the product graph (Figure 7b), all possible pairs of nodes from the initial graphs coloured with the same property produce a new node. Edges between the new nodes are assigned if edges between the nodes in the original graphs have an equivalent label. In the case presented in Figure 7b, nodes Aa and Bb are linked by an edge because the distances between AB and ab in the original graphs are equivalent. Two distances are considered equivalent if their difference is below 1Å. This threshold allows for including a certain degree of fuzziness to the comparison, otherwise only perfect matches would be successfully detected. This is absolutely necessary in surface-based binding site comparisons, as even the surfaces of the binding sites from the same protein may suffer variations in different structures, native or co-crystallised with a variety of ligands.

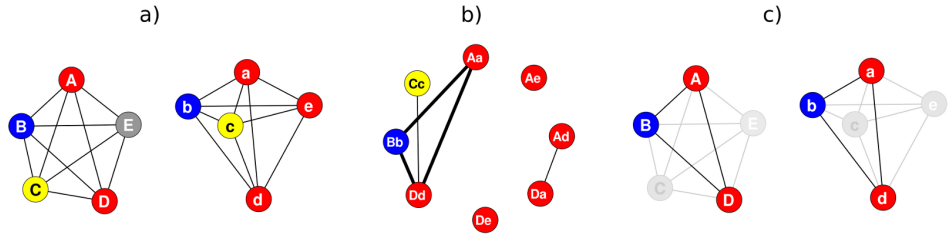


Figure 7: Scheme illustrating the procedure used to locate similarities between binding sites using a maximum common subgraph isomorphism approach. It comprises the following steps of: **a)** binding sites are described as graphs with coloured nodes and labelled edges; **b)** construction of a product graph by clique detection; and **c)** assignment of node equivalency between graphs.

Once the product graph is built, a clique detection algorithm[18] is used to find the maximal cliques, which represent the matches between the original graphs. A clique is defined as a completely connected subgraph and usually only the largest one is retained as the matching solution (Figure 7b). In this case, the largest clique corresponds to the nodes Aa, Bb and Dd, indicating the correspondences between the original graphs (Figure 7c). It is worth stressing here that clique detection is a NP-complete problem meaning that, although any given solution can be quickly verified, there is no known efficient way to locate a solution in the first place. This implies that the time required to solve the problem quickly increases as the size of the product graphs grows. In practice, this means that binding sites characterized by around 300 feature points start to be computationally too demanding to compare. Fortunately, the number of pharmacophore centroids that describe most ligand binding sites are below this size and thus, clique detection is applicable in most of the cases of interest for local binding site comparisons. This may not be always the case when comparing protein-protein interaction sites, which are much bigger in most cases.[6]

As discussed in the previous paragraph, the clique detection algorithm yields the best matching between the feature points of the two binding sites being compared. One of the advantages of this method over other faster methods such as fingerprint matching[19], is that a three-dimensional superimposition of the binding sites can be subsequently obtained from the clique matching of surface feature points. A least square estimation of parameters between the matching feature points[20] results in the transformation matrix that produces the 3D alignment between two binding sites. However, at this stage, matches between feature points from a concave area and a convex one are possible, which would produce unrealistic alignments. To avoid accepting those matches as solutions, an angle threshold is introduced after the superposition is performed. Each pair of matched surface feature points is evaluated regarding

the angle between the vectors from the point to each of the actual protein atoms that produce the feature point. If the angle is below a given threshold (set to 60º in this case), the match is considered permissible, otherwise it is discarded and the size of the match reduced. The largest 100 cliques obtained are hereby evaluated, and the biggest one after discounting illicit matches is ultimately kept.

The last step in the comparison is the assignment of a scoring. A cosine score is adopted, which takes into account the size of the match as well the sizes of both compared binding sites. The similarity (S) is obtained as a function of the number of matched feature points (c), and the number of feature points of the two binding sites that were compared (a,b) according to the following formula: $S = c / (a \times b)^{1/2}$

Retrospective validations

In order to validate the performance of our methodology, a comparative assessment of the results obtained against those published by other methods, such as Cavbase[17], in several example cases is presented.

A dataset of 113 binding cavities extracted from X-ray structures of proteins belonging to 13 diverse functional enzyme families was compared and clustered. All entries from an enzyme family belong to the same SCOP subfamily and thus, have similar sequences and folds. Cavities were extracted using LIGSITEcs and those located in the catalytic domain were manually selected. The clustering was performed using the CLUTO Toolkit[21], with the partitional rb algorithm and default parameters. The number of 16 predefined output clusters produced the best clustering, as some cavities are divided into subpockets by the LIGSITEcs algorithm. As shown in Figure 8, a good

clustering was obtained. Only one cavity was clustered with proteins of other enzyme families, this misclassification being a consequence of a much bigger binding site detected. The obtained clustering was very similar to the one reported using Cavbase, demonstrating that our approach can successfully classify binding sites from different protein families. Although this is not our ultimate goal, it is a necessary and important achievement that requires proper validation.

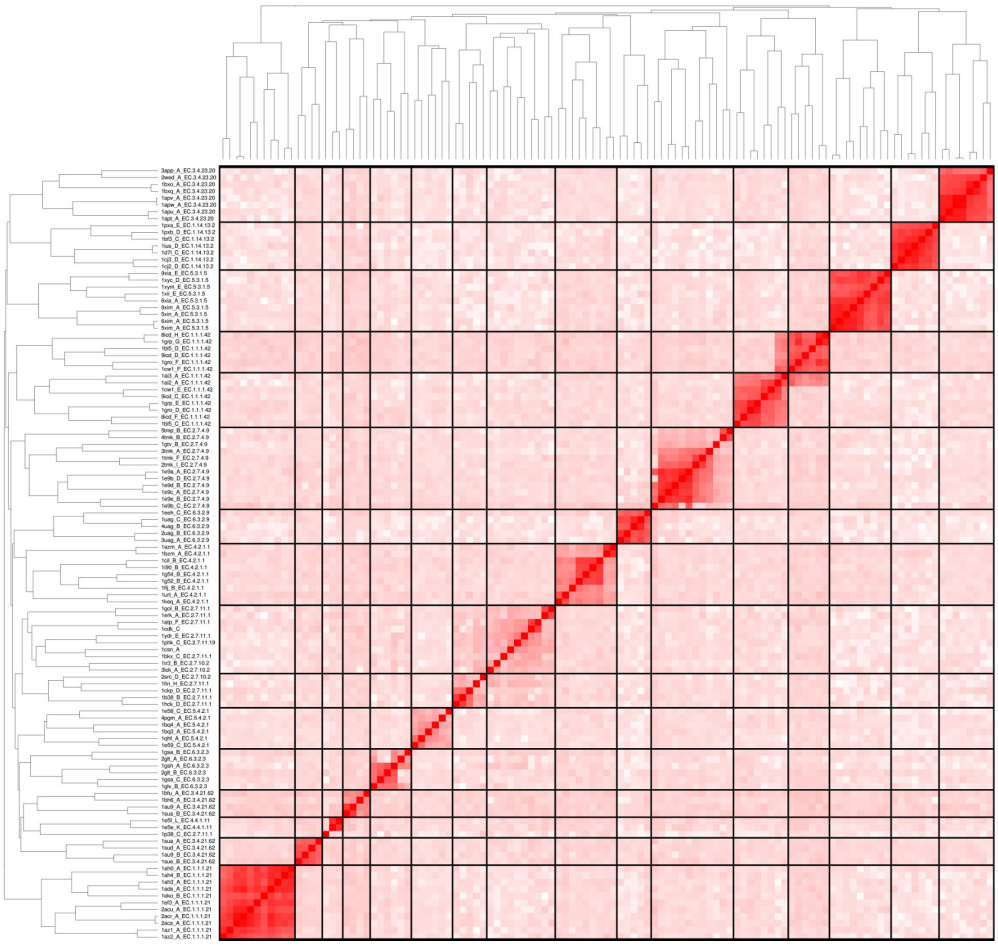


Figure 8: Example of binding site similarity clustering for 113 cavities from 13 diverse functional enzyme families.[17] A good classification according to the enzyme family is achieved.

Continuing on the comparison against Cavbase, a second dataset of 24 α-carbonic anhydrase binding sites, belonging to 6 isozyme classes, was extracted from the PDB. The same procedure detailed above was followed and the results revealed that a convincing clustering was obtained (Figure 9), consistent with isoform classification, and eve performing a bit better than Cavbase in some cases.

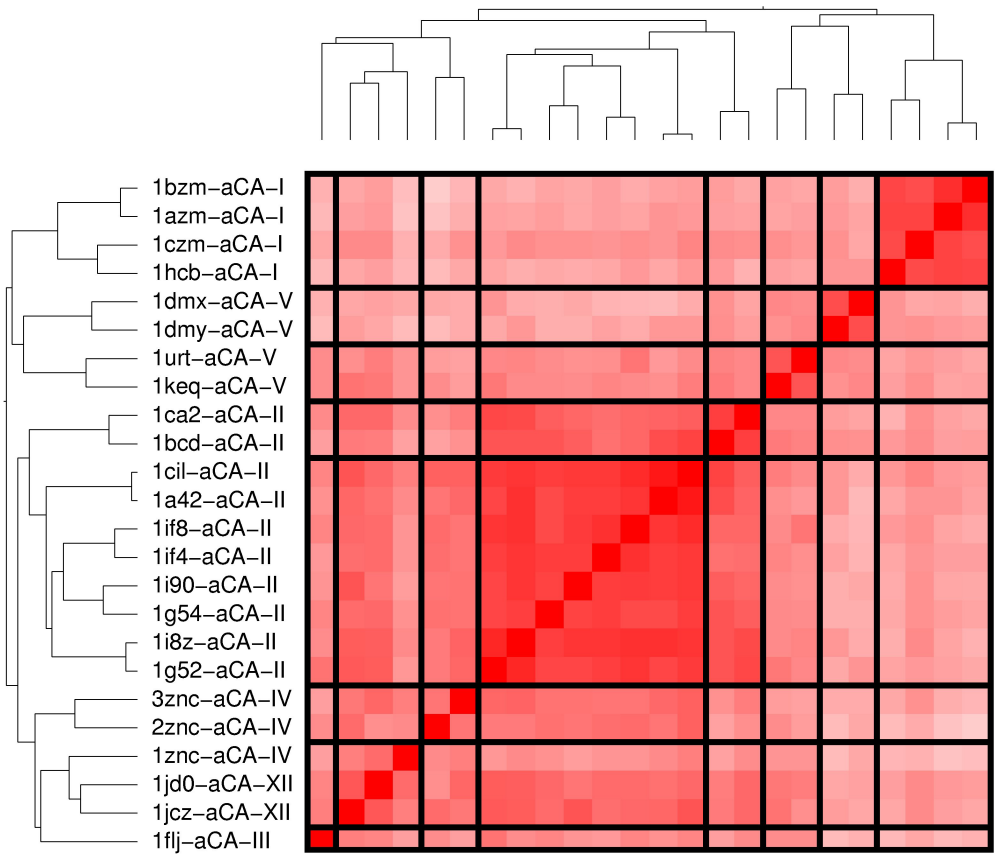


Figure 9: Example of binding site similarity clustering for α-carbonic anhydrase binding sites.[17] A convincing classification according enzyme isoforms is obtained.

As a third validation experiment, a dataset assembled by Kuhn et al.[17] to evaluate the performance of Cavbase, but used also as benchmark for other methods,[22] was tested. It consists on a set of protein kinases and thus, the challenge consists on evaluating the ability of a given method to differentiate between closely related binding sites. It contains five p38 MAP kinases, seven Cell Division Protein Kinase 2, five Glycogen Synthase Kinase-3 beta and five LCK Tyrosine kinases.

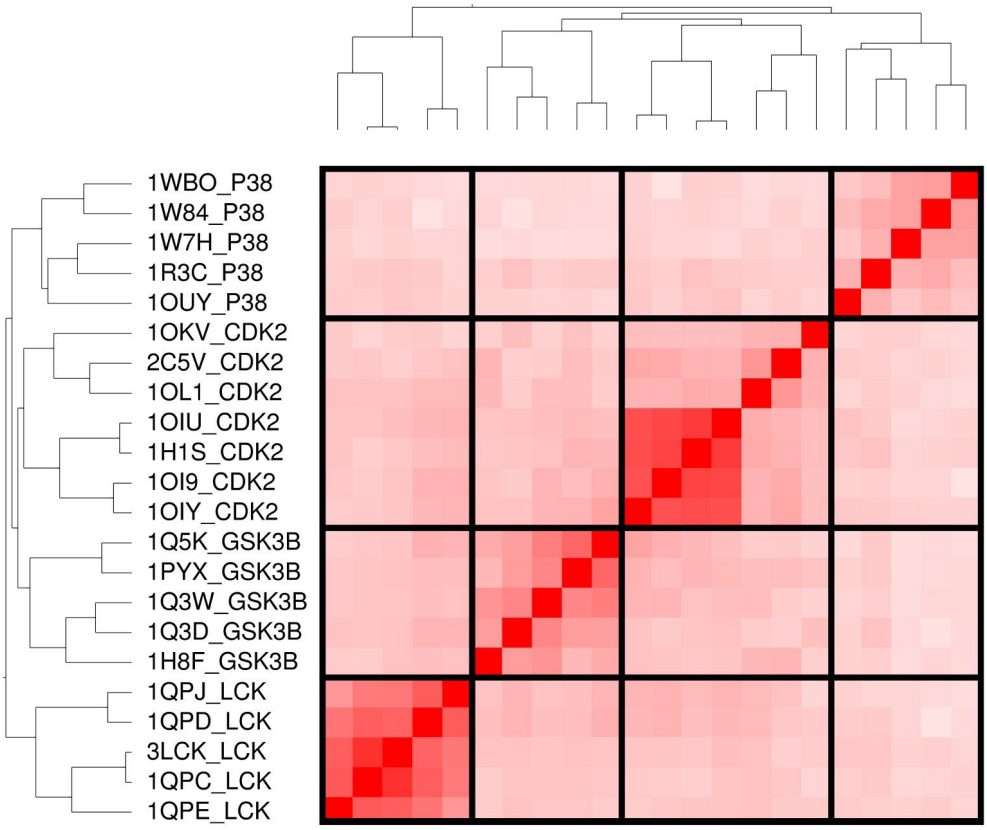


Figure 10: Clustering of 22 ATP-binding sites from 4 protein kinase subfamilies.[23]

As observed in Figure 10, the similarities obtained are able to perfectly organize the different kinase binding sites. Finally, a rather small set of eight binding site pairs that was used to evaluate the performance of different

methods in several publications[22–24] was also tested. It consist of pairs of binding sites that, although they bind to the same (or very similar) ligands, the similarity is not evident and thus, they are considered difficult cases. Results are shown in Table 1.

Table 1: Results from different comparison methods to detect significant similarities between difficult pairs of binding sites.[23]

PDB1	PDB2	Sites Base[25]	SuMo[26]	Site Engine[12]	Pocket Match[24]	BS Align[27]	Site Align[28]	FuzCav[23]	ProBis[29]	TrixP[22]	
1GJC	1V2Q	**29.38**	**89**	**32.56**	**50.17**	**31.77**	**0.03**	**0.19**	**3.74**	0.18	**0.35**
	2AYW	NA	**40**	x	**52.29**	**31.51**	**0.02**	**0.18**	**3.75**	0.27	**0.31**
	1O3P	**54.80**	NA	**38.48**	**88.01**	**42.26**	**0.01**	**0.18**	**4.95**	**0.65**	**0.53**
1ECM	4CSM	**54.65**	x	x	**55.56**	x	x	**0.18**	**1.65**	0.16	**0.35**
1V07	1HBI	**46.81**	x	x	**61.42**	x	**0.20**	**0.18**	**6.04**	**0.43**	0.19
1M6Z	1LGA	x	x	x	**63.85**	x	x	x	0.58	0.24	0.11
1ZID	2CIG	NA	NA	x	**56.01**	x	x	x	0.29	0.19	0.10
6COX	1OQ5	x	x	**33.14**	x	x	x	**0.16**	0.67	x	0.16

The first two columns indicate the pdb codes for the binding sites that are compared. The following columns contain the particular scores obtained from the different comparison methods, the last one (shaded in gray) corresponding to those obtained with the method described here. Highlighted in bold face are the similarity values that are considered convincingly significant according to the characteristics of each of the methodologies. The table is divided in two sections, namely, a first block containing pairs of proteins sharing structural fold, according to SCOP[30] or CATH[31], and a second one containing pairs of *a priori* completely unrelated proteins. It is clear from the table that successfully locating similarities between binding sites belonging to proteins sharing the same fold is a relatively easy task that most methods can achieve, including the

one described here (despite failing in successfully identifying as similar the binding sites from 1V07 and 1HBI). Regarding unrelated binding sites (second block of the table), the scene changes radically. In this case, only PocketMatch is able to find similarities in two cases. The pair of binding sites 6COX-1OQ5, corresponding to a cyclooxigenase-2 and a carbonic anhydrase, deserves special consideration. It was experimentally shown that both binding sites can bind to celecoxib, a polypharmacology event that Cavbase could attribute to a partial match between their binding sites.[32] Remarkably, most of the binding site comparison methods included in Table 1 fail to locate this similarity. It seems clear that detecting binding site similarities in unrelated proteins is a more challenging task that the devised methodology, as many others, fails to achieve, at least in the few examples included in Table 1.

Pharmacophore signatures of binding sites

Extending the pair-wise comparisons described in the preceding sections to multiple comparisons of a set of binding sites can also be done. The procedure is as follows: A single binding site is chosen as a template and all other binding sites are compared and superimposed to it. Independently of their origin, all feature points are agglomeratively clustered by their position and property. This strategy was adopted over clustering based on the results of the feature point matching obtained during comparison validations, as it revealed to produce consistent results irrespectively of the binding site that was chosen as template. The clustering itself produces a multiple alignment, a sample part of which is shown in Figure 11.

	161	162	163	164	165	166	167	168	169
	A	A	A	D	D	D	D	D	D
1VZQ	SER'195	GLY'216	-	SER'195	SER'195	GLY'216	GLY'219	-	-
1T4V	SER'235	GLY'258	-	-	SER'235	GLY'258	GLY'260	PHE'269	-
2UUF	-	GLY'216	-	-	SER'195	GLY'216	GLY'219	PHE'227	HIS'57
1VIT	-	GLY'216	-	-	SER'195	GLY'216	GLY'219	PHE'227	-
1DWD	-	GLY'237	-	-	SER'214	GLY'237	GLY'239	PHE'248	-
2FEQ	SER'195	GLY'216	-	-	SER'195	GLY'216	GLY'219	PHE'227	-
2BXT	-	GLY'216	-	-	SER'195	GLY'216	GLY'219	PHE'227	-
1SB1	-	GLY'216	-	-	SER'195	GLY'216	GLY'219	PHE'227	HIS'57
2ANK	-	GLY'216	-	-	SER'195	GLY'216	GLY'219	PHE'227	-
1ETS	-	-	ARG'221	SER'195	SER'195	GLY'216	GLY'219	PHE'227	HIS'57

Figure 11: Sample fragment of a residue multiple alignment based on feature surface points matches for 10 thrombin binding sites.

Each position of the sequence alignment shown in Figure 11 corresponds to a cluster, and feature points of each binding site assigned to the cluster (if any) are shown in the column caption. It is easily seen from such alignments that some surface feature points are more conserved than others. Unfortunately, as the number of binding sites and their sized grow, such alignments can be difficult to interpret. To alleviate this limitation, appropriate visualization of the binding site features is required. Figure 12a shows a visual representation for the alignment of 25 thrombin binding sites. Each cluster is represented by a single point, coloured according to its property, which is located at the average position of all feature points forming the cluster. The size of the point is relative to the number of structures that have a feature point in that cluster, the largest ones corresponding to those feature points common to all binding sites under evaluation. A line joining each of the points to all members of the cluster is also shown. This provides information on the positional variability of the feature points in each cluster. A large point with short lines indicates a highly

conserved feature point, both in terms of presence and position, with high probabilities of being relevant for ligand binding.

