Capítol 8.

Resum.

[*Aquesta pàgina ha estat deixada en blanc intencionadament*]

I. RESUM

L'obtenció del primer esborrany del genoma humà porta també associada l'aparició d'una quantitat ingent de dades. D'aquestes cal destacar les relacionades a la variació intraespecífica, en particular les variacions puntuals o SNP (*single nucleotide polimorphisms*). La rellevància d'aquest tipus de variabilitat radica en què es creu que explica un 90% de la variabilitat entre individus (Cargill et al., 1999; Chakravarti, 2001; Collins et al., 1998; Collins et al., 1997; Sachidanandam et al., 2001). Aquesta variabilitat és la base de la transmissió de malalties monogèniques hereditàries, així com la susceptibilitat a patir malalties poligèniques amb un patró d'heredabilitat complex. Així, la teoria CVCD (*common variant-common disease*) (Cargill et al., 1999; Sachidanandam et al., 2001), explicaria que, SNPs poc freqüents serien responsables de malalties rares, que solen ser les monogèniques o mendelianes. Altrament, diverses mutacions més abundants en la població, explicarien la susceptibilitat a patir malalties més comuns. Una comprensió profunda de la relació entre patologia i mutacions puntuals ha de provenir de la comprensió de l'impacte que tenen aquestes mutacions a nivell molecular. De fet, abunden els estudis centrats en l'efecte de les mutacions puntuals sobre l'estructura de les proteïnes i les seves conseqüències sobre la funció. En aquest sentit, caldria destacar estudis sobre els efectes dels mutants puntuals sobre l'estructura de la proteïna mitjançant la mutagènesi dirigida realitzats especialment per Mathews i Fersht (Fersht & Serrano, 1993; Matthews, 1987; Matthews, 1993; Matthews, 1995).

Juntament amb aquestes dades existeix una quantitat important d'estudis de caracterització a nivell clínic de mutacions en pacients de malalties monogèniques. Així es van dirigir els nostres interessos en la caracterització de mutacions puntuals patològiques amb la idea de poder classificar les mutacions

puntuals segons fossin patològiques o neutres. Amb aquest objectiu es va construir una base de dades de mutacions puntuals en proteïna associades a malaltia. Com a font primària de dades es va usar la base de dades de SwissProt (Apweiler et al., 2004) degut a la qualitat contrastada de les seves anotacions. Es van buscar mutacions puntuals associades a proteïnes humanes i amb estructura tridimensional coneguda.

Una vegada construïda la base de dades de mutacions patològiques es va establir un model de mutació neutra, degut a que encara no existeix una base de dades contrastada que contingui un número important de SNPs no sinònims neutres. El model triat es deriva de l'anàlisi dels alineaments múltiples de seqüència. Així, de les seqüències que comparteixen més d'un 95% d'identitat de seqüència amb la proteïna estudiada es consideren mutacions neutres totes aquestes que confeccionen aquest 5% de dissimilaritat amb la proteïna d'interès. Aquest model de neutralitat també l'han usat altres autors (Santibañez-Koref et al., 2003; Sunyaev et al., 2001). Aquest assumeix que si dues proteïnes han divergit lleugerament aquest canvis han estat acceptats de manera silenciosa o neutre per l'evolució.

Amb els dos grups es va fer una anàlisi comparativa de les distribucions de fins 23 paràmetres diferents per les mutacions, que podem agrupar en tres grans blocs: paràmetres estructurals, evolutius i basats en propietats d'aminoàcids. L'observació de moltes d'aquestes distribucions mostren que hi ha un clar comportament diferencial entre les mutacions neutres i les patològiques. Així els valors de matrius de mutacions, especialment la matriu de mutació Blosum62 (veure fig1), mostra un biax en les dues distribucions on les mutacions neutres acumulen més valors positius, o sigui permesos evolutivament, mentre que les mutacions patològiques acumulen valors negatius. Així sembla que les mutacions patològiques tendeixen a