Retaining only those positions in the alignment (clusters) that contain members of most binding sites results in what will be referred to as the binding site signature (Figure 12b). This signature is a simplified representation of an ensemble of binding sites (25 thrombin binding sites in this case), that highlights the most conserved features in the binding site. Additionally, it can be treated as a regular binding site feature point representation, so all previously described methodologies can be applied to compare signatures against binding sites or against other signatures, with the advantage that the signatures are smaller than binding sites and thus faster and easier to compare.

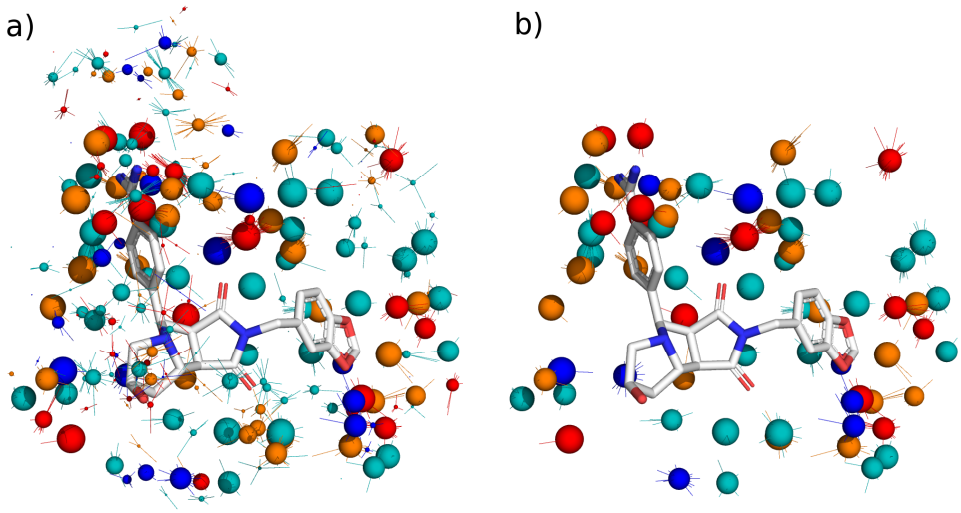


Figure 12: a) Thrombin binding site signature derived from 25 x-ray thrombin structures (1vzq used as template). b) Only surface feature points present in at least a 75% of the structures are shown.

Conclusions

A new methodology to perform binding site comparisons based on surface feature points has been devised and implemented. Its ability to discern between binding sites of both different and closely related proteins has been shown. Also, its performance when detecting similarities between protein binding sites was comparable to that of other existing methods. The methodology was further extended to extract those feature centroids in binding cavities most conserved among multiple entries of the same protein or among entries for different proteins of the same family. Full exploitation of these binding site signatures, both for proteins and protein families, is currently underway in our research group.

References

(1) Cock, P. J. A.; Antao, T.; Chang, J. T.; Chapman, B. A.; Cox, C. J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; de Hoon, M. J. L. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **2009**, *25*, 1422–1423.

(2) O'Boyle, N.; Morley, C.; Hutchison, G. Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit. *Chem. Cent. J.* **2008**, *2*, 5.

(3) *The PyMOL Molecular Graphics System*; Schrödinger, LLC.

(4) Hendlich, M.; Rippmann, F.; Barnickel, G. LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *J. Mol. Graph. Model.* **1997**, *15*, 359–63, 389.

(5) Levitt, D. G.; Banaszak, L. J. POCKET: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. *J. Mol. Graph.* **1992**, *10*, 229–34.

(6) Laskowski, R. A.; Luscombe, N. M.; Swindells, M. B.; Thornton, J. M. Protein clefts in molecular recognition and function. *Protein Sci.* **1996**, *5*, 2438–52.

(7) Edelsbrunner, H.; Facello, M.; Fu, P.; Liang, J. Measuring proteins and voids in proteins. In *Proceedings of the 28th Hawaii International Conference on System Sciences*; IEEE Computer Society, 1995; p. 256.

(8) Huang, B.; Schroeder, M. LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC Struct. Biol.* **2006**, *6*, 19.

(9) Zauhar, R. J. SMART: a solvent-accessible triangulated surface generator for molecular graphics and boundary element applications. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 149–159.

(10) Schmitt, S.; Kuhn, D.; Klebe, G. A new method to detect related function among proteins independent of sequence and fold homology. *J. Mol. Biol.* **2002**, *323*, 387–406.

(11) Kalliokoski, T.; Vulpetti, A. Large-Scale Evaluation of CavBase for Analyzing the Polypharmacology of Kinase Inhibitors. *Mol. Inf.* **2011**, *30*, 923–925.

(12) Shulman-Peleg, A.; Nussinov, R.; Wolfson, H. J. SiteEngines: recognition and comparison of binding sites and protein-protein interfaces. *Nucleic Acids Res.* **2005**, *33*, W337–341.

(13) Shatsky, M.; Shulman-Peleg, A.; Nussinov, R.; Wolfson, H. J. Recognition of Binding Patterns Common to a Set of Protein Structures. In *Research in Computational Molecular Biology*; 2005; pp. 440–455.

(14) Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. GROMACS: fast, flexible, and free. *J. Comput. Chem.* **2005**, *26*, 1701–1718.

(15) Gardiner, E. J.; Artymiuk, P. J.; Willett, P. Clique-detection algorithms for matching three-dimensional molecular structures. *J. Mol. Graph. Model.* **1997**, *15*, 245–253.

(16) Kinoshita, K.; Furui, J.; Nakamura, H. Identification of protein functions from a molecular surface database, eF-site. *J. Struct. Funct. Genomics* **2002**, *2*, 9–22.

(17) Kuhn, D.; Weskamp, N.; Schmitt, S.; Hüllermeier, E.; Klebe, G. From the similarity analysis of protein cavities to the functional classification of protein families using cavbase. *J. Mol. Biol.* **2006**, *359*, 1023–44.

(18) Niskanen, S.; Östergård, P. R. J. *Cliquer User's Guide, Version 1.0*; Tech. Rep. T48; Communications Laboratory, Helsinki University of Technology: Espoo, Finland, 2003.

(19) Mason, J. S.; Morize, I.; Menard, P. R.; Cheney, D. L.; Hulme, C.; Labaudiniere, R. F. New 4-point pharmacophore method for molecular similarity and diversity applications: overview of the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures. *J. Med. Chem.* **1999**, *42*, 3251–3264.

(20) Umeyama, S. Least-Squares Estimation of Transformation Parameters Between Two Point Patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **1991**, *13*, 376–380.

(21) Karypis, G. *CLUTO - Software for Clustering High-Dimensional Datasets*; University of Minnesota, Minneapolis, 2002.

(22) Von Behren, M. M.; Volkamer, A.; Henzler, A. M.; Schomburg, K. T.; Urbaczek, S.; Rarey, M. Fast Protein Binding Site Comparison via an Index-Based Screening Technology. *J. Chem. Inf. Model.* **2013**, *53*, 411–422.

(23) Weill, N.; Rognan, D. Alignment-Free Ultra-High-Throughput Comparison of Druggable Protein−Ligand Binding Sites. *J. Chem. Inf. Model.* **2010**, *50*, 123–135.

(24) Yeturu, K.; Chandra, N. PocketMatch: A new algorithm to compare binding sites in protein structures. *BMC Bioinf.* **2008**, *9*, 543.

(25) Gold, N. D.; Jackson, R. M. SitesBase: a database for structure-based protein-ligand binding site comparisons. *Nucleic Acids Res.* **2006**, *34*, D231–4.

(26) Jambon, M.; Imberty, A.; Deléage, G.; Geourjon, C. A new bioinformatic approach to detect common 3D sites in protein structures. *Proteins* **2003**, *52*, 137–145.

(27) Aung, Z.; Tong, J. C. BSAlign: a rapid graph-based algorithm for detecting ligand-binding sites in protein structures. *Genome Inform.* **2008**, *21*, 65–76.

(28) Schalon, C.; Surgand, J.-S.; Kellenberger, E.; Rognan, D. A simple and fuzzy method to align and compare druggable ligand-binding sites. *Proteins* **2008**, *71*, 1755–1778.

(29) Konc, J.; Janezic, D. ProBiS: a web server for detection of structurally similar protein binding sites. *Nucleic Acids Res.* **2010**, *38*, W436–440.

(30) Lo Conte, L.; Brenner, S. E.; Hubbard, T. J. P.; Chothia, C.; Murzin, A. G. SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res.* **2002**, *30*, 264–267.

(31) Orengo, C. A.; Michie, A. D.; Jones, S.; Jones, D. T.; Swindells, M. B.; Thornton, J. M. CATH--a hierarchic classification of protein domain structures. *Structure* **1997**, *5*, 1093–1108.

(32) Weber, A.; Casini, A.; Heine, A.; Kuhn, D.; Supuran, C. T.; Scozzafava, A.; Klebe, G. Unexpected nanomolar inhibition of carbonic anhydrase by COX-2-selective celecoxib: new pharmacological opportunities due to related binding site recognition. *J. Med. Chem.* **2004**, *47*, 550–557.

III.2: Identification of similar binding sites to detect distant polypharmacology

Jalencas, X.; Mestres, J., Identification of similar binding sites to detect distant polypharmacology. Mol. Inf. Submitted minf.201300082.

Journal Impact Factor: 2.39

Molecular Informatics

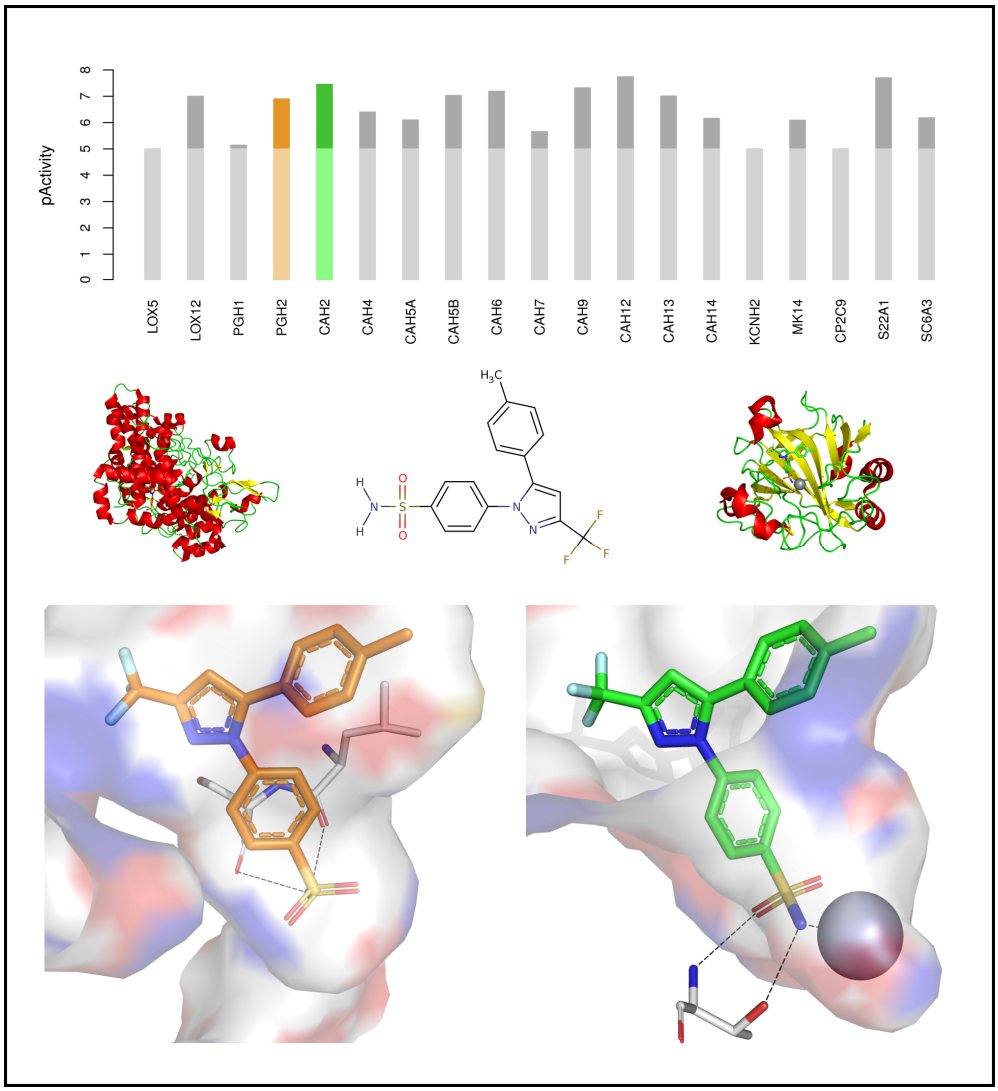
Identification of similar binding sites to detect distant polypharmacology

Journal:	*Molecular Informatics*
Manuscript ID:	minf.201300082
Wiley - Manuscript type:	Review
Date Submitted by the Author:	01-May-2013
Complete List of Authors:	Jalencas, Xavier; IMIM Hospital del Mar Research Institute, Research Programme on Biomedical Informatics Mestres, Jordi; Institut Municipal dInvestigacio Medica, Research Unit in Biomedical Informatics
Keywords:	Chemogenomics, Drug profiling, Protein structures

In order to develop a binding-site centred structure-based approach useful in drug discovery, as stated in the objectives of this Thesis, a necessary first step corresponds to acquire knowledge on the state of the art of the field. This knowledge, by the side of the gained expertise on the field during the Thesis, has been ultimately used for the preparation of a review on how binding site similarities can be exploited to identify unexpected protein-ligand interactions.

Identification of similar binding sites to detect distant polypharmacology

Xavier Jalencas and Jordi Mestres*



Corresponding Author:
* Email: jmestres@imim.es; tel: +34 93 3160540; fax: +34 93 3160550

Chemogenomics Laboratory, Research Programme on Biomedical Informatics (GRIB), IMIM Hospital del Mar Research Institute and University Pompeu Fabra, Parc de Recerca Biomèdica, Doctor Aiguader 88, 08003 Barcelona, Catalonia, Spain. E-mail: jmestres@imim.es; Fax: +34 93 3160550; Tel: +34 93 3160540.

Abstract: The ability of small molecules to interact with multiple proteins is referred to as polypharmacology. This property is often linked to the therapeutic action of drugs but it is known also to be responsible for many of their side effects. Because of its importance, the development of computational methods than can predict drug polypharmacology has become an important line of research that led recently to the identification of many novel targets for known drugs. Nowadays, the majority of these methods are based on measuring the similarity of a query molecule against the hundreds of thousands of molecules for which pharmacological data on thousands of proteins are available in public sources. However, similarity-based methods are inherently biased by the chemical coverage offered by the active molecules present in those public repositories, which limits significantly their capacity to predict interactions with proteins structurally and functionally unrelated to any of the already known targets for drugs. It is in this respect that structure-based methods aiming at identifying similar binding sites may offer an alternative complementary means to ligand-based methods for detecting distant polypharmacology. The different existing approaches to binding site detection, representation, comparison, and fragmentation are reviewed and recent successful applications presented.

Keywords: polypharmacology, target profiling, binding site alignment, fragment-based drug discovery, chemoisosterism

1.Introduction

The increasing availability in the public domain of pharmacological data for small molecules[1–9] has promoted the recent development of ligand-based computational approaches to predicting the interaction of molecules against thousands of protein targets based mainly on chemical similarity principles.[10–12] In the last five years, these methods have predicted and then confirmed experimentally a total of 249 new drug-target interactions, which represent an increase of almost 7% of all drug-target interactions currently known for those drugs.[13] However, a close inspection at the novelty of the new drug-target interactions identified (Figure 1) reveals that, on one hand, for the vast majority of them there was already an interaction to a target of the same protein family already known in the public domain and, on the other hand, most of the new targets identified for old drugs belong to the class of aminergic G protein-coupled receptors, a family known to have levels of cross-pharmacology among their members significantly higher than those found on average in other large protein families of therapeutical relevance.[13–16] Therefore, even though examples of distant polypharmacology relationships have been reported recently using ligand-based methods,[17–21] more efforts should be devoted to exploring alternative approaches that go beyond mere molecular similarity.

The amount of protein structure information available in the public domain continues to grow exponentially and today there are over 72,000 X-ray structure entries in the Protein Data Bank (PDB).[22] Not only the number of structures is increasing but their functional coverage is being expanded as well. There is still a strong bias for recognised therapeutic targets,[23,24] but it is slowly being corrected thanks to recent structural genomics initiatives.[25,26] Consequently, even though the main body of information on protein structures is still devoted to enzymes,[23] members of protein families traditionally difficult to crystallize, such as G protein-coupled receptors (GPCRs) and other membrane proteins,

have seen their first structures crystallized in recent years.[27–33] Therefore, with such a vast and increasingly diverse structural information on proteins, structure-based methods re-emerge as a less-biased complementary alternative to well-established ligand-based methods for detecting distant cross-pharmacology relationships among targets.

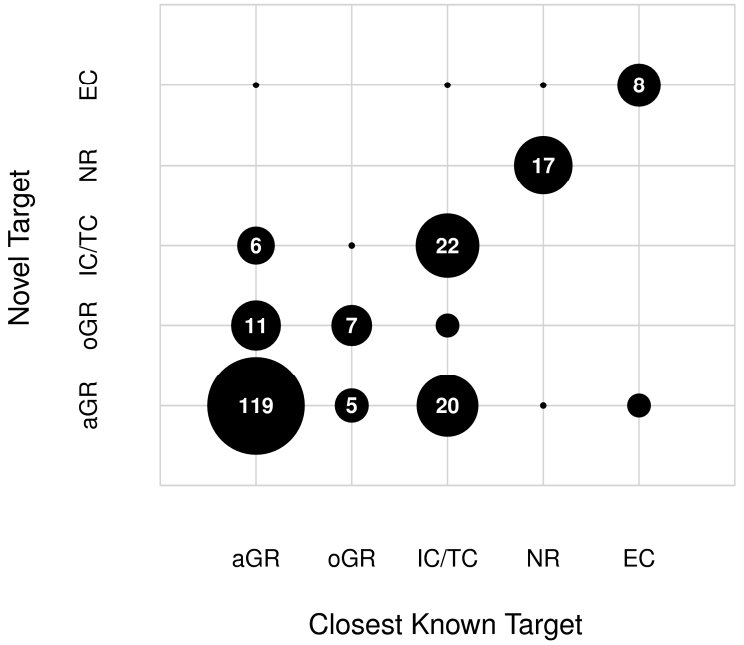


Figure 1. Relationships between the novel targets predicted by similarity-based methods and confirmed experimentally for drugs in the last five years and the family of the closest target already known in the public domain at the time of the discovery. aGR: aminergic GPCRs; oGR: other GPCRs; IC/TC: ligand-gated ion channels/ transporters; NR: nuclear receptors; EC: enzymes

Protein structures have long been exploited computationally to probe binding cavities with ligands at an atomic level.[34] However, compared to ligand-based methods, docking of ligands across thousands of protein binding

sites is a computationally demanding task, limited by the functional coverage of protein structures, and technically challenging because it requires some automated protocols for setting up large collections of heterogeneous binding sites.[10] In spite of all these inherent difficulties, several examples have illustrated recently the potential of structure-based approaches to identify the targets of small molecules.[35–41]

In recent years, a diverse range of structure-based approaches to detecting, representing, and comparing binding sites in a computationally more efficient manner has emerged as promising new tools to identify phylogenetically-distant targets to which small molecules may show polypharmacology.[13,42] For example, a promising alternative to avoid having to dock ligands on thousands of proteins, is to organise binding sites based on their similarities and assume that they will bind to similar ligands.[43] Accordingly, this review aims at collecting the most significant developments in the various aspects involved in the identification of similar binding sites and its use to detect distant polypharmacology.

2 Identification of similar binding sites

Binding site similarity is easily detected between members of the same protein family, consistent with the relatively high levels of intra-family polypharmacology observed among drugs.[14] Also, similarity within binding sites has been shown to be in many cases more conserved than sequence and the overall protein structure.[44,45] Full sequences may have diverged and structures may have led to different overall topologies, but local portions of the structure essential to protein function may have been retained. Indeed, it is well known that proteins with low sequence identity or fold similarity can still share a common function and bind to the same or highly similar ligands.[42,46] For

example, serine proteases are a group of proteins with the same function that span over several sequence and structure families. They share a common trypsin-like catalytic triad that in some instances cannot be detected by full sequence comparison or structural alignment. P-loops (a phosphate binding motif) are also conserved among otherwise unrelated nucleotide binding proteins.[47–49] Conversely, proteins with common folds can also perform different functions, as it is the case of TIM-barrels or immunoglobulin-like structures.

The fact that the function of proteins is, in many instances, dependant on a limited number of residues has promoted the development of methods that look for conserved patterns in protein structures.[50–52] Among those, local structural pattern screening methods, such as PINTS[53], SPASM[54], TESS[55], and ASSAM,[56] query a structure database with a three-dimensional pattern to detect similarities in the absence of fold and sequence similarity, as a means to infer protein function. Basically, they try to identify groups of common amino acids in two protein structures independently of their sequence order.[57] Alternatively, a protein structure can also be compared to a database of predefined motifs. The catalytic triad in serine proteases is a paradigmatic example of such structural templates. In contrast, surface matching methods aim at identifying common properties independently of residue and three-dimensional atom coordinates, based on the fact that it is the protein surface what is ultimately involved in the interaction with ligands.[58] These methods are specially suited to detect cases of convergent evolution, where protein structures can differ largely. For example, de Rinaldis *et al.*[47] adapted structural pattern searches to surface comparison to estimate the similarity between the surface profile generated by a multiple structural alignment to a second protein surface in an attempt to ascertain whether two surfaces with similar geometric and chemical properties appear in different folds.

2.1 Binding site detection

As small molecules interact with well-formed internal cavities or concrete binding clefts on the surface of proteins, only residues delineating those regions need to be considered to define protein-ligand binding. Identifying the boundaries of such regions is therefore a necessary step previous to any kind of binding site comparison. In this respect, there is an increasing amount of protein structures having a co-crystallized ligand in the binding cavity.[4,63,64] In these cases, identifying the binding cavity is straightforward and the residues defining it can be readily selected by its proximity to the interacting ligand. However, in many other instances, protein structure entries do not contain any ligand. To address these situations, computational methods have been developed to predict putative binding sites.[60,65]

Many approaches are based on pure geometrical criteria and rely on the fact that binding sites tend to be in the largest deep cleft cavity of the protein surface, especially in enzymes.[22,66] Methods such as SURFNET[67], LIGSITE[68], PASS,[69] CASTp[70] and PocketPicker[71] are representative of geometry-based approaches to identify buried volumes on protein surfaces. Among them, LIGSITE[68] embeds the protein in a three-dimensional regular grid. Each point of the grid found outside the protein volume is scanned along multiple orthogonal axes and checked whether they hit at some point any of the grid points inside the protein. The larger the number of axes hitting the protein surface, the more buried the grid point is considered to be. Clustering of contiguous buried grid points is then applied to define any surface pocket

located. The size of the grid and the number of scanned axes define the level of resolution of the pockets predicted. PocketPicker[71] uses a conceptually similar strategy and scans the molecular surroundings of grid points close to the protein surface along 30 search rays to obtain a buriedness index for each grid point. Inclusion of evolutionary information, by considering conserved residues

among a protein family likely to be involved in ligand binding, has been shown to improve the detection ability of geometry-based methods and reduce noise.[72,73]

Beyond pure geometry-based approaches, some energy-based methods have also emerged recently based on interaction energies between the protein and small chemical probes.[60] Among them, Q-SiteFinder[74] uses criteria based on van der Waals interaction energies of a methyl probe with the protein and it was shown to be able to detect smaller binding sites, in a more accurate manner, than geometry-based approaches. Also, PocketFinder[75] is based on a transformation of the Lennard-Jones potential to predict envelopes representing the shape and size of putative binding volumes, and SiteMap[76] is an example of an approach that combines geometry- and energy-based criteria to cluster the grid points that will be ultimately selected to define surface cavities. Finally, FINDSITE[77] takes a completely different approach and has the additional advantage that it can be used in low-resolution models. From a target sequence, template structures are identified using a threading algorithm. Then, structures containing a bound ligand among all templates are superimposed to the target structure and any cluster of ligands observed is used to define putative binding sites.

Binding site detection is not a simple task and the availability of such a large and diverse number of approaches is illustrative of the inherent difficulties encountered to clearly define what constitutes a binding pocket. Pockets may vary widely in shape and size and just defining the actual extent of a pocket is not trivial, each method having its own criteria. Nonetheless, Schmidtke *et al.*[78] recently conducted a comparison between several methods concluding that both geometry- and energy-based methods exhibit similar performances, with over 95% of the cases detecting the true binding site among the top-5 pockets identified.

2.2 Binding site representation

Any method that aims at comparing binding sites shall first represent them using some mathematical description that captures the essential features relevant for binding. This is a key step in the entire process, since the choice for a particular binding site representation will strongly determine the actual perception of similarity among binding sites. In this respect, a simplified representation of the binding site is desired in order to reduce the complexity of subsequent pair-wise comparisons and save computational cost. At this stage, it is important to preserve all potentially relevant information and retain all key features in the binding site. A large amount of inadequately located information will only add noise and lead to poor performance. Thus, a useful binding site representation should include an optimal level of fuzziness that allows obtaining the same representation within a certain degree of protein flexibility. Indeed, exact atomic positions may differ slightly among various structures of the same protein due to side chain rearrangements upon ligand binding. Ideally, these subtle variations between structures of the same protein binding to different ligands should have a minimal effect in the binding site representation.

Taking these° considerations into account, the representation of a binding site is usually reduced to a set of meaningful pseudo-centres that capture the geometric, pharmacophoric and/or physicochemical information of its surrounding amino acids. Perhaps the simplest and straightforward representation is to use the atomic positions of the Cα carbons from the residues defining the binding site, labelled with generic residue properties.[79] This may be complemented with the atomic positions of the Cβ carbons from the residue side chains, labelled also with the respective pharmacophoric properties. In an attempt to move away from exact atomic positions, other strategies are to place property-labelled pseudo-atoms in the center of mass of the residue side chains or in predefined geometric positions relative to binding

site amino acids.[80,81] Alternatively, all atoms of the flanking residues[82,83] or vertices of the triangulated surface[84] defining the binding site can also be used if a more detailed representation is desired at the expense of a much higher computational cost during binding site comparisons. Some recent approaches have tried to escape from the positional bias of using atom coordinates. Among them, Baroni *et al.*[85] use GRID force field analysis to locate minimal energy points, Hoffman *et al.*[86] use of atomic densities, and Jalencas and Mestres[43] place pseudo-centres directly on the protein surface defining the binding site.

One of the first methods that took care of these considerations, and that has since been an inspiration to others, is Cavbase[87]. It condenses the physicochemical properties of the residues delineating the binding site in a restricted set of generic pseudo-centres corresponding to properties relevant to ligand-protein interactions, namely, hydrophobic aliphatic (H), aromatic (R), hydrogen-bond acceptor (A), donor (D) and mixed donor/acceptor (AD). The inclusion of a mixed donor/acceptor feature is justified because protonation states can sometimes be difficult to determine. Such pseudo-centres are subsequently filtered based on their surface exposure. These pseudo-centres are then used to align binding sites by clique detection (*vide infra*). The binding site representation used in Cavbase[80,87] has been adopted by many other methods, such as MolLoc[88] or ProBiS[89] (see Table 1). In some cases, the features of the pseudo-centres have been expanded to include partial charges,[86] electrostatic potential,[84] positively (P) and negatively (N) charged regions,[90] or to differentiate between aromatic features in face and edge positions.[91]

The use of pure geometric representations such as shape curvatures, spherical harmonics or wavelet coefficients has also been explored.[92] For example, Kahraman *et al.*[93] used pure shape descriptors to compare binding sites and ligands and arrived to the conclusion that the assumption that binding sites that interact with similar ligands have similar geometries is only partially true. It was found that similarity is more closely related to the flexibility of

molecules, being the shapes of binding sites more variable than those of the ligands. This fact suggests that shape complementarity alone is not sufficient for molecular recognition, especially with highly flexible ligands, and thus binding site representations should include additional physicochemical properties. In addition, binding pockets tend to be larger than ligands, leaving spaces that are either left empty or being occupied by water molecules. Also, An *et al.* used five shape descriptors (volume, surface area, and the 3 principal axes of the binding pocket) together with two physicochemical descriptors (hydrophobicity and electrostatic charge) to compare what they refer to as ligand binding envelopes.[75] Finally, Nayal and Honig characterized surface cavities by a set of 408 physicochemical, geometrical and structural attributes (SCREEN) and then used a random forest classifier to successfully discriminate drug-binding cavities.[94]

Deriving a mathematical representation of thousands of protein binding sites can be a computationally demanding task. However, it ought to be stressed that, once computed, representations can then be stored in a database prior to binding site comparisons. It can take longer to build the database of binding site representations than to perform binding site comparisons themselves, but it only needs to be done once.

2.3 Binding site comparison

Comparing binding sites involves often finding the best rigid-body transformation that leads to an optimal three-dimensional superposition of protein environments. The results of this process are however strongly dependent on three key aspects: the binding site representation (*vide supra*), the similarity metric, and the particular algorithm used to compare the binding sites.

Defining a metric or score to quantify the degree of binding site similarity is an essential requirement to the entire process. A wide range of symmetric and asymmetric scores are currently in use.[95] In addition, binding site similarities account usually for the number of matching binding-site features to which different weighting schemes can be applied. Sometimes, it can be difficult to compare binding sites of different sizes[96] and thus, it is important to use scoring schemes that can be applied also to assess local similarities between small protein environments within binding sites.[97,98] Finally, some methods provide also a measure of the statistical significance of the similarity, like Z-scores, E-values[89] or p-values, which in a number of cases are derived from an extreme value distribution (EVD).[99–101] For example, eF-site combines a Z-score and coverage to evaluate pairwise similarities,[102] and Davies *et al.* use a Poisson index as a probabilistic model devised specially for binding site comparisons in the context of the SitesBase database.[82,96]

As the degree of similarity between binding sites relies often on the alignment procedure, failing to find the best alignment solution may produce an underestimation of their similarity. This is perhaps one of the reasons why a wide variety of alignment methods exist currently. Popular algorithms include iterative search, geometric matching, geometric hashing, and clique detection. All these methods have long been used in other scientific disciplines and adapted to binding site comparisons. Iterative search algorithms, for example, can be applied when the binding site has been simplified to an extent where evaluating all possible alignments is feasible. Only the top scoring alignment is finally retained. It is a straightforward procedure but also relatively slow.

Geometric matching algorithms compare groups of features between the binding sites. Feature triplets are commonly used as the minimal representation of local protein environments, although pairs and quadruplets are also employed. A triplet is characterised by the features of the vertices and the length of the edges connecting them. Two triplets are considered equivalent if

both have the same vertex features and similar edge lengths (within a given threshold). All triplets defining one binding site are then matched to all their equivalents in another binding site. The possibility of matching symmetries may also be considered. A transformation is performed by each triplet match mapping one triplet to the other by a least squares fitting routine or using quaternions.[103] Each transformation generates an alignment between a pair of binding sites, the quality of which is then evaluated and scored. The mapping with the highest score is conserved. When pure geometric binding site representations are used, surface matching is found to be able to find correct solutions when binding sites are highly similar, but performance drops significantly as similarity decreases.[104] Among these pure geometric approaches, MolLoc adapts spin image representations used in three-dimensional object recognition to locate similar regions in protein surfaces containing matching pairs of atoms with similar physicochemical properties.[88]

Geometric hashing is a more efficient variation of geometric matching, developed originally in computer vision and later adapted for binding site comparisons.[105,106] In contrast to geometric matching, the features of the binding site are converted to a hash table.[49] Each key in the table is a group of features, often a feature triplet defined by the properties of the vertices and the length of the edges connecting them. The features of the other binding site to be compared are grouped in the same way and used to access the hash table to obtain matches. Each match defines a transformation representing a potential solution that can be globally scored and stored. A sample of those matches (seed matches) is first obtained and clustered. Those representing a similar transformation are grouped together, and a representative of each cluster is then selected. The most common transformation corresponds to the largest alignment that will allow obtaining subsequent feature matching and scoring for the entire binding site. This approach of expanding local matches to the full

cavity makes the methodology more suited to detect local similarities. Methods based on three-dimensional patterns, such as TESS,[55] use geometric hashing.

Clique detection algorithms aim at finding the maximum subgraph isomorphism between two binding sites represented as graphs.[107] A product graph is then obtained by pairing nodes with compatible labels between query and target graphs.[56] Node pairs are linked if the edges connecting both elements in the original graphs are similar. Nodes are usually labelled with descriptors associated with the pharmacophoric properties of local protein environments, whereas edges are labelled with distances that reflect the shape of the binding site. Cliques of product graphs correspond to subsets of adjacent pairs of target nodes that satisfy both geometrical and physicochemical constrains. The largest common subgraph identified will then be used to generate a three-dimensional superimposition and its size to provide a rough measure of the similarity between binding sites. The Bron-Kerbosch algorithm is often used for this purpose.[108] However, clique detection is a NP-complete problem meaning that, although possible solutions can be quickly evaluated, there is no efficient way to locate the best solution. This implies that the time required to solve the problem increases very quickly as the size of the problem grows. Depending on the size and labels of input binding site graphs, the product graph can easily grow to dimensions that cannot be processed at affordable running times.

Several smart strategies have been devised to address this issue. For example, eF-sites decomposes the entire binding site surface in small portions, so sub-cliques are obtained first between those surface patches that are then combined into a binding site clique solution if the transformation they produce is similar.[84] Also, IsoCleft incorporates two innovations that allow including in the model all the atoms of the binding site. First, an initial superimposition is obtained using clique detection on C_α atoms only. The resulting superimposition is used to filter the pairs of the product graph constructed

from all non-hydrogen atoms. Then, a modified Bron-Kerbosch algorithm is used to select the first clique solution to be explored instead of generating all cliques and retaining the largest clique only in the end. This strategy allows for obtaining a nearly optimal solution with a significant gain of time. Along the same lines, Weskamp *et al.*[109] introduced the k-clique hashing algorithm that combines the advantages of clique detection with the speed of geometric hashing. They replaced the relatively slow clique detection step in Cavbase[87] by a clique hashing approach consisting on applying clique detection to a simplified product graph whose nodes represent larger local matches of size k that are finally assembled. Alternatively, Hoffman *et al.* used a convolution kernel approach between two clouds of points.[86] Its main advantage is that it does not require a pair-wise alignment of those points, capturing instead similarities between atom densities. This allows for a smoother alignment and reflects the fact that atoms in different positions can have equivalent roles in ligand binding. 3D-Zernike descriptors have also been shown suitable for local or global binding site description and comparisons.[110–112] As a final example, SiteAlign maps binding site properties on a faceted sphere located at its centre of mass. The alignment of such spheres is intended to provide a better tolerance of atomic variations and rotameric states than the rather crisp descriptions offered by pseudo-atom methods.[98]

Since the search for the best three-dimensional match between binding sites can be computationally highly demanding, a variety of alignment-free methods have been also developed to allow for large scale binding site comparisons. For example, one can use emergent self-organizing map (ESOM) to project feature vectors in a two-dimensional space.[92] Alternatively, Stahl *et al.*[91] used a self–organizing neural network on correlation vectors encoding atom types and surface shape to successfully classify pockets of zinc-containing metalloproteases according their enzymatic class. Anderson *et al.*[113] replaced direct geometry comparisons between binding sites for a principal component

analysis on a wide set of cavity properties. They were able to show that the most important general features for differentiating ligand-binding cavities were size/shape, polarity, charge, depth/shape, electrostatic field and aromaticity. A popular approach is the use of fingerprints, where binding site properties are projected in a numerical high dimensional vector. In this direction, FLAP uses a GRID force field analysis to locate minimum energy points in molecular interaction fields that are then used to generate binding site four-point pharmacophore fingerprints complementary to those of ligands.[114] PocketPicker[71] compares binding sites by describing the shape of the pocket with the buriedness of grid probes. Binned distances between grid points classified according their buriedness in six bins leads to a 420-dimensional shape descriptor. The similarity between binding sites is assessed by computing the Euclidean distance of their respective shape descriptors. FuzCav encodes binding sites in a fingerprint of triplets by labelling C_α atoms with pharmacophoric properties and binning the distances between them[115] Ito *et al.* introduced a particularly fast and scalable variation of FuzCav based on Structural Sketches, which are bit strings created by random projections of triplet descriptors allowing similar pairs to be found by multiple sorting.[116] To minimize the impact of discretised distance ranges in the fingerprint generation, KRIPO adopts fuzzy fingerprints.[80]

As a counterpart of their speed, alignment-free methods lack the interpretability that otherwise alignment methods provide. It is for that reason that some methods, such as SubCav, use a fingerprint approach to rapidly compare binding sites that is complemented with a final alignment step. In this case, the assignment of matching atoms required for the alignment is based on atoms that share the largest amount of fingerprint elements.[117] For the sake of clarity, Table 1 summarizes the main characteristics of all methodologies reviewed in this work.

Table 1. List of existing computational approaches for the structure-based comparison of protein binding sites.

Method	Features	Feature position	Alignment	Scoring
Stahl *et al.*[91] 2000	Accessibility, Aliphatic, Rface, Redge, A, D.	Surface points	Topological correlation vectors.	Self-organizing neural network.
eF-seek [84,102,118] 2002	Electrostatic potential, hydrophobicity and curvature	Triangulated surface vertices	Modified clique detection	Weighted count of matches.
Cavbase [87,119] 2002	H, R, A, D, AD	Residue pseudo-atoms	Clique detection	Overlapping surface points.
CSC [120] 2003	Atom element	Representative side chain atoms	Pairing of 4-atom local environments and merging.	Number of matching atoms
SuMo [79] 2003	Predefined chemical groups and atom burial	Chemical group mass centre.	Geometric matching (triplets) on a graph of pairs of adjacent similar chemical group triangles.	Number of matching groups
pvSOAR [99,121] 2003	Sequence fragment of exposed residues	Amino acid centre of mass	Sequence alignment followed by structural alignment, coordinates and orientation RMSD	p-value (EVD)
SitesBase [49,122] 2004	Atom elements C, N, O, S, P	Binding site atoms	Geometric matching (triplets)	Number of coincident mapped atoms
SiteEngine [90] 2004	H, R, A, D, AD, P, N Surface patch shape	Residue pseudo-atoms	Geometric hashing (triplets)	Hierarchical scoring scheme
SURFACE [123] 2004	Residue (substitution matrix)	Residue C_α and side chain geometric centre	Geometric matching (pairs of residues)	Number of overlapped residues
CPASS [97] 2006	Residue type and shortest distance to ligand atom.	C_α	Iterative search	RMSD weighted BLOSUM62

MultiBind [124] 2006	Cavbase + topological	Cavbase	Geometric hashing. Branch-and-bound algorithm for multiple alignments.	Sum of similarities for the matched pseudocenters
Park and Kim[125,126] 2006	Residue type	C_α	Clique detection	BLOSUM62
Zhang *et al.*[127] 2006	Residue class.	Residue side chain geometric centre.	Clique detection	Tanimoto
FLAP [85] 2007	Grid force field.	Pharmacophoric points	4-point pharmacophore fingerprints.	Cosine-like similarity
PocketPicker 2007	Buriedness	Pocket grid probes	2-point fingerprints	Euclidean distance
Ramensky *et. al.*[128] 2007.	43 force-field chemical types	Binding site atoms around a ligand atom	Clique detection	Proportion of matched atoms.
3D-Surfer [110] 2008	Electrostatic potential and hydrophobicity	Grid-discretised surface	3D Zernike Descriptor	Euclidean distance
BSAlign [129] 2008	Physicochemical and geometric	C_α	Clique detection	Number of aligned residues balanced by RMSD
IsoCleft [83] 2008	H, R, A, D, neutral, neutral-donor, and neutral-acceptor.	Binding site non-hydrogen atoms.	Clique detection	Tanimoto
PocketMatch [130] 2008	Amino acid groups	C_α, C_β and side-chain centroid	Sorted list of distances between pairs of points (2D fingerprints)	Petke similarity
PROSURFER [131] 2008	H, A, D, P, N and “other”	Feature vector in each surface atom describing the local environment.	Geometric matching (atom triplets)	Tanimoto-like
SiteAlign [98] 2008	Topological and chemical for each residue.	Projected from the Cα to a faceted sphere at the centre of the site.	Iterative rotation and translation of a sphere over another one.	Average of normalised differences along a fingerprint.

SOIPPA [132,133] 2008	Geometric potential	C_α Delaunay tessellation	Clique detection	Weighted sum of profile distances for each aligned Cα pair.
SurfaceScreen [134] 2008	Global surface shape and physicochemical texture	Residue mass centre.	Iterative search for the best superimposition of common residues.	Surface Volume Overlap Tanimoto (SVOT)
WaveGeoMap [92] 2008	Shape (wavelet) and physicochemical (H, R, A, D, AD	Feature vector assigned to each surface patch.	Common orientation of surface patches	Emergent self - organizing map on feature vectors
MED-SuMo[135] 2009	Physicochemical	Surface Chemical Features. (SCF)	Geometric matching (triplets) on a graph of pairs of adjacent similar chemical group triangles.	Number of common features and local shape similarity
MolLoc [88] 2009	Geometric spin images	Connolly surface points.	Geometrically consistent correspondences refined by atom types.	Corresponding surface area.
PESD [136] 2009	Electrostatic potential, polar, hydrophobic and hydrogen-bonding	Sample of Gauss-Connolly surface points.	Property encoded D2 shape distribution. Binned distances between pairs of points.	Signature comparison
VA [137] 2009	Principal components of 29 physicochemical amino acid properties	C_α	Clique detection	Clique size
Yin *et al.*[138] 2009	Curvature	Surface points	Geometric fingerprints (distance-dependant distribution of curvatures)	Root-mean deviation of each fingerprint bin.
Anderson *et al.* [113] 2010	SCREEN[94]	Vector assigned to each binding site	Principal Component Analysis	PLS-DA
BSSF[139] 2010	Physicochemical features	Residue fragments centroids	Fingerprint of distances	Canberra distance and z-score

FuzCav [115] 2010	HRADPN	C_α	3-point pharmacophore fingerprint	Simpson
Hoffmann *et al*. [86] 2010	Partial charge	Cloud of atoms	Convolution kernel	Convolution kernel
ProBiS [89] 2010	Cavbase	Functional groups pseudo-atoms	Clique detection on overlapping subgraphs	E-value
PocketFeature [140] 2011	Physicochemical features (based on FEATURE)	Spherical microenvironments around residues (6 shells)	Exhaustive for all microenvironment pairs	Normalized Tanimoto
Structural Sketches [116] 2011	Physicochemical and geometrical properties (8 sets of 4 properties)	C_α	Bit strings of 3-point pharmacophore fingerprint random projections (SketchSort)	Cosine
KRIPO [80] 2012	HRADPN	Geometric positions relative to amino acids	3-point fuzzy pharmacophore fingerprint	Modified Tanimoto
APoc [101] 2013	Side chain orientation (Cα-Cβ vector) and residue type.	C_α	iAlign [141]	p-value (EVD) for PS-score
Jalencas *et al.* [43]	HRAD	Surface feature points	Clique detection	Cosine
SubCav [117] 2013	Ppharmacophoric features (D, CA, C, P, H, D=, A=, H=)	Binding site atoms	3-point pharmacophoric fingerprint	Tanimoto and Cosine
TrixP [142] 2013	HAD + pocket shape	hydrophilic atoms and hydrophobic grid points	3-point pharmacophoric fingerprint + geometric matching like	Combined score for all features.

2.4 Binding site fragmentation

Protein flexibility, which can range from slight side-chain rotations to relatively large backbone rearrangements, may result in significant variations in the size and shape of binding cavities, thus hindering the detection of binding site similarities.[143] Accordingly, the simplistic lock and key model of the molecular recognition event has been gradually refined to account for protein flexibility. In this respect, the induced-fit model takes into consideration the adaptive conformational changes that the protein undergoes when binding to ligands, whereas the selected-fit model looks at the event from the perspective of the ligand, which stabilises a complementary protein conformation among the many different conformations a protein can adopt in equilibrium.[144] The direct consequences of these various aspects of protein flexibility are that structure-based methods to comparing entire binding sites may have difficulties detecting similarities among the different holo and apo structures of the same protein but also among protein structures interacting with different ligands. Therefore, strategies to detect local binding-site similarities may represent an alternative to global binding-site comparisons, potentially less sensitive to protein conformational changes.

Several methods have implemented different strategies to address the issue of protein flexibility, such as increasing the degree of fuzziness of the descriptors, focusing on local binding-site environments, or using an ensemble of protein conformations derived from molecular dynamics simulations. The sensitivity of a particular method to atom rearrangements depends very much on the type of the descriptors used to define the binding site. For example, a binding site representation based on the positions of the C_α carbons, although less accurate, will most probably remain invariant to side-chain conformational changes and be affected only by significant backbone rearrangements. On the other hand, a representation based on the atomic positions of side chains or on

surface points will be more sensitive to conformational variations. The challenge is thus to have the right balance between accuracy of the binding-site description and sensitivity to conformational changes.

Perhaps the most common approach to deal with protein flexibility is to search for similar local protein environments instead of comparing entire binding sites. Ramensky *et al.* were amongst the first to introduce a local approach to binding site similarity and use the assumption that similar protein environments will interact with the same chemical fragment in a similar orientation.[128] However, assessing the true relevance of local binding site similarities is not straightforward as they can be scattered across the entire binding cavity in completely disconnected patches. Because of that, a balance between the similarity among several small surface patches and that of the global binding cavity should be taken into consideration.[92] In this respect, how sub-cavities are defined will certainly influence the final outcome. DoGSite, for example, allows to define surface regions in apo structures based solely on the topological characteristics of the binding cavity.[145] However, binding site fragmentation is more straightforward when holo structures are available and can be defined on the basis of a ligand fragmentation.[146,147] Wallach *et al.* used this strategy to deconstruct binding cavities in a set of potentially overlapping sub-cavities according to chemical groups of co-crystallized ligands.[148] More recently, Jalencas and Mestres[43] used also chemical fragments to define protein environments as interacting binding-site surface regions of consistent pharmacophoric features.

Methods performing local binding-site comparisons between protein environments defined from interacting chemical fragments are particularly well suited for fragment-based drug discovery.[43] It is commonly accepted that some fragments are prone to interact more frequently than expected with certain amino acids. For instance Chan *et al.* retrieved from the Protein Data Bank all chemical fragments forming hydrogen bonds with the most common residues

in the binding site (i.e. aspartic, glutamic, arginine and histidine).[149] Wang *et al.* characterized chemical fragments from the Protein Data Bank by counting their closest residues. Those fragment-residue interaction profiles showed that chemical fragments have specific preferences for certain types of residues.[150] Along the same lines, Soga *et al.* introduced the term chemocavity as a specific concavity in a protein where a specific group of small molecules (a canonical molecular group) is inclined to be bound. Moreover, they introduced a chemocavity index based on the amino acid concurrence rate and were able to correlate a chemocavity with its corresponding molecular group, thus reinforcing the idea that there may be specific sites for particular chemical fragments.[151] These approaches can be used to quantify the optimality of the interaction between a protein environment and a chemical fragment.

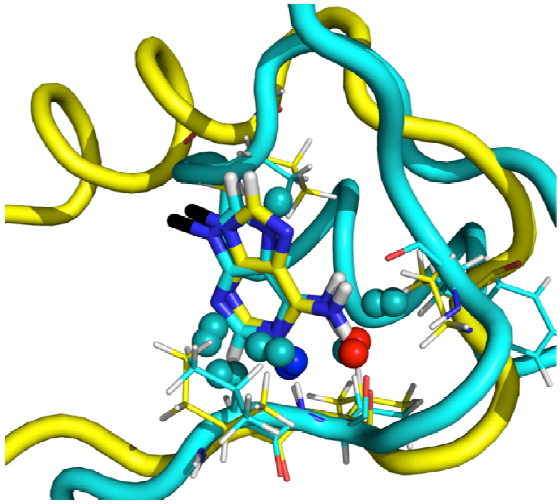


Figure 2. An illustrative example of a pair of chemoisosteric protein environments from structurally diverse proteins that interact with the adenyl fragment, namely, citrate synthase (PDB code 1csi, cyan) and glucokinase (PDB code 1ua4, yellow).

Some methods offer the possibility of providing the three-dimensional alignment of local protein environments.[43,80,135,152] Since protein environments are usually defined based on the interacting chemical fragments present in the native ligand, alignment of protein environments translates directly into superposition of chemical fragments. This offers great opportunities for the use of these methodologies in fragment-based drug discovery.

3. Detection of distant polypharmacology

In recent years, it has become apparent that selective drugs are more the exception rather than the rule and that most therapeutically effective molecules bind to multiple proteins.[15] This ability of small molecules to interact with multiple proteins is commonly referred to as polypharmacology.[13] Since experimental testing of molecules on *in vitro* binding assays for thousands of proteins is currently unfeasible, *in silico* methods based on ligand similarity have proven very useful in predicting novel targets for known drugs.[17,153–160] However, as emphasised above, for the vast majority of the new drug-target interactions identified there was already a known interaction to a target of the same protein family. In fact, detection of polypharmacology across members of the same protein family is not surprising. From a structure-based viewpoint, Kinjo and Nakamura showed that the majority of similar binding sites are confined within the same protein family, with the exception of nucleotide and ion binding sites.[161] In fact, very few examples actually exist on the use of these ligand-based methods to predict affinities between molecules and proteins distantly related to any of their already known targets.[17,19–21] In this respect, structure-based methods may complement ligand-based methods to detect distant polypharmacology.

Indeed, the PDB offers structural evidence of distant polypharmacology examples.[22] We took a set of 1,358 approved drugs (excluding nutraceuticals) from DrugBank 3.0[8] and searched them using their InChI keys in Ligand Expo.[63] Of those, 387 compounds involving 2,655 structures were located in the PDB, 234 of them present in more than one PDB entry. Detection of distant polypharmacology examples was performed by assigning structural[162,163] and functional[24,164] classification codes to each structure. A total of 138 drugs were found to be co-crystallized in at least two structurally and functionally unrelated proteins. Filtering for drugs with known affinity in PDBbind 2011[64] for those unrelated protein structures resulted in a list of 20 drugs for which structural evidence of distant polypharmacology currently exists. Details for four of them are provided in Table 2. For each drug, PDB entries are grouped according to their functional code and a single group representative was selected. The percentage of sequence identity, as obtained with the Needle routine from the EMBOSS package),[165] between the drug's primary target and any other binding protein is also provided as a metric of distant polypharmacology.

Among those four drugs, acetazolamide is a carbonic anhydrase inhibitor used in the treatment of a wide variety of diseases. In the PDB, it is found co-crystallised with carbonic anhydrase 2, for which nanomolar affinity is reported, but also with endochitinase, to which it binds with low micromolar affinity. Another example is indomethacin, a non-steroidal anti-inflammatory drug that acts as a prostaglandin inhibitor and that it is found co-crystallised in the PDB with 9 additional unrelated proteins and known to bind to them with low micromolar affinity. A third example is estradiol, a sex hormone with potent estrogenic effects that nonetheless it has been co-crystallised with 6 additional unrelated proteins, for some of which binding with potent affinity has been reported. Finally, caffeine is a central nervous system stimulant that acts as adenosine receptor A_{2a} antagonist but it is known to bind with low micromolar

affinity to at least two additional proteins, for which structural evidence of their interaction is also available in the PDB.

Having shown that structural evidence exists in the PDB, the challenge is now to show whether distant polypharmacology can be predicted with structure-based methods. In this respect, detection of similar local surface regions can be used to identify proteins that could interact with the same molecule.[10,166] One of the first examples where similarities between binding sites of unrelated proteins could explain polypharmacology was reported by Weber *et al.*[167] They realized that the unsubstituted arylsulfonamide moiety present in COX-2 selective inhibitors, such as celecoxib and valdecoxib, was also commonly present in carbonic anhydrase (CA) inhibitors. Using enzyme kinetics and X-ray crystallography, they were able to confirm an unexpected nanomolar affinity of celecoxib and valdecoxib for isoenzymes of the CA family. Cavbase[87] was able to find surface property similarities between COX-2 and CA-II binding sites for the regions accommodating the sulphonamide and trifluoromethyl groups, although no overall binding site similarity was detected. Other examples were reported by Minai *et al.*, who used local atomic environments to locate structures with similar regions to those known to be co-crystallised with a drug.[131] They obtained a list of more than one million pairs of such similar regions, that were made available online. In a retrospective analysis, some of the predicted interactions could be confirmed experimentally, such as captopril and matrix metalloproteinase-12. Other predictions, like the binding of lovastatin to RXR and flurbiproben to phospholipase A2, suggested plausible novel mechanisms of action.

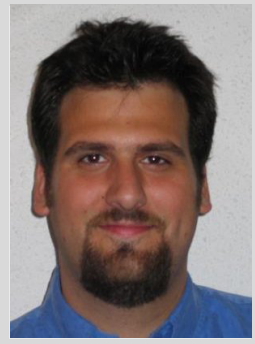
Another method that has been applied to detecting distant polypharmacology is SOIPPA (Sequence Order-Independent Profile-Profile Alignment).[133] Based on a shape descriptor initially devised to locate protein binding sites,[132] it includes also features from sequence alignment methodologies. Starting from a geometric potential, defined to quantitatively

describe the shape of a structure dependent on both the global shape and the local environment of each residue, SOIPPA represents the protein by a Delaunay tessellation of C_α carbon atoms for the whole protein, each of them labelled with its geometric potential, as well as with a position specific scoring matrix (PSSM) obtained with PSI-Blast. This triangulation is treated as a graph and clique detection techniques can be used for pair-wise order-independent local sequence alignment. This approach allowed to detect significant binding site similarities between the Estrogen Receptor subtype alpha (ERα) and a Sacroplamatic Reticulum Ca^{2+} ion channel ATPase protein (SERCA) and thus provide a potential mechanism of action by which on could explain some of the adverse effects known for selective estrogen receptor modulators (SERMs).[168] The same approach was able also to predict highly significant similarities between the NAD binding site of the Rossman fold and the S-adenosyl-methionine (SAM) binding site of SAM methyltranferases.[133] This allowed to anticipate the interaction between entacapone (a drug targeting a SAM-dependant methyltransferase, COMT) and enoyl-acil carrier protein reductase (InhA), a NAD-binding protein, which is a target for anti-tubercular drugs. This prediction was subsequently validated by *in vitro* testing of Comptan (whose active component is entacapone), showing an IC_{50} of about 80 μM on InhA, well beyond its toxicity concentration.[169] Finally, SOIPPA was used to successfully identify off-targets for the *T. brucei* RNA editing ligase 1 inhibitor drug candidate NSC-45208.[170]

In addition, SiteAlign was used to predict the potential interaction of protein kinase inhibitors with synapsin I, an ATP binding protein.[171] The prediction was validated by an *in vitro* competition assay, giving an affinity of staurosporine to bovine synapsin I of about 0.3μM. Other more specific kinase inhibitors, such as roscovitine and quercetagetin, were also found to be synapsin I nanomolar inhibitors.[171] Also, PocketPicker was used by Stauch *et al.* to predict putative binding pockets in a model of the APOBEC3C protein, which has

All methods aiming at identifying similarities among binding sites involve performing tasks related to binding site i) detection, to focus on the concrete region to be compared, ii) representation, with appropriate descriptors relevant to the geometric and pharmacophoric features, iii) comparison, to align binding sites and score their similarity, and iv) fragmentation, to offer a more local perspective and reduce complexity.[59] For each of these tasks a rich variety of approaches exist.[59–62] The following sections contain a detailed survey of those methods, many of which are compiled in Table 1, including concrete examples of identification of similar binding sites and how these gave rise, in some cases, to detection of distant drug polypharmacology.

Xavier Jalencas was born in Terrassa, Catalonia. He received a Master's Degree in Biology from the Universitat Pompeu Fabra in 2005. He is in his final year as a PhD student in the Chemogenomics Laboratory at IMIM (Barcelona). His thesis focuses on the detection, fragmentation, comparison, and classification of protein binding sites as a means to better understand drug polypharmacology.



Jordi Mestres was born in Girona, Catalonia. He received a PhD in Computational Chemistry from the University of Girona in 1996. After seven years in pharmaceutical industry (Pharmacia&Upjohn and Organon), he took on his current position as Head of Chemogenomics at IMIM. His research interests focus on the development of computational approaches to systems chemical biology and drug discovery.

strong antilentiviral activity.[172] Comparing the largest of those pockets with a set of pockets with known ligands suggested that nucleic acids could act as substrates. This was experimentally demonstrated to be true for RNA, whose binding was required for the incorporation of APOBEC3C into viral particles.

Cavbase was used to compare and cluster a set of ATP binding sites from 258 protein kinases spanning 48 SCOP families. As expected, high sequence similarity was correlated with high binding site similarity, and they obtained a proper subfamily classification. Nevertheless they observed cases where kinases with high sequence similarity exhibited a modest binding site similarity as well as binding site similarities between sequence-unrelated kinases.[173] Some examples of the later, supported by the existence of a molecule inhibiting both kinases, are provided by Kalliokoski *et al.* [174] In a similar experiment, SitesBase was used to cluster a non-redundant set of binding sites from 354 protein kinases. The clustering was compared to that obtained from sequence alignment and unexpected binding site similarities and dissimilarities were reported.[175] Experimental interaction profiles were also compared with binding site similarity profiles, yielding high enrichment factors. Milleti *et al.* used a shape-context based descriptor to predict kinase inhibition maps by pocket similarity.[176] Finally, Brylinsky e*t al.* combined a modified version of PocketMatch with sequence, ligand and experimental data in a machine-learning approach to compute a putative map of cross-interactions within the human kinome.[177]

Finally, Jalencas and Mestres[43] introduced recently the concept of chemoisosterism of protein environments as the complementary property to bioisosterism of chemical fragments. In the same way that two chemical fragments are considered bioisosteric if they can bind to the same protein environment, two protein environments will be considered chemoisosteric if they can interact with the same chemical fragment. Accordingly, the degree of chemoisosterism of protein environments is directly related to the level of

polypharmacology of chemical fragments. An interaction network between 1072 chemical fragments and 3177 clusters of protein environments was constructed. An analysis of this fragment-environment network revealed the existence of local chemoisosteric relationships among binding cavities of completely unrelated proteins. In particular, it was shown that one can obtain a reasonably correct fragment mapping of the binding cavity of a nuclear receptor structure by direct superimposition of protein environments from enzyme structures only and without any further computational energy optimisation to refine the placement and orientation of the associated chemical fragments. The presence of promiscuous chemical fragments in molecular structures, such as phenol, 3,4-dihydroxyfuran, pyridine, and chlorophenyl, should enhance the chances of a molecule to have affinity for multiple targets.

Table 2. Examples of approved drugs co-crystallized with at least two unrelated proteins. *CID*: PDB compound identifier, *N1*: number of structures for a particular functional code. *N2*: number of structures for a particular Uniprot accession number, *AC*: Uniprot Acession number, *Seq. id.*: sequence identity to primary target (%).[165] Primary targets are highlighted in bold.

Drug name CID	*Functional code*	*N1*	*N2*	*PDB*	*AC*	*Affinity*	*Seq. id.*	*Name*
Acetazolamide AZM	EC4.2.1.1	18	**8**	**3dc3**	**P00918**	**Ki=4.9nM**	NA	**Carbonic anhydrase 2**
	EC.3.2.1.14	2	1	2uy4	P29029	Ki=21μM	8.6	Endochitinase
Indomethacin IMN	EC.1.14.99.1	1	**1**	**4cox**	**Q05769**	**IC50= 109.57nM** [b]	NA	**Prostaglandin G/H synthase 2**
	EC.1.11.1.7	1	1	3ogw	P80025	-	18.4	Lactoperoxidase
	EC.1.3.1.48	2	1	2zb8	Q8N8N7	-	11.1	Prostaglandin reductase 2
	EC.1.3.1.20 [a]	3	3	1s2a	P42330	IC50= 4.10 μM[b]	9.8	Aldo-keto reductase family 1 C3

	NR.1.C.3	2	2	3ads	P37231	Kd=9.73μM	9.8	PPAR-γ
	EC.3.4.21.-	1	1	3ib1	P24627	-	7.1	Lactotransferrin
	EC.3.1.1.4	4	2	3h1x	P59071	Kd=1.3μM	5.7	Phospholipase A2 VRV-PL-VIIIa
	TC.9.B.35.1.1	2	2	4ik7	P02766	-	5.0	Transthyretin
	EC.5.3.99.3	1	1	1z9h	Q9N0A4	IC50=1mM	4.1	Prostaglandin E synthase 2
		3	3	2bxm	P02768	Ki=4.26μM [b]	2.9	Serum albumin
	NR.3.A.1	**10**	**10**	**1qku**	**P03372**	**Kd= 0.26nM**	NA	**Estrogen receptor**
	NR.3.A.2	**3**	**2**	**3oll**	**Q92731**	**Ki= 4.69nM [b]**	44.7	**Estrogen receptor beta**
	E.1.1.1.35	1	1	1e6w	O70351	-	5.9	3-hydroxyacyl-CoA dehydrogenase-2
Estradiol EST		2	2	1jgl	P01837	Kd=2nM	4.8	Ig kappa chain C region
	EC.1.1.1.62	6	6	1fdu	P14061	Km=30nM	3.6	Estradiol 17-beta-dehydrogenase 1
		1	1	1lhu	P04278	IC50=50.0nM [b]	2.6	Sex hormone-binding globulin
	EC.2.8.2.1	1	1	2d06	P50225	Ki=83.2μM	2.4	Sulfotransferase 1A1
	EC.2.8.2.4	1	1	1aqu	P49891	-	0.8	Estrogen sulfotransferase, testis
	GR.1.10.2.2	1	**1**	**3rfm**	**P29274**	**Ki= 19.80μM**	NA	**Adenosine receptor A2a**
Caffeine CFF	EC.2.4.1.1	7	5	1l7x	P06737	Kd=92μM	11.8	Glycogen phosphorylase, liver form
	EC.3.2.1.14	2	1	2a3b	Q873X9	IC50= 469μM	3.3	Chitinase

[a] multiple codes, only one is shown

[b] affinity values from ChEMBL[9]

Conclusion

Despite the large number of currently available structure-based methods to perform binding-site comparisons, the number of success stories with prospective experimental validations of distant polypharmacology predictions is still relatively low. The wide diversity of existing structure-based approaches to comparing protein binding sites indicates that the problem is not yet successfully resolved and the resulting signal-to-noise ratio of the predictions is far to be optimal. In fact, protein flexibility remains still a challenging issue and structural water molecules, known to have in many instances a key role in bridging protein-ligand interactions, are seldom accounted for in the binding site description and the subsequent comparison process. Each of the methods reviewed has its own strengths and limitations, and their relative performance is, at present, very much dependent on the nature of the particular problem to be solved. The potential of using protein structure data for detecting distant polypharmacology remains to be fully exploited, particularly as an integral part of the ligand-based methods for predicting ligand-protein interactions.

Acknowledgements

This research was funded by the Spanish Ministerio de Ciencia e Innovación (BIO2011-26669 and PTA2009-1865-P) and the Instituto de Salud Carlos III.

References

[1] C. M. Krejsa, D. Horvath, S. L. Rogalski, J. E. Penzotti, B. Mao, F. Barbosa, J. C. Migeon, *Curr. Opin. Drug Discovery Dev.* **2003**, *6*, 470–480.

[2] M. Olah, M. Mracec, L. Ostopovici, R. Rad, A. Bora, N. Hadaruga, I. Olah, M. Banda, Z. Simon, M. Mracec, T. I. Oprea, in *Chemoinformatics in Drug Discovery* (Ed.: T.I. Oprea), Wiley-VCH, New York, **2005**, pp. 221–239.

[3] T. Liu, Y. Lin, X. Wen, R. N. Jorissen, M. K. Gilson, *Nucleic Acids Res.* **2007**, *35*, D198–201.

[4] M. L. Benson, R. D. Smith, N. A. Khazanov, B. Dimcheff, J. Beaver, P. Dresslar, J. Nerothin, H. A. Carlson, *Nucleic Acids Res.* **2008**, *36*, D674–678.

[5] N. H. Jensen, B. L. Roth, *Comb. Chem. High Throughput Screen.* **2008**, *11*, 420–426.

[6] A. J. Harmar, R. A. Hills, E. M. Rosser, M. Jones, O. P. Buneman, D. R. Dunbar, S. D. Greenhill, V. A. Hale, J. L. Sharman, T. I. Bonner, W. A. Catterall, A. P. Davenport, P. Delagrange, C. T. Dollery, S. M. Foord, G. A. Gutman, V. Laudet, R. R. Neubig, E. H. Ohlstein, R. W. Olsen, J. Peters, J.-P. Pin, R. R. Ruffolo, D. B. Searls, M. W. Wright, M. Spedding, *Nucleic Acids Res.* **2009**, *37*, D680–685.

[7] Y. Wang, E. Bolton, S. Dracheva, K. Karapetyan, B. A. Shoemaker, T. O. Suzek, J. Wang, J. Xiao, J. Zhang, S. H. Bryant, *Nucleic Acids Res.* **2010**, *38*, D255–266.

[8] C. Knox, V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, A. Pon, K. Banco, C. Mak, V. Neveu, Y. Djoumbou, R. Eisner, A. C. Guo, D. S. Wishart, *Nucleic Acids Res.* **2011**, *39*, D1035–1041.

[9] A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, J. P. Overington, *Nucleic Acids Res.* **2012**, *40*, D1100–1107.

[10] D. Rognan, *Mol. Inf.* **2010**, *29*, 176–187.

[11] J. A. Allen, B. L. Roth, *Annu. Rev. Pharmacol. Toxicol.* **2011**, *51*, 117–144.

[12] L. Xie, L. Xie, S. L. Kinnings, P. E. Bourne, *Annu. Rev. Pharmacol. Toxicol.* **2012**, *52*, 361–379.

[13] X. Jalencas, J. Mestres, *Med. Chem. Comm.* **2013**, *4*, 80–87.

[14] J. Mestres, E. Gregori-Puigjané, S. Valverde, R. V. Solé, *Mol. BioSyst.* **2009**, *5*, 1051–1057.

[15] I. Vogt, J. Mestres, *Mol. Inf.* **2010**, *29*, 10–14.

[16] F. Briansó, M. C. Carrascosa, T. I. Oprea, J. Mestres, *Curr. Top. Med. Chem.* **2011**, *11*, 1956–1963.

[17] A. J. DeGraw, M. J. Keiser, J. D. Ochocki, B. K. Shoichet, M. D. Distefano, *J. Med. Chem.* **2010**, *53*, 2464–2471.

[18] A. A. Antolín, X. Jalencas, J. Yélamos, J. Mestres, *ACS Chem. Biol.* **2012**, *7*, 1962–1967.

[19] R. Hajjo, V. Setola, B. L. Roth, A. Tropsha, *J. Med. Chem.* **2012**, *55*, 5704–5719.

[20] X. Lin, X.-P. Huang, G. Chen, R. Whaley, S. Peng, Y. Wang, G. Zhang, S. X. Wang, S. Wang, B. L. Roth, N. Huang, *J. Med. Chem.* **2012**, *55*, 5749–5759.

[21] H. Lin, M. F. Sassano, B. L. Roth, B. K. Shoichet, *Nat. Methods* **2013**, *10*, 140–146.

[22] P. W. Rose, B. Beran, C. Bi, W. F. Bluhm, D. Dimitropoulos, D. S. Goodsell, A. Prlic, M. Quesada, G. B. Quinn, J. D. Westbrook, J. Young, B. Yukich, C. Zardecki, H. M. Berman, P. E. Bourne, *Nucleic Acids Res.* **2011**, *39*, D392–401.

[23] J. Mestres, *Drug Discovery Today* **2005**, *10*, 1629–37.

[24] R. Garcia-Serna, L. Opatowski, J. Mestres, *Bioinformatics* **2006**, *22*, 1792–1793.

[25] B. H. Dessailly, R. Nair, L. Jaroszewski, J. E. Fajardo, A. Kouranov, D. Lee, A. Fiser, A. Godzik, B. Rost, C. Orengo, *Structure* **2009**, *17*, 869–881.

[26] J. Weigelt, *Exp. Cell Res.* **2010**, *316*, 1332–1338.

[27] S. G. F. Rasmussen, H.-J. Choi, D. M. Rosenbaum, T. S. Kobilka, F. S. Thian, P. C. Edwards, M. Burghammer, V. R. P. Ratnala, R. Sanishvili, R. F. Fischetti, G. F. X. Schertler, W. I. Weis, B. K. Kobilka, *Nature* **2007**, *450*, 383–7.

[28] T. Warne, M. J. Serrano-Vega, J. G. Baker, R. Moukhametzianov, P. C. Edwards, R. Henderson, A. G. W. Leslie, C. G. Tate, G. F. X. Schertler, *Nature* **2008**, *454*, 486–491.

[29] E. Y. T. Chien, W. Liu, Q. Zhao, V. Katritch, G. W. Han, M. A. Hanson, L. Shi, A. H. Newman, J. A. Javitch, V. Cherezov, R. C. Stevens, *Science* **2010**, *330*, 1091–1095.

[30] B. Wu, E. Y. T. Chien, C. D. Mol, G. Fenalti, W. Liu, V. Katritch, R. Abagyan, A. Brooun, P. Wells, F. C. Bi, D. J. Hamel, P. Kuhn, T. M. Handel, V. Cherezov, R. C. Stevens, *Science* **2010**, *330*, 1066–1071.

[31] T. Shimamura, M. Shiroishi, S. Weyand, H. Tsujimoto, G. Winter, V. Katritch, R. Abagyan, V. Cherezov, W. Liu, G. W. Han, T. Kobayashi, R. C. Stevens, S. Iwata, *Nature* **2011**, *475*, 65–70.

[32] M. A. Hanson, C. B. Roth, E. Jo, M. T. Griffith, F. L. Scott, G. Reinhart, H. Desale, B. Clemons, S. M. Cahalan, S. C. Schuerer, M. G. Sanna, G. W. Han, P. Kuhn, H. Rosen, R. C. Stevens, *Science* **2012**, *335*, 851–855.

[33] H. Wu, D. Wacker, M. Mileni, V. Katritch, G. W. Han, E. Vardy, W. Liu, A. A. Thompson, X.-P. Huang, F. I. Carroll, S. W. Mascarella, R. B. Westkaemper, P. D. Mosier, B. L. Roth, V. Cherezov, R. C. Stevens, *Nature* **2012**, *485*, 327–332.

[34] I. D. Kuntz, J. M. Blaney, S. J. Oatley, R. Langridge, T. E. Ferrin, *J. Mol. Biol.* **1982**, *161*, 269–288.

[35] A. M. Aronov, N. R. Munagala, I. D. Kuntz, C. C. Wang, *Antimicrob. Agents Chemother.* **2001**, *45*, 2571–2576.

[36] Q.-T. Do, I. Renimel, P. Andre, C. Lugnier, C. D. Muller, P. Bernard, *Curr. Drug Discovery Technol.* **2005**, *2*, 161–167.

[37] J. Cai, C. Han, T. Hu, J. Zhang, D. Wu, F. Wang, Y. Liu, J. Ding, K. Chen, J. Yue, X. Shen, H. Jiang, *Protein Sci.* **2006**, *15*, 2071–2081.

[38] P. Muller, G. Lena, E. Boilard, S. Bezzine, G. Lambeau, G. Guichard, D. Rognan, *J. Med. Chem.* **2006**, *49*, 6768–6778.

[39] Q.-T. Do, C. Lamy, I. Renimel, N. Sauvan, P. André, F. Himbert, L. Morin-Allory, P. Bernard, *Planta Med.* **2007**, *73*, 1235–1240.

[40] S. Zahler, S. Tietze, F. Totzke, M. Kubbutat, L. Meijer, A. M. Vollmar, J. Apostolakis, *Chem. Biol.* **2007**, *14*, 1207–1214.

[41] L. Yang, J. Chen, L. He, *PLoS Comput. Biol.* **2009**, *5*, e1000441.

[42] I. Nobeli, A. D. Favia, J. M. Thornton, *Nat. Biotechnol.* **2009**, *27*, 157–167.

[43] X. Jalencas, J. Mestres, *J. Chem. Inf. Model.* **2013**, *53*, 279–292.

[44] M. Y. Galperin, E. V. Koonin, *J. Biol. Chem.* **2012**, *287*, 21–28.

[45] N. Sturm, J. Desaphy, R. J. Quinn, D. Rognan, E. Kellenberger, *J. Chem. Inf. Model.* **2012**, *52*, 2410–2421.

[46] P. F. Gherardini, M. N. Wass, M. Helmer-Citterich, M. J. E. Sternberg, *J. Mol. Biol.* **2007**, *372*, 817–845.

[47] M. de Rinaldis, G. Ausiello, G. Cesareni, M. Helmer-Citterich, *J. Mol. Biol.* **1998**, *284*, 1211–1221.

[48] K. Kinoshita, K. Sadanami, A. Kidera, N. Go, *Protein Eng.* **1999**, *12*, 11–14.

[49] A. Brakoulias, R. M. Jackson, *Proteins* **2004**, *56*, 250–260.

[50] T. Hamelryck, *Proteins* **2003**, *51*, 96–108.

[51] P. P. Wangikar, A. V. Tendulkar, S. Ramya, D. N. Mali, S. Sarawagi, *J. Mol. Biol.* **2003**, *326*, 955–978.

[52] P. F. Gherardini, G. Ausiello, M. Helmer-Citterich, *PLoS ONE* **2010**, *5*, e11988.

[53] A. Stark, R. B. Russell, *Nucleic Acids Res.* **2003**, *31*, 3341–3344.

[54] G. J. Kleywegt, *J. Mol. Biol.* **1999**, *285*, 1887–1897.

[55] A. C. Wallace, N. Borkakoti, J. M. Thornton, *Protein Sci.* **1997**, *6*, 2308–2323.

[56] R. V. Spriggs, P. J. Artymiuk, P. Willett, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 412–421.

[57] R. B. Russell, *J. Mol. Biol.* **1998**, *279*, 1211–27.

[58] A. Via, F. Ferrè, B. Brannetti, M. Helmer-Citterich, *Cell. Mol. Life Sci.* **2000**, *57*, 1970–1977.

[59] E. Kellenberger, C. Schalon, D. Rognan, *Curr. Comp. Aided Drug Des.* **2008**, *4*, 209–220.

[60] S. Pérot, O. Sperandio, M. A. Miteva, A.-C. Camproux, B. O. Villoutreix, *Drug Discovery Today* **2010**, *15*, 656–667.

[61] B. Nisius, F. Sha, H. Gohlke, *J. Biotechnol.* **2012**, *159*, 123–134.

[62] A. Vulpetti, T. Kalliokoski, F. Milletti, *Future Med Chem* **2012**, *4*, 1971–1979.

[63] Z. Feng, L. Chen, H. Maddula, O. Akcan, R. Oughtred, H. M. Berman, J. Westbrook, *Bioinformatics* **2004**, *20*, 2153–2155.

[64] R. Wang, X. Fang, Y. Lu, C.-Y. Yang, S. Wang, *J. Med. Chem.* **2005**, *48*, 4111–4119.

[65] S. Leis, S. Schneider, M. Zacharias, *Curr. Med. Chem.* **2010**, *17*, 1550–1562.

[66] R. A. Laskowski, N. M. Luscombe, M. B. Swindells, J. M. Thornton, *Protein Sci.* **1996**, *5*, 2438–52.

[67] R. A. Laskowski, *J. Mol. Graph.* **1995**, *13*, 323–30, 307–8.

[68] M. Hendlich, F. Rippmann, G. Barnickel, *J. Mol. Graph. Model.* **1997**, *15*, 359–63, 389.

[69] G. P. Brady, P. F. Stouten, *J. Comput. Aided Mol. Des.* **2000**, *14*, 383–401.

[70] T. A. Binkowski, S. Naghibzadeh, J. Liang, *Nucleic Acids Res.* **2003**, *31*, 3352–3355.

[71] M. Weisel, E. Proschak, G. Schneider, *Chem. Cent. J.* **2007**, *1*, 7.

[72] B. Huang, M. Schroeder, *BMC Struct. Biol.* **2006**, *6*, 19.

[73] F. Glaser, R. J. Morris, R. J. Najmanovich, R. A. Laskowski, J. M. Thornton, *Proteins* **2006**, *62*, 479–488.

[74] A. T. R. Laurie, R. M. Jackson, *Bioinformatics* **2005**, *21*, 1908–1916.

[75] J. An, M. Totrov, R. Abagyan, *Mol. Cell. Proteomics.* **2005**, *4*, 752–61.

[76] T. A. Halgren, *J. Chem. Inf. Model.* **2009**, *49*, 377–389.

[77] M. Brylinski, J. Skolnick, *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 129–134.

[78] P. Schmidtke, C. Souaille, F. Estienne, N. Baurin, R. T. Kroemer, *J. Chem. Inf. Model.* **2010**, *50*, 2191–2200.

[79] M. Jambon, A. Imberty, G. Deléage, C. Geourjon, *Proteins* **2003**, *52*, 137–145.

[80] D. J. Wood, J. de Vlieg, M. Wagener, T. Ritschel, *J. Chem. Inf. Model.* **2012**, *52*, 2031–2043.

[81] M. L. Waters, *Curr. Opin. Chem. Biol.* **2002**, *6*, 736–741.

[82] N. D. Gold, R. M. Jackson, *Nucleic Acids Res.* **2006**, *34*, D231–4.

[83] R. Najmanovich, N. Kurbatova, J. Thornton, *Bioinformatics* **2008**, *24*, i105–111.

[84] K. Kinoshita, J. Furui, H. Nakamura, *J. Struct. Funct. Genomics* **2002**, *2*, 9–22.

[85] M. Baroni, G. Cruciani, S. Sciabola, F. Perruccio, J. S. Mason, *J. Chem. Inf. Model.* **2007**, *47*, 279–94.

[86] B. Hoffmann, M. Zaslavskiy, J.-P. Vert, V. Stoven, *BMC Bioinf.* **2010**, *11*, 99.

[87] S. Schmitt, D. Kuhn, G. Klebe, *J. Mol. Biol.* **2002**, *323*, 387–406.

[88] S. Angaran, M. E. Bock, C. Garutti, C. Guerra, *Nucleic Acids Res.* **2009**, *37*, W565–570.

[89] J. Konc, D. Janezic, *Bioinformatics* **2010**, *26*, 1160–1168.

[90] A. Shulman-Peleg, R. Nussinov, H. J. Wolfson, *J. Mol. Biol.* **2004**, *339*, 607–33.

[91] M. Stahl, C. Taroni, G. Schneider, *Protein Eng.* **2000**, *13*, 83–88.

[92] K. Kupas, A. Ultsch, G. Klebe, *Proteins* **2008**, *71*, 1288–1306.

[93] A. Kahraman, R. J. Morris, R. A. Laskowski, J. M. Thornton, *J. Mol. Biol.* **2007**, *368*, 283–301.

[94] M. Nayal, B. Honig, *Proteins* **2006**, *63*, 892–906.

[95] J. Mestres, G. Maggiora, *J. Math. Chem.* **2006**, *39*, 107–118.

[96] J. R. Davies, R. M. Jackson, K. V. Mardia, C. C. Taylor, *Bioinformatics* **2007**, btm470.

[97] R. Powers, J. C. Copeland, K. Germer, K. A. Mercier, V. Ramanathan, P. Revesz, *Proteins* **2006**, *65*, 124–135.

[98] C. Schalon, J.-S. Surgand, E. Kellenberger, D. Rognan, *Proteins* **2008**, *71*, 1755–1778.

[99] T. A. Binkowski, L. Adamian, J. Liang, *J. Mol. Biol.* **2003**, *332*, 505–526.

[100] N. D. Gold, R. M. Jackson, *J. Chem. Inf. Model.* **2006**, *46*, 736–742.

[101] M. Gao, J. Skolnick, *Bioinformatics* **2013**, *29*, 597–604.

[102] K. Kinoshita, H. Nakamura, *Protein Sci.* **2005**, *14*, 711–718.

[103] B. K. P. Horn, *J. Opt. Soc. Am. A* **1987**, *4*, 629–642.

[104] M. Rosen, S. L. Lin, H. Wolfson, R. Nussinov, *Protein Eng.* **1998**, *11*, 263–277.

[105] O. Bachar, D. Fischer, R. Nussinov, H. Wolfson, *Protein Eng.* **1993**, *6*, 279–288.

[106] X. Pennec, N. Ayache, *Bioinformatics* **1998**, *14*, 516–522.

[107] E. J. Gardiner, P. J. Artymiuk, P. Willett, *J. Mol. Graph. Model.* **1997**, *15*, 245–253.

[108] C. Bron, J. Kerbosch, *Commun. ACM* **1973**, *16*, 577, 575.

[109] N. Weskamp, D. Kuhn, E. Hüllermeier, G. Klebe, *Bioinformatics* **2004**, *20*, 1522–1526.

[110] L. Sael, D. La, B. Li, R. Rustamov, D. Kihara, *Proteins* **2008**, *73*, 1–10.

[111] D. La, J. Esquivel-Rodríguez, V. Venkatraman, B. Li, L. Sael, S. Ueng, S. Ahrendt, D. Kihara, *Bioinformatics* **2009**, *25*, 2843–2844.

[112] R. Chikhi, L. Sael, D. Kihara, *Proteins* **2010**, *78*, 2007–2028.

[113] C. D. Andersson, B. Y. Chen, A. Linusson, *Proteins* **2010**, *78*, 1408–1422.

[114] J. S. Mason, A. C. Good, E. J. Martin, *Curr. Pharm. Des.* **2001**, *7*, 567–97.

[115] N. Weill, D. Rognan, *J. Chem. Inf. Model.* **2010**, *50*, 123–135.

[116] J.-I. Ito, Y. Tabei, K. Shimizu, K. Tomii, K. Tsuda, *Proteins* **2012**, *80*, 747–763.

[117] T. Kalliokoski, T. S. G. Olsson, A. Vulpetti, *J. Chem. Inf. Model.* **2013**, *53*, 131–141.

[118] K. Kinoshita, H. Nakamura, *Protein Sci.* **2003**, *12*, 1589–1595.

[119] D. Kuhn, N. Weskamp, S. Schmitt, E. Hüllermeier, G. Klebe, *J. Mol. Biol.* **2006**, *359*, 1023–44.

[120] M. Milik, S. Szalma, K. A. Olszewski, *Protein Eng.* **2003**, *16*, 543–552.

[121] T. A. Binkowski, P. Freeman, J. Liang, *Nucleic Acids Res.* **2004**, *32*, W555–W558.

[122] N. D. Gold, R. M. Jackson, *J. Mol. Biol.* **2006**, *355*, 1112–24.

[123] F. Ferrè, G. Ausiello, A. Zanzoni, M. Helmer-Citterich, *Nucleic Acids Res.* **2004**, *32*, D240–244.

[124] M. Shatsky, A. Shulman-Peleg, R. Nussinov, H. J. Wolfson, *J. Comput. Biol.* **2006**, *13*, 407–428.

[125] K. Park, D. Kim, *Genome Inform.* **2006**, *17*, 216–225.

[126] K. Park, D. Kim, *Proteins* **2008**, *71*, 960–71.

[127] Z. Zhang, M. G. Grigorov, *Proteins* **2006**, *62*, 470–478.

[128] V. Ramensky, A. Sobol, N. Zaitseva, A. Rubinov, V. Zosimov, *Proteins* **2007**, *69*, 349–357.

[129] Z. Aung, J. C. Tong, *Genome Inform.* **2008**, *21*, 65–76.

[130] K. Yeturu, N. Chandra, *BMC Bioinf.* **2008**, *9*, 543.

[131] R. Minai, Y. Matsuo, H. Onuki, H. Hirota, *Proteins* **2008**, *72*, 367–381.

[132] L. Xie, P. E. Bourne, *BMC Bioinf.* **2007**, *8 Suppl 4*, S9.

[133] L. Xie, P. E. Bourne, *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 5441–5446.

[134] T. A. Binkowski, A. Joachimiak, *BMC Struct. Biol.* **2008**, *8*, 45.

[135] F. Moriaud, O. Doppelt-Azeroual, L. Martin, K. Oguievetskaia, K. Koch, A. Vorotyntsev, S. A. Adcock, F. Delfaud, *J. Chem. Inf. Model.* **2009**, *49*, 280–294.

[136] S. Das, A. Kokardekar, C. M. Breneman, *J. Chem. Inf. Model.* **2009**, *49*, 2863–2872.

[137] A. McGready, A. Stevens, M. Lipkin, B. D. Hudson, D. C. Whitley, M. G. Ford, *J. Mol. Model.* **2009**, *15*, 489–498.

[138] S. Yin, E. A. Proctor, A. A. Lugovskoy, N. V. Dokholyan, *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 16622–16626.

[139] B. Xiong, J. Wu, D. Burk, M. Xue, H. Jiang, J. Shen, *BMC Bioinf.* **2010**, *11*, 47.

[140] T. Liu, R. B. Altman, *PLoS Comput Biol* **2011**, *7*, e1002326.

[141] M. Gao, J. Skolnick, *Bioinformatics* **2010**, *26*, 2259–2265.

[142] M. M. von Behren, A. Volkamer, A. M. Henzler, K. T. Schomburg, S. Urbaczek, M. Rarey, *J. Chem. Inf. Model.* **2013**, *53*, 411–422.

[143] P. Cozzini, G. E. Kellogg, F. Spyrakis, D. J. Abraham, G. Costantino, A. Emerson, F. Fanelli, H. Gohlke, L. A. Kuhn, G. M. Morris, M. Orozco, T. A. Pertinhez, M. Rizzi, C. A. Sotriffer, *J. Med. Chem.* **2008**, *51*, 6237–6255.

[144] S. J. Teague, *Nat. Rev. Drug Discovery* **2003**, *2*, 527–541.

[145] A. Volkamer, A. Griewel, T. Grombacher, M. Rarey, *J. Chem. Inf. Model.* **2010**, *50*, 2041–2052.

[146] X. Q. Lewell, D. B. Judd, S. P. Watson, M. M. Hann, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511–522.

[147] M. Wagener, J. P. M. Lommerse, *J. Chem. Inf. Model.* **2006**, *46*, 677–685.

[148] I. Wallach, R. H. Lilien, *Bioinformatics* **2009**, *25*, i296–304.

[149] A. W. E. Chan, R. A. Laskowski, D. L. Selwood, *J. Med. Chem.* **2010**, *53*, 3086–3094.

[150] L. Wang, Z. Xie, P. Wipf, X.-Q. Xie, *J. Chem. Inf. Model.* **2011**, *51*, 807–815.

[151] S. Soga, H. Shirai, M. Kobori, N. Hirayama, *J. Chem. Inf. Model.* **2008**, *48*, 1679–85.

[152] J. D. Durrant, A. J. Friedman, J. A. McCammon, *J. Chem. Inf. Model.* **2011**, *51*, 2573–2580.

[153] M. J. Keiser, B. L. Roth, B. N. Armbruster, P. Ernsberger, J. J. Irwin, B. K. Shoichet, *Nat. Biotechnol.* **2007**, *25*, 197–206.

[154] M. Campillos, M. Kuhn, A.-C. Gavin, L. J. Jensen, P. Bork, *Science* **2008**, *321*, 263–266.

[155] M. J. Keiser, V. Setola, J. J. Irwin, C. Laggner, A. I. Abbas, S. J. Hufeisen, N. H. Jensen, M. B. Kuijer, R. C. Matos, T. B. Tran, R. Whaley, R. A. Glennon, J. Hert, K. L. H. Thomas, D. D. Edwards, B. K. Shoichet, B. L. Roth, *Nature* **2009**, *462*, 175–181.

[156] C. Moneriz, J. Mestres, J. M. Bautista, A. Diez, A. Puyet, *FEBS J.* **2011**, *278*, 2951–2961.

[157] J. Mestres, S. A. Seifert, T. I. Oprea, *Clin. Pharmacol. Ther.* **2011**, *90*, 662–665.

[158] A. Schlessinger, E. Geier, H. Fan, J. J. Irwin, B. K. Shoichet, K. M. Giacomini, A. Sali, *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 15810–15815.

[159] E. Lounkine, M. J. Keiser, S. Whitebread, D. Mikhailov, J. Hamon, J. L. Jenkins, P. Lavan, E. Weber, A. K. Doak, S. Côté, B. K. Shoichet, L. Urban, *Nature* **2012**, *486*, 361–367.

[160] E. Gregori-Puigjané, V. Setola, J. Hert, B. A. Crews, J. J. Irwin, E. Lounkine, L. Marnett, B. L. Roth, B. K. Shoichet, *Proc. Natl. Acad. Sci. U. S. A.* **2012**, DOI 10.1073/pnas.1204524109.

[161] A. R. Kinjo, H. Nakamura, *Structure* **2009**, *17*, 234–246.

[162] C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, J. M. Thornton, *Structure* **1997**, *5*, 1093–1108.

[163] L. Lo Conte, S. E. Brenner, T. J. P. Hubbard, C. Chothia, A. G. Murzin, *Nucleic Acids Res.* **2002**, *30*, 264–267.

[164] A. C. R. Martin, *Bioinformatics* **2004**, *20*, 986–988.

[165] P. Rice, I. Longden, A. Bleasby, *Trends Genet.* **2000**, *16*, 276–277.

[166] V. J. Haupt, M. Schroeder, *Briefings Bioinf.* **2011**, *12*, 312–326.

[167] A. Weber, A. Casini, A. Heine, D. Kuhn, C. T. Supuran, A. Scozzafava, G. Klebe, *J. Med. Chem.* **2004**, *47*, 550–557.

[168] L. Xie, J. Wang, P. E. Bourne, *PLoS Comput. Biol.* **2007**, *3*, e217.

[169] S. L. Kinnings, N. Liu, N. Buchmeier, P. J. Tonge, L. Xie, P. E. Bourne, *PLoS Comput. Biol.* **2009**, *5*, DOI 10.1371/journal.pcbi.1000423.

[170] J. D. Durrant, R. E. Amaro, L. Xie, M. D. Urbaniak, M. A. J. Ferguson, A. Haapalainen, Z. Chen, A. M. Di Guilmi, F. Wunder, P. E. Bourne, J. A. McCammon, *PLoS Comput. Biol.* **2010**, *6*, e1000648.

[171] E. Defranchi, E. De Franchi, C. Schalon, M. Messa, F. Onofri, F. Benfenati, D. Rognan, *PLoS ONE* **2010**, *5*, e12214.

[172] B. Stauch, H. Hofmann, M. Perkovic, M. Weisel, F. Kopietz, K. Cichutek, C. Münk, G. Schneider, *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 12079–12084.

[173] D. Kuhn, N. Weskamp, E. Hullermeier, G. Klebe, *ChemMedChem* **2007**, *2*, 1432–1447.

[174] T. Kalliokoski, A. Vulpetti, *Mol. Inf.* **2011**, *30*, 923–925.

[175] S. L. Kinnings, R. M. Jackson, *J. Chem. Inf. Model.* **2009**, DOI 10.1021/ci800289y.

[176] F. Milletti, A. Vulpetti, *J. Chem. Inf. Model.* **2010**, *50*, 1418–1431.

[177] M. Brylinski, J. Skolnick, *Mol. Pharm.* **2010**, *7*, 2324–2333.

III.3: Chemoisosterism in the proteome

Jalencas, X.; Mestres, J. Chemoisosterism in the Proteome.
J. Chem. Inf. Model. **2013**, 53, 279-292

Journal Impact Factor: 4.675

The tool for binding site characterization and comparison detailed in Chapter III.1 was tailored in this paper by a fragment-based approach with the aim of gaining ability to find similarities in unrelated proteins. The new approach was used in this paper to analyse protein environments binding chemical fragments. The term "chemoisosterism" was coined to describe the phenomenon of different environments binding the same chemical fragments. Some examples were provided on possible applications of chemoisosterism to drug design, focusing in the target PPAR-γ. Its complementarity to ligand-based methods and other structure-based methods like docking revealed chemoisosterism as a valuable tool for drug design projects.

A poster and an oral communication were also presented on this topic.

- Jalencas, X.; Mestres, J. A knowledge-based approach to assessing the target promiscuity of chemical fragments. Oral communication presented at 9th International Conference on Chemical Structures. 2011 Jun 5-9. Noordwijkerhout. Netherlands.

- Jalencas, X.; Mestres, J. Indexing cavities in protein structures. Poster presented at 21st International Symposium on Medicinal Chemistry. 2010 Sep 5-9. Brussels. Belgium

III.4: On the Origins of Drug Polypharmacology

Jalencas, X.; Mestres, J. On the origins of drug polypharmacology. Med. Chem. Comm. **2013**, 4, 80-87.

Journal Impact Factor: 2.8; Citations: 5

After exploring chemoisosterism and its ability to detect suitable chemical fragments for a particular target of know structure in **Chapter III.3**, its implications to drug polypharmacology are discussed in this section. A wide-ranging review of drug polypharmacology is elaborated, focusing on all possible sources for this phenomenon, including, but not limited to chemoisosterism. Special interest has been put in the implications that drug polypharmacology has in drug discovery and some hypothesis about its evolutionary origin are provided.

Part IV: Discussion

IV.1 Comparing binding sites

In this Thesis, I have pursued the main objective of developing a novel structure-based methodology based on protein binding site comparisons that is able to provide valuable information on the polypharmacology of drugs.

Comparison of binding sites was a hot and promising topic at the beginning of this Thesis. Several highly cited papers on the field had been published in the preceding years[97–101] and, for the first time with physicochemical properties of the binding site were starting to be used alongside with shape in binding site comparisons. Although at that time it was a time-consuming approach, not suitable for large-scale screenings, it was nonetheless able to unveil similarities between binding sites where no other approaches succeeded[102,103], even being able to eventually explain unexpected protein-ligand relationships.[104]

Taking as a starting point the state of the art on binding site comparison methodologies, a novel approach has been developed and implemented (Chapter III.I). The main differences against other existing methods lie in the three-dimensional location of the descriptors, which are placed directly on the protein surface and encode for its local pharmacophoric properties. In contrast to other approaches that only consider the alpha carbons of the residues defining the binding cavity, the utilization of such feature surface points removes the need of a pair of binding sites to have residues in equivalent positions to be considered as similar. As long as the physicochemical properties they expose in the surface are equivalent and available in equivalent directions, similarities can still be detected irrespectively of the coordinates of the protein atoms.[105] As a counterpart, more complexity is incorporated in the comparisons, and its sensitivity to protein flexibility is also increased.[106] This apparent limitation has been addressed by implementing a fragmental approach to binding site comparisons, which partially alleviates this conformational sensitivity.

The methodology has been demonstrated to successfully being able to discriminate between binding sites from different proteins, as well as subgroups within a protein family. Unfortunately, like other methods, its ability to detect similarities between unrelated protein binding sites has proved to be limited when full protein cavities are considered. As similarities between related proteins can be established in most cases by other more efficient approaches such as sequence, fold or amino acid pattern comparisons[107,108], a better power to detect similarities in unrelated binding sites is desirable as it provides new data that cannot be obtained by any other means.

The bibliography and expertise gathered during this Thesis has been further exploited for the elaboration of a comprehensive an updated review on the topic (Chapter III.2). Special attention has been put in the selected cases where a new protein-ligand interaction has been inferred from a similarity found between binding sites of different proteins, and even more, this new interaction could be subsequently confirmed experimentally. Surprisingly, despite the large number of existing methodologies to compare binding sites and the wide range of approaches they use, only a handful of such cases exist. Locating unexpected similarities in unrelated proteins that can be further exploited to predict new protein ligand interactions is by no means a simple task.

IV.2 Chemoisosterism

To gain capacity in detecting similarities between unrelated proteins that lead to the identification of new targets for a particular ligand (or new ligands for a particular target) beyond the members of a protein family, a fragment-based approach has been developed to address the main issues associated with binding site comparisons highlighted above. As protein cavities binding similar or even the same ligands are rarely found to be similar if they are not related by sequence or fold, attention was put in detecting local similarities.[109]

Consequently, binding sites are decomposed in protein environments and ligands in chemical fragments (Chapter III.3).

We early noted that when comparing protein environments, most chemical fragments (being the benzene ring the most extreme case) appeared to be found co-crystallised in different protein environments that could not be related by similarity. Accordingly, the term chemoisosterism was coined during the development of this Thesis to describe the ability of different protein environments to be compatible with the same chemical fragment. The term "chemoisosterism" is inspired by shifting the point of view of bioisosterism (that describes chemical fragments that can bind the same protein environment) from the protein (bio) to the chemical (chemo) perspective.

To exploit protein environment similarities to its maximum extent, a database of chemoisosteric environments was built. Given a set of structures of protein-ligand complexes, unique protein environments are associated to their corresponding chemical fragments. The construction of such a database for the PDBbind[66] data set is detailed in Chapter III.3. As the process of building a database of chemoisosteric environments is completely automatic, specific databases focused in a particular group of protein structures can be easily obtained to better adapt to the needs of a particular project.

Applications of chemoisosterism in drug design are thoroughly described in Chapter III.3. Some retrospective applications are shown using as example Peroxisome Proliferator-Aactivated Receptor gamma (PPARγ), a member of the nuclear receptor family.[110] Remarkably, we were able to reconstruct most of its ligand chemical fragments based solely on chemical fragments that are bound to similar protein environments in enzymes structures. It is worth stressing here the level of protein environment hopping achieved, between largely distinct protein families such as nuclear receptors and enzymes. This

would allow applying the same technique even in the hypothetical case where no other nuclear receptors had any available structure.

Simultaneously, chemoisosteric relationships have been used to construct a focused chemical library[111] of suitable chemical fragments for PPARγ. Many of those fragments could be found in molecules known to be active against PAPRγ, but with no available resolved structure. Interestingly, the composition of the fragment library was compared against ensembles of fragments obtained using other approaches, namely, ligand-based similarity[22] and a structure-based docking[112], and it was found that the overlap of the results obtained by the three methods was very small. This strongly supports chemoisosterism as a tool that is able to provide relevant, complementary, information that would probably not be obtained by any other means. Chemoisosterism can therefore be of utility to complement other widely used methods and even more, in cases where the problem to address lies out of their applicability domain but has a resolved protein structure available.

IV.3 Polypharmacology

One of the sought endeavours when screening for similarities between protein binding sites is the discovery of new targets for already known ligands. The ability of small molecules to interact with multiple protein targets is usually referred to as polypharmacology.[75] An updated overview of drug polypharmacology has been addressed in Chapter III.4, putting special stress on its potential causes. Sources of polypharmacology can be primarily described as chemical or biological. Chemical sources of polypharmacology include molecular properties and fragment composition. In this respect, the experience and expertise on chemical fragments gained in the tasks described in Chapter III.3 were exploited to find a correlation between the complexity of chemical

fragments and their protein promiscuity (the number of targets that bind that fragment). Biological sources of polypharmacology, including target phylogeny and binding site similarities, have been equally addressed. Overall, the main conclusion drawn from this analysis is that there is still a highly conservative view of drug polypharmacology, mainly attributable to the lack of completeness of drug-target interaction data.[113]

IV.4 Future directions of research

Several different perspectives arise at different points of this Thesis. To start with, binding site signatures presented in Chapter III.1 have many potential utilities that are yet to be exploited. If comparing signatures was demonstrated to be equivalent to compare the binding sites themselves, the complexity of the comparison would be greatly reduced, resulting in considerable savings of time and resources. Under a chemogenomics perspective, binding site descriptors can be attached to similar ligand descriptors to be able to directly predict if a particular ligand has a probability to bind to a given binding site.[114,115] In the present case, feature surface descriptors used for binding sites could be adapted to describe ligand surfaces. Complementarity between binding sites and ligands could be then estimated, and hence their interaction predicted. Predicting protein-ligand interaction is a challenging task with many pitfalls.[116] Protein, and specially, ligand flexibility issues would be need to be solved. Even though, a similar approach could be also used to refine docking poses[117] or homology models.

Regarding the potential applications of chemoisosterism in drug discovery presented in Chapter III.3, here a wide array of possibilities emerges. First and foremost, its successful application in a real drug design project would definitely validate the used approach and provide insight on its real value besides the retrospective validations that have been performed. Chemoisosterism has been

demonstrated to be useful for predicting chemical fragments for a protein binding site, but many variations of this can be explored. By applying chemoisosterism under similarity constraints, an empty binding site with no co-crystallized ligand can be populated with putative chemical fragments by comparing it against a set of protein environments and chemical fragments. In this case, a scoring scheme penalizing less the differences in size of the compared binding sites than the described cosine score would be desirable. Once a binding site is populated with chemical fragments, another interesting line worth to be explored is the assembly and linking of such fragments to complete molecules.[118] This *de novo* drug design based on fragments[119] would provide an easy-to-interpret and more direct output of molecules that chemical fragments alone. This approach, due to its hopping abilities, would be especially useful in proteins with little structural information, such as GPCRs.

In a similar way, the structure of a complex of a given molecule can be used to locate other binding sites compatible with its chemical fragments and therefore predict a new target for the molecule, what is commonly known as target profiling.[120] Difficulties here are expected to be found regarding the fact that some chemical fragments are found to bind to many different proteins, and unless a very discriminative chemical fragment appears, it can be difficult to predict the targets for a given set of chemical fragments merely based on an enumeration of them. Fragment connectivity and assemblage would probably need to be addressed also in this case. It is worth noting that this feature would be of special interest in drug repurposing.[121,122] Many structures in the Protein Data Bank containing drugs could be used as a starting point.

Finally, being a knowledge-based approach, chemoisosterism relies on the structural data that is exploited. In this direction, a complete database of all chemical fragments and protein environments available in the Protein Data Bank would greatly improve both its applicability domain as well as its hopping ability and precision. Even more, a collection of all putative protein pockets

extracted from the PDB would also provide a valuable source of information. With around 90,000 structures in the PDB, the construction of such databases would require great computational and storage resources. Efforts in this direction are currently underway in our research group and will hopefully lead to an interesting continuation of the work presented in this Thesis.

Part V: Conclusions

The main contributions of this Thesis can be summarized as follows:

i) An updated and exhaustive review on available computational binding site comparison methodologies has been performed putting special stress on the successful cases where the detection of a similarity between binding sites led to the identification of a new case of distant polypharmacology that could be experimentally tested.

ii) A novel binding site comparison methodology based on surface feature points has been devised developed and implemented.

iii) The developed software has been demonstrated to be able to successfully discriminate binding sites from different proteins, achieving a similar degree of performance as other existing methodologies. A finer discrimination, exemplified by the classification of different kinase types and carbonic anhydrase isoforms has also been achieved.

iv) A fragment-based approach involving protein environments and chemical fragments has been implemented to gain predictive power in the difficult task of locating similarities between binding sites of completely unrelated proteins.

v) The term "chemoisosterism" has been coined as a counterpart to bioisosterism, to describe protein environments which are compatible with the same chemical fragment.

vi) A database of chemoisosteric protein environments related to their compatible chemical fragments has been constructed including 16,533 different chemical fragments linked 73,931 different protein environments in 87,711 interactions.

vii) Applicability of chemoisosterism in drug design processes and its value in complementing other existing methodologies has been illustrated by retrospective analysis concerning the PPARγ receptor. Most of the fragments of a PPARγ native ligand have been reconstructed by hopping fragments from enzyme protein environments.

viii) A focused library of chemical fragments for PPARγ has been constructed. Its contents are supported by the presence of such fragments in known PPARγ inhibitors.

ix) The origins of drug polypharmacology are yet to be clearly understood. We propose that the levels of polypharmacology observed in current drugs may just be a latent signature of the exploitation of chemoisosterism during evolution.

References

(1) DiMasi, J. A.; Hansen, R. W.; Grabowski, H. G. The price of innovation: new estimates of drug development costs. *J. Health. Econ.* **2003**, *22*, 151–185.

(2) Adams, C. P.; Brantner, V. V. Spending on new drug development. *Health Econ.* **2010**, *19*, 130–141.

(3) Paul, S. M.; Mytelka, D. S.; Dunwiddie, C. T.; Persinger, C. C.; Munos, B. H.; Lindborg, S. R.; Schacht, A. L. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat. Rev. Drug Discovery* **2010**, *9*, 203–214.

(4) Kola, I.; Landis, J. Can the pharmaceutical industry reduce attrition rates? *Nat. Rev. Drug Discovery* **2004**, *3*, 711–716.

(5) Kodadek, T. The rise, fall and reinvention of combinatorial chemistry. *Chem. Commun.* **2011**, *47*, 9757–9763.

(6) Bibette, J. Gaining confidence in high-throughput screening. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 649–650.

(7) Shendure, J.; Lieberman Aiden, E. The expanding scope of DNA sequencing. *Nat. Biotechnol.* **2012**, *30*, 1084–1094.

(8) Scannell, J. W.; Blanckley, A.; Boldon, H.; Warrington, B. Diagnosing the decline in pharmaceutical R&D efficiency. *Nat. Rev. Drug Discovery* **2012**, *11*, 191–200.

(9) Li, Q.; Cheng, T.; Wang, Y.; Bryant, S. H. PubChem as a public resource for drug discovery. *Drug Discovery Today* **2010**, *15*, 1052–1057.

(10) Knox, C.; Law, V.; Jewison, T.; Liu, P.; Ly, S.; Frolkis, A.; Pon, A.; Banco, K.; Mak, C.; Neveu, V.; Djoumbou, Y.; Eisner, R.; Guo, A. C.; Wishart, D. S. DrugBank 3.0: a comprehensive resource for "omics" research on drugs. *Nucleic Acids Res.* **2011**, *39*, D1035–1041.

(11) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40*, D1100–1107.

(12) Ekins, S.; Mestres, J.; Testa, B. In silico pharmacology for drug discovery: methods for virtual ligand screening and profiling. *Br. J. Pharmacol.* **2007**, *152*, 9–20.

(13) Ekins, S.; Mestres, J.; Testa, B. In silico pharmacology for drug discovery: applications to targets and beyond. *Br. J. Pharmacol.* **2007**, *152*, 21–37.

(14) Keiser, M. J.; Setola, V.; Irwin, J. J.; Laggner, C.; Abbas, A. I.; Hufeisen, S. J.; Jensen, N. H.; Kuijer, M. B.; Matos, R. C.; Tran, T. B.; Whaley, R.; Glennon, R. A.; Hert, J.; Thomas, K. L. H.; Edwards, D. D.; Shoichet, B. K.; Roth, B. L. Predicting new molecular targets for known drugs. *Nature* **2009**, *462*, 175–181.

(15) DeGraw, A. J.; Keiser, M. J.; Ochocki, J. D.; Shoichet, B. K.; Distefano, M. D. Prediction and evaluation of protein farnesyltransferase inhibition by commercial drugs. *J. Med. Chem.* **2010**, *53*, 2464–2471.

(16) Mestres, J.; Seifert, S. A.; Oprea, T. I. Linking pharmacology to clinical reports: cyclobenzaprine and its possible association with serotonin syndrome. *Clin. Pharmacol. Ther.* **2011**, *90*, 662–665.

(17) Moneriz, C.; Mestres, J.; Bautista, J. M.; Diez, A.; Puyet, A. Multi-targeted activity of maslinic acid as an antimalarial natural compound. *FEBS J.* **2011**, *278*, 2951–2961.

(18) Antolín, A. A.; Jalencas, X.; Yélamos, J.; Mestres, J. Identification of Pim Kinases as Novel Targets for PJ34 with Confounding Effects in PARP Biology. *ACS Chem. Biol.* **2012**, *7*, 1962–1967.

(19) Gregori-Puigjané, E.; Setola, V.; Hert, J.; Crews, B. A.; Irwin, J. J.; Lounkine, E.; Marnett, L.; Roth, B. L.; Shoichet, B. K. Identifying mechanism-of-action targets for drugs and probes. *Proc. Natl. Acad. Sci. U. S. A.* **2012**.

(20) Lounkine, E.; Keiser, M. J.; Whitebread, S.; Mikhailov, D.; Hamon, J.; Jenkins, J. L.; Lavan, P.; Weber, E.; Doak, A. K.; Côté, S.; Shoichet, B. K.; Urban, L. Large-scale prediction and testing of drug activity on side-effect targets. *Nature* **2012**, *486*, 361–367.

(21) Vidal, D.; Mestres, J. In Silico Receptorome Screening of Antipsychotic Drugs. *Mol. Inf.* **2010**, *29*, 543–551.

(22) Vidal, D.; Garcia-Serna, R.; Mestres, J. Ligand-based approaches to in silico pharmacology. *Methods Mol. Biol.* **2011**, *672*, 489–502.

(23) Bender, A.; Glen, R. C. Molecular similarity: a key technique in molecular informatics. *Org. Biomol. Chem.* **2004**, *2*, 3204–3218.

(24) Eckert, H.; Bajorath, J. Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug Discov. Today* **2007**, *12*, 225–233.

(51) Venter, J. C. *et al.,* The sequence of the human genome. *Science* **2001**, *291*, 1304–1351.

(26) Hopkins, A. L.; Groom, C. R. The druggable genome. *Nat. Rev. Drug Discovery* **2002**, *1*, 727–730.

(27) Russ, A. P.; Lampel, S. The druggable genome: an update. *Drug Discovery Today* **2005**, *10*, 1607–1610.

(28) Paolini, G. V.; Shapland, R. H. B.; Hoorn, W. P. van; Mason, J. S.; Hopkins, A. L. Global mapping of pharmacological space. *Nat. Biotechnol.* **2006**, *24*, 805–815.

(29) Roth, B. L.; Sheffler, D. J.; Kroeze, W. K. Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia. *Nat. Rev. Drug Discovery* **2004**, *3*, 353–359.

(30) Roth, B. L. Drugs and valvular heart disease. *N. Engl. J. Med.* **2007**, *356*, 6–9.

(31) Caron, P. R.; Mullican, M. D.; Mashal, R. D.; Wilson, K. P.; Su, M. S.; Murcko, M. A. Chemogenomic approaches to drug discovery. *Curr Opin Chem Biol* **2001**, *5*, 464–470.

(32) Harris, C. J.; Stevens, A. P. Chemogenomics: structuring the drug discovery process to gene families. *Drug Discovery Today* **2006**, *11*, 880–888.

(33) Mestres, J. Computational chemogenomics approaches to systematic knowledge-based drug discovery. *Curr Opin Drug Discov Devel* **2004**, *7*, 304–313.

(34) Rognan, D. Chemogenomic approaches to rational drug design. *Br. J. Pharmacol.* **2007**, *152*, 38–52.

(35) Cases, M.; Mestres, J. A chemogenomic approach to drug discovery: focus on cardiovascular diseases. *Drug Discovery Today* **2009**, *14*, 479–485.

(36) Magrane, M.; Consortium, U. UniProt Knowledgebase: a hub of integrated protein data. *Database* **2011**, *2011*, bar009.

(37) Artimo, P.; Jonnalagedda, M.; Arnold, K.; Baratin, D.; Csardi, G.; de Castro, E.; Duvaud, S.; Flegel, V.; Fortier, A.; Gasteiger, E.; Grosdidier, A.; Hernandez, C.; Ioannidis, V.; Kuznetsov, D.; Liechti, R.; Moretti, S.; Mostaguir, K.; Redaschi, N.; Rossier, G.; Xenarios, I.; Stockinger, H. ExPASy: SIB bioinformatics resource portal. *Nucleic Acids Res.* **2012**, *40*, W597–603.

(38) Kotera, M.; Hirakawa, M.; Tokimatsu, T.; Goto, S.; Kanehisa, M. The KEGG databases and tools facilitating omics analysis: latest developments involving human diseases and pharmaceuticals. *Methods Mol. Biol.* **2012**, *802*, 19–39.

(39) Punta, M.; Coggill, P. C.; Eberhardt, R. Y.; Mistry, J.; Tate, J.; Boursnell, C.; Pang, N.; Forslund, K.; Ceric, G.; Clements, J.; Heger, A.; Holm, L.; Sonnhammer, E. L. L.; Eddy, S. R.; Bateman, A.; Finn, R. D. The Pfam protein families database. *Nucleic Acids Res.* **2012**, *40*, D290–301.

(40) Orengo, C. A.; Michie, A. D.; Jones, S.; Jones, D. T.; Swindells, M. B.; Thornton, J. M. CATH--a hierarchic classification of protein domain structures. *Structure* **1997**, *5*, 1093–1108.

(41) Lo Conte, L.; Brenner, S. E.; Hubbard, T. J. P.; Chothia, C.; Murzin, A. G. SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res.* **2002**, *30*, 264–267.

(42) Pieper, U.; Eswar, N.; Webb, B. M.; Eramian, D.; Kelly, L.; Barkan, D. T.; Carter, H.; Mankoo, P.; Karchin, R.; Marti-Renom, M. A.; Davis, F. P.; Sali, A. modbase, a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res* **2009**, *37*, D347–D354.

(43) Gregori-Puigjané, E.; Garriga-Sust, R.; Mestres, J. Indexing molecules with chemical graph identifiers. *J. Comput. Chem.* **2011**, *32*, 2638–2646.

(44) Heller, S.; McNaught, A.; Stein, S.; Tchekhovskoi, D.; Pletnev, I. InChI - the worldwide chemical structure identifier standard. *J. Cheminf.* **2013**, *5*, 7.

(45) Ferenczy, G. G.; Keseru, G. M. Chapter 2: Thermodynamics of Ligand Binding. In *Physico-Chemical and Computational Approaches to Drug Discovery*; Luque, J.; Barril, X., Eds.; Royal Society of Chemistry, 2012.

(46) Bissantz, C.; Kuhn, B.; Stahl, M. A Medicinal Chemist's Guide to Molecular Interactions. *J. Med. Chem.* **2010**, *53*, 5061–5084.

(47) Ermondi, G.; Caron, G. Recognition forces in ligand-protein complexes: blending information from different sources. *Biochem. Pharmacol.* **2006**, *72*, 1633–1645.

(48) Waters, M. L. Aromatic interactions in model systems. *Curr. Opin. Chem. Biol.* **2002**, *6*, 736–741.

(49) Baum, B.; Muley, L.; Smolinski, M.; Heine, A.; Hangauer, D.; Klebe, G. Non-additivity of functional group contributions in protein-ligand binding: a comprehensive study by crystallography and isothermal titration calorimetry. *J. Mol. Biol.* **2010**, *397*, 1042–1054.

(50) Stierand, K.; Rarey, M. Drawing the PDB: Protein−Ligand Complexes in Two Dimensions. *ACS Med. Chem. Lett.* **2010**, *1*, 540–545.

(51) Koshland, D. E. Application of a Theory of Enzyme Specificity to Protein Synthesis. *Proc. Natl. Acad. Sci. U.S.A.* **1958**, *44*, 98–104.

(52) Tsai, C. J.; Kumar, S.; Ma, B.; Nussinov, R. Folding funnels, binding funnels, and protein function. *Protein Sci.* **1999**, *8*, 1181–1190.

(53) Exner, T. E.; Keil, M.; Brickmann, J. Pattern recognition strategies for molecular surfaces. I. Pattern generation using fuzzy set theory. *J. Comput. Chem.* **2002**, *23*, 1176–1187.

(54) Rose, P. W.; Beran, B.; Bi, C.; Bluhm, W. F.; Dimitropoulos, D.; Goodsell, D. S.; Prlic, A.; Quesada, M.; Quinn, G. B.; Westbrook, J. D.; Young, J.; Yukich, B.; Zardecki, C.; Berman, H. M.; Bourne, P. E. The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic Acids Res.* **2011**, *39*, D392–401.

(55) Bruno, I. J.; Cole, J. C.; Lommerse, J. P.; Rowland, R. S.; Taylor, R.; Verdonk, M. L. IsoStar: a library of information about nonbonded interactions. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 525–537.

(56) Hendlich, M.; Bergner, A.; Günther, J.; Klebe, G. Relibase: design and development of a database for comprehensive analysis of protein-ligand interactions. *J. Mol. Biol.* **2003**, *326*, 607–620.

(57) Shin, J.-M.; Cho, D.-H. PDB-Ligand: a ligand database based on PDB for the automated and customized classification of ligand-binding structures. *Nucleic Acids Res.* **2005**, *33*, D238–241.

(58) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.* **2007**, *35*, D198–201.

(59) Schreyer, A.; Blundell, T. CREDO: A Protein-Ligand Interaction Database for Drug Discovery. *Chem. Biol. Drug. Des.* **2009**, *73*, 157–167.

(60) Verdonk, M. L.; Cole, J. C.; Taylor, R. SuperStar: a knowledge-based approach for identifying interaction sites in proteins. *J. Mol. Biol.* **1999**, *289*, 1093–1108.

(61) Block, P.; Sotriffer, C. A.; Dramburg, I.; Klebe, G. AffinDB: a freely accessible database of affinities for protein-ligand complexes from the PDB. *Nucleic Acids Res.* **2006**, *34*, D522–526.

(62) Benson, M. L.; Smith, R. D.; Khazanov, N. A.; Dimcheff, B.; Beaver, J.; Dresslar, P.; Nerothin, J.; Carlson, H. A. Binding MOAD, a high-quality protein-ligand database. *Nucleic Acids Res.* **2008**, *36*, D674–678.

(63) Yamaguchi, A.; Iida, K.; Matsui, N.; Tomoda, S.; Yura, K.; Go, M. Het-PDB Navi.: a database for protein-small molecule interactions. *J. Biochem.* **2004**, *135*, 79–84.

(64) Feng, Z.; Chen, L.; Maddula, H.; Akcan, O.; Oughtred, R.; Berman, H. M.; Westbrook, J. Ligand Depot: a data warehouse for ligands bound to macromolecules. *Bioinformatics* **2004**, *20*, 2153–2155.

(65) Stuart, A. C.; Ilyin, V. A.; Sali, A. LigBase: a database of families of aligned ligand binding sites in known protein sequences and structures. *Bioinformatics* **2002**, *18*, 200–201.

(66) Wang, R.; Fang, X.; Lu, Y.; Yang, C.-Y.; Wang, S. The PDBbind database: methodologies and updates. *J. Med. Chem.* **2005**, *48*, 4111–4119.

(67) Kellenberger, E.; Muller, P.; Schalon, C.; Bret, G.; Foata, N.; Rognan, D. sc-PDB: an annotated database of druggable binding sites from the Protein Data Bank. *J. Chem. Inf. Model.* **2006**, *46*, 717–27.

(68) Mestres, J. Representativity of target families in the Protein Data Bank: impact for family-directed structure-based drug discovery. *Drug Discovery Today* **2005**, *10*, 1629–37.

(69) Chen, Y. Z.; Zhi, D. G. Ligand-protein inverse docking and its potential use in the computer search of protein targets of a small molecule. *Proteins* **2001**, *43*, 217–226.

(70) Stark, J. L.; Powers, R. Application of NMR and molecular docking in structure-based drug discovery. *Top. Curr. Chem.* **2012**, *326*, 1–34.

(71) Erlanson, D. A. Fragment-based lead discovery: a chemical update. *Curr. Opin. Biotechnol.* **2006**, *17*, 643–652.

(72) Warr, W. A. Fragment-based drug discovery. *J. Comput.-Aided Mol. Des.* **2009**, *23*, 453–458.

(73) Sun, C.; Petros, A. M.; Hajduk, P. J. Fragment-based lead discovery: challenges and opportunities. *J. Comput.-Aided Mol. Des.* **2011**, *25*, 607–610.

(74) Hann, M. M.; Leach, A. R.; Harper, G. Molecular complexity and its impact on the probability of finding leads for drug discovery. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 856–864.

(75) Jalencas, X.; Mestres, J. On the origins of drug polypharmacology. *Med. Chem. Comm.* **2013**, *4*, 80–87.

(76) Kuntz, I. D.; Chen, K.; Sharp, K. A.; Kollman, P. A. The maximal affinity of ligands. *Proc. Natl. Acad. Sci. U. S. A.* **1999**, *96*, 9997–10002.

(77) Congreve, M.; Carr, R.; Murray, C.; Jhoti, H. A "rule of three" for fragment-based lead discovery? *Drug Discovery Today* **2003**, *8*, 876–877.

(78) Baker, M. Fragment-based lead discovery grows up. *Nat. Rev. Drug Discovery* **2013**, *12*, 5–7.

(79) Marcou, G.; Rognan, D. Optimizing fragment and scaffold docking by use of molecular interaction fingerprints. *J. Chem. Inf. Model.* **2007**, *47*, 195–207.

(80) Kennewell, E. A.; Willett, P.; Ducrot, P.; Luttmann, C. Identification of target-specific bioisosteric fragments from ligand-protein crystallographic data. *J. Comput.-Aided Mol. Des.* **2006**, *20*, 385–394.

(81) Evans, B. E.; Rittle, K. E.; Bock, M. G.; DiPardo, R. M.; Freidinger, R. M.; Whitter, W. L.; Lundell, G. F.; Veber, D. F.; Anderson, P. S.; Chang, R. S. Methods for drug discovery: development of potent, selective, orally effective cholecystokinin antagonists. *J. Med. Chem.* **1988**, *31*, 2235–2246.

(82) Welsch, M. E.; Snyder, S. A.; Stockwell, B. R. Privileged scaffolds for library design and drug discovery. *Curr. Opin. Chem. Biol.* **2010**, *14*, 347–361.

(83) Barelier, S.; Krimm, I. Ligand specificity, privileged substructures and protein druggability from fragment-based screening. *Curr. Opin. Chem. Biol.* **2011**, *15*, 469–474.

(84) DeSimone, R. W.; Currie, K. S.; Mitchell, S. A.; Darrow, J. W.; Pippin, D. A. Privileged structures: applications in drug discovery. *Comb. Chem. High Throughput Screening* **2004**, *7*, 473–494.

(85) Hajduk, P. J.; Bures, M.; Praestgaard, J.; Fesik, S. W. Privileged Molecules for Protein Binding Identified from NMR-Based Screening. *J. Med. Chem.* **2000**, *43*, 3443–3447.

(86) Wagener, M.; Lommerse, J. P. M. The Quest for Bioisosteric Replacements. *J. Chem. Inf. Model.* **2006**, *46*, 677–685.

(87) Southall, N. T.; Ajay Kinase patent space visualization using chemical replacements. *J. Med. Chem.* **2006**, *49*, 2103–2109.

(88) Langmuir, I. ISOMORPHISM, ISOSTERISM AND COVALENCE. *J. Am. Chem. Soc.* **1919**, *41*, 1543–1559.

(89) Patani, G.; LaVoie, E. Bioisosterism: A Rational Approach in Drug Design. *Chem. Rev.* **1996**, *96*, 3147–3176.

(90) Lima, L. M.; Barreiro, E. J. Bioisosterism: a useful strategy for molecular modification and drug design. *Curr. Med. Chem.* **2005**, *12*, 23–49.

(91) Meanwell, N. A. Synopsis of Some Recent Tactical Application of Bioisosteres in Drug Design. *J. Med. Chem.* **2011**, *54*, 2529–2591.

(92) Ujvary, I. BIOSTER - a database of structurally analogous compounds. *Pesticide Science* **1997**, *51*, 92–95.

(93) Birchall, K.; Gillet, V. J.; Willett, P.; Ducrot, P.; Luttmann, C. Use of reduced graphs to encode bioisosterism for similarity-based virtual screening. *J. Chem. Inf. Model.* **2009**, *49*, 1330–1346.

(94) Maggiora, G. M. On outliers and activity cliffs--why QSAR often disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535.

(95) Wassermann, A. M.; Bajorath, J. Chemical substitutions that introduce activity cliffs across different compound classes and biological targets. *J. Chem. Inf. Model.* **2010**, *50*, 1248–1256.

(96) Jalencas, X.; Mestres, J. Chemoisosterism in the Proteome. *J. Chem. Inf. Model.* **2013**, *53*, 279–292.

(97) Kinoshita, K.; Furui, J.; Nakamura, H. Identification of protein functions from a molecular surface database, eF-site. *J. Struct. Funct. Genomics* **2002**, *2*, 9–22.

(98) Schmitt, S.; Kuhn, D.; Klebe, G. A new method to detect related function among proteins independent of sequence and fold homology. *J. Mol. Biol.* **2002**, *323*, 387–406.

(99) Jambon, M.; Imberty, A.; Deléage, G.; Geourjon, C. A new bioinformatic approach to detect common 3D sites in protein structures. *Proteins* **2003**, *52*, 137–145.

(100) Shulman-Peleg, A.; Nussinov, R.; Wolfson, H. J. Recognition of functional sites in protein structures. *J. Mol. Biol.* **2004**, *339*, 607–33.

(101) Gold, N. D.; Jackson, R. M. Fold independent structural comparisons of protein-ligand binding sites for exploring functional relationships. *J. Mol. Biol.* **2006**, *355*, 1112–24.

(102) Ferrè, F.; Ausiello, G.; Zanzoni, A.; Helmer-Citterich, M. Functional annotation by identification of local surface similarities: a novel tool for structural genomics. *BMC Bioinf.* **2005**, *6*, 194.

(103) Gherardini, P. F.; Wass, M. N.; Helmer-Citterich, M.; Sternberg, M. J. E. Convergent evolution of enzyme active sites is not a rare phenomenon. *J. Mol. Biol.* **2007**, *372*, 817–845.

(104) Weber, A.; Casini, A.; Heine, A.; Kuhn, D.; Supuran, C. T.; Scozzafava, A.; Klebe, G. Unexpected nanomolar inhibition of carbonic anhydrase by COX-2-selective celecoxib: new pharmacological opportunities due to related binding site recognition. *J. Med. Chem.* **2004**, *47*, 550–557.

(105) Feldman, H. J.; Labute, P. Pocket Similarity: Are α Carbons Enough? *J. Chem. Inf. Model.* **2010**, *50*, 1466–1475.

(106) Kellenberger, E.; Schalon, C.; Rognan, D. How to Measure the Similarity Between Protein Ligand-Binding Sites? *Curr. Comp. Aided Drug Des.* **2008**, *4*, 209–220.

(107) Holm, L.; Sander, C. Dali/FSSP classification of three-dimensional protein folds. *Nucleic Acids Res.* **1997**, *25*, 231–234.

(108) Stark, A.; Russell, R. B. Annotation in three dimensions. PINTS: Patterns in Non-homologous Tertiary Structures. *Nucleic Acids Res.* **2003**, *31*, 3341–3344.

(109) Moriaud, F.; Doppelt-Azeroual, O.; Martin, L.; Oguievetskaia, K.; Koch, K.; Vorotyntsev, A.; Adcock, S. A.; Delfaud, F. Computational fragment-based approach at PDB scale by protein local similarity. *J. Chem. Inf. Model.* **2009**, *49*, 280–294.

(110) Gampe, R. T., Jr; Montana, V. G.; Lambert, M. H.; Miller, A. B.; Bledsoe, R. K.; Milburn, M. V.; Kliewer, S. A.; Willson, T. M.; Xu, H. E. Asymmetry in the PPARgamma/RXRalpha crystal structure reveals the molecular basis of heterodimerization among nuclear receptors. *Mol. Cell* **2000**, *5*, 545–555.

(111) Gozalbes, R. Rational generation of focused chemical libraries: an update on computational approaches. *Comb. Chem. High Throughput Screen.* **2011**, *14*, 428.

(112) Morris, G. M.; Huey, R.; Olson, A. J. Using AutoDock for ligand-receptor docking. *Curr. Protoc. Bioinformatics* **2008**, *Chapter 8*, Unit 8.14.

(113) Mestres, J.; Gregori-Puigjané, E.; Valverde, S.; Solé, R. V. Data completeness--the Achilles heel of drug-target networks. *Nat. Biotechnol.* **2008**, *26*, 983–4.

(114) Baroni, M.; Cruciani, G.; Sciabola, S.; Perruccio, F.; Mason, J. S. A common reference framework for analyzing/comparing proteins and ligands. Fingerprints for Ligands and Proteins (FLAP): theory and application. *J. Chem. Inf. Model.* **2007**, *47*, 279–94.

(115) Meslamani, J.; Rognan, D. Enhancing the Accuracy of Chemogenomic Models with a Three-Dimensional Binding Site Kernel. *J. Chem. Inf. Model.* **2011**.

(116) Parenti, M. D.; Rastelli, G. Advances and applications of binding affinity prediction methods in drug discovery. *Biotechnol. Adv.* **2012**, *30*, 244–250.

(117) Zsoldos, Z.; Reid, D.; Simon, A.; Sadjad, S. B.; Johnson, A. P. eHiTS: a new fast, exhaustive flexible ligand docking system. *J. Mol. Graph. Model.* **2007**, *26*, 198–212.

(118) Pfeffer, P.; Fober, T.; Hüllermeier, E.; Klebe, G. GARLig: A Fully Automated Tool for Subset Selection of Large Fragment Spaces via a Self-Adaptive Genetic Algorithm. *J. Chem. Inf. Model.* **2010**, *50*, 1644–1659.

(119) Dey, F.; Caflisch, A. Fragment-based de novo ligand design by multiobjective evolutionary optimization. *J. Chem. Inf. Model.* **2008**, *48*, 679–690.

(120) Nonell-Canals, A.; Mestres, J. In Silico Target Profiling of One Billion Molecules. *Mol. Inf.* **2011**, *30*, 405–409.

(121) Moriaud, F.; Richard, S. B.; Adcock, S. A.; Chanas-Martin, L.; Surgand, J.-S.; Ben Jelloul, M.; Delfaud, F. Identify drug repurposing candidates by mining the Protein Data Bank. *Briefings Bioinf.* **2011**, *12*, 336–340.

(122) Oprea, T. I.; Mestres, J. Drug repurposing: far beyond new targets for old drugs. *AAPS J.* **2012**, *14*, 759–763.

Appendix A

Contributions to other publications not included in this Thesis:

Antolín, A. A.; Jalencas, X.; Yélamos, J.; Mestres, J. Identification of Pim Kinases as Novel Targets for PJ34 with Confounding Effects in PARP Biology. ACS Chem. Biol. 2012, 7, 1962–1967.

Journal Impact Factor: 6.446; Citations: 3

ACS chemical biology

Letters

pubs.acs.org/acschemicalbiology

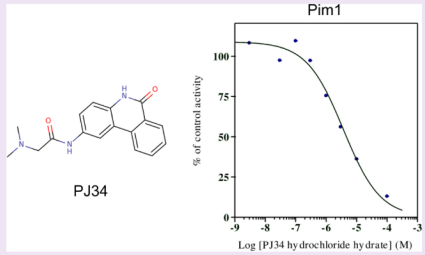
Identification of Pim Kinases as Novel Targets for PJ34 with Confounding Effects in PARP Biology

Albert A. Antolín,[†] Xavier Jalencas,[†] José Yélamos,[‡] and Jordi Mestres*[,†]

[†]Chemogenomics Laboratory, Research Program on Biomedical Informatics and [‡]Department of Immunology, Research Program on Cancer, IMIM Hospital del Mar Research Institute and Universitat Pompeu Fabra, Doctor Aiguader 88, 08003 Barcelona, Catalonia, Spain

Supporting Information

ABSTRACT: Small molecules are widely used in chemical biology without complete knowledge of their target profile, at risk of deriving conclusions that ignore potential confounding effects from unknown off-target interactions. The prediction and further experimental confirmation of novel affinities for PJ34 on Pim1 (IC_{50} = 3.7 μM) and Pim2 (IC_{50} = 16 μM) serine/threonine kinases, together with their involvement in many of the processes relevant to PARP biology, questions the appropriateness of using PJ34 as a chemical tool to probe the biological role of PARP1 and PARP2 at the high micromolar concentrations applied in most studies.



The main contribution to this work consisted on the modelling an in-vitro confirmed interaction between a common PARP chemical probe (PJ34) and the serine/threonine-protein kinase PIM-1. The binding mode for PJ34 to the binding site of PIM-1 was initially predicted through a flexible three-dimensional ligand alignment tool (MIMIC) applied to template PIM-1 structures co-crystallized with different ligands. A docking tool (AutoDock) was subsequently used to confirm and refine the obtained models.

Appendix B

Pharmacophoric properties (HRADPN) and atom types assigned to amino acid atoms. Hydrogen atoms inherit the properties to the atom they are bound.

ALA	N	N.amh	RD
ALA	CA	C.3	H
ALA	C	C.2	HR
ALA	O	O.2	RA
ALA	CB	C.3	H
ARG	N	N.amh	RD
ARG	CA	C.3	H
ARG	C	C.2	HR
ARG	O	O.2	RA
ARG	CB	C.3	H
ARG	CG	C.3	H
ARG	CD	C.3	H
ARG	NE	N.plh	RDP
ARG	CZ	C.cat	R
ARG	NH1	N.plh	RDP
ARG	NH2	N.plh	RDP
ASN	N	N.amh	RD
ASN	CA	C.3	H
ASN	C	C.2	HR
ASN	O	O.2	RA
ASN	CB	C.3	H
ASN	CG	C.2	HR
ASN	OD1	O.2	RA
ASN	ND2	N.amh	RD
ASP	N	N.amh	RD
ASP	CA	C.3	H
ASP	C	C.2	HR
ASP	O	O.2	RA
ASP	CB	C.3	H
ASP	CG	C.2	HR
ASP	OD1	O.co2	RAN
ASP	OD2	O.co2	RAN
CYS	N	N.amh	RD
CYS	CA	C.3	H
CYS	C	C.2	HR
CYS	O	O.2	RA
CYS	CB	C.3	H
CYS	SG	S.3h	AD
GLN	N	N.amh	RD
GLN	CA	C.3	H
GLN	C	C.2	HR
GLN	O	O.2	RA
GLN	CB	C.3	H
GLN	CG	C.3	H
GLN	CD	C.2	HR
GLN	OE1	O.2	RA
GLN	NE2	N.amh	RDA
GLU	N	N.amh	RD
GLU	CA	C.3	H
GLU	C	C.2	HR
GLU	O	O.2	RA
GLU	CB	C.3	H
GLU	CG	C.3	H
GLU	CD	C.2	HR
GLU	OE1	O.co2	RAN
GLU	OE2	O.co2	RAN
GLY	N	N.amh	RD
GLY	CA	C.3	H
GLY	C	C.2	HR
GLY	O	O.2	RA
HIS	N	N.amh	RD
HIS	CA	C.3	H
HIS	C	C.2	HR
HIS	O	O.2	RA
HIS	CB	C.3	H
HIS	CG	C.2	HR
HIS	ND1	N.ar	RA
HIS	CD2	C.2	HR
HIS	CE1	C.2	HR
HIS	NE2	N.arh	RD
ILE	N	N.amh	RD
ILE	CA	C.3	H
ILE	C	C.2	HR
ILE	O	O.2	RA
ILE	CB	C.3	H
ILE	CG1	C.3	H
ILE	CG2	C.3	H
ILE	CD1	C.3	H
LEU	N	N.amh	RD
LEU	CA	C.3	H
LEU	C	C.2	HR
LEU	O	O.2	RA
LEU	CB	C.3	H
LEU	CG	C.3	H
LEU	CD1	C.3	H
LEU	CD2	C.3	H
LYS	N	N.amh	RD
LYS	CA	C.3	H
LYS	C	C.2	HR
LYS	O	O.2	RA
LYS	CB	C.3	H
LYS	CG	C.3	H
LYS	CD	C.3	H
LYS	CE	C.3	H
LYS	NZ	N.4h	RDP
MET	N	N.amh	RD
MET	CA	C.3	H
MET	C	C.2	HR
MET	O	O.2	RA
MET	CB	C.3	H
MET	CG	C.3	H
MET	SD	S.3	H
MET	CE	C.3	H
PHE	N	N.amh	RD
PHE	CA	C.3	H
PHE	C	C.2	HR
PHE	O	O.2	RA
PHE	CB	C.3	H
PHE	CG	C.ar	HR
PHE	CD1	C.ar	HR
PHE	CD2	C.ar	HR
PHE	CE1	C.ar	HR
PHE	CE2	C.ar	HR
PHE	CZ	C.ar	HR
PRO	N	N.am	HR
PRO	CA	C.3	H
PRO	C	C.2	HR
PRO	O	O.2	RA
PRO	CB	C.3	H
PRO	CG	C.3	H
PRO	CD	C.3	H
SER	N	N.amh	RD
SER	CA	C.3	H
SER	C	C.2	HR
SER	O	O.2	RA
SER	CB	C.3	H
SER	OG	O.3h	AD
THR	N	N.amh	RD
THR	CA	C.3	H
THR	C	C.2	HR
THR	O	O.2	RA
THR	CB	C.3	H
THR	OG1	O.3h	AD
THR	CG2	C.3	H
TRP	N	N.amh	RD
TRP	CA	C.3	H
TRP	C	C.2	HR
TRP	O	O.2	RA
TRP	CB	C.3	H
TRP	CG	C.2	HR
TRP	CD1	C.2	HR
TRP	CD2	C.ar	HR
TRP	NE1	N.arh	RD
TRP	CE2	C.ar	HR
TRP	CE3	C.ar	HR
TRP	CZ2	C.ar	HR
TRP	CZ3	C.ar	HR
TRP	CH2	C.ar	HR
TYR	N	N.amh	RD
TYR	CA	C.3	H
TYR	C	C.2	HR
TYR	O	O.2	RA
TYR	CB	C.3	H
TYR	CG	C.ar	HR
TYR	CD1	C.ar	HR
TYR	CD2	C.ar	HR
TYR	CE1	C.ar	HR
TYR	CE2	C.ar	HR
TYR	CZ	C.ar	HR
TYR	OH	O.3h	AD
VAL	N	N.amh	RD
VAL	CA	C.3	H
VAL	C	C.2	HR
VAL	O	O.2	RA
VAL	CB	C.3	H
VAL	CG1	C.3	H
VAL	7CG2	C.3	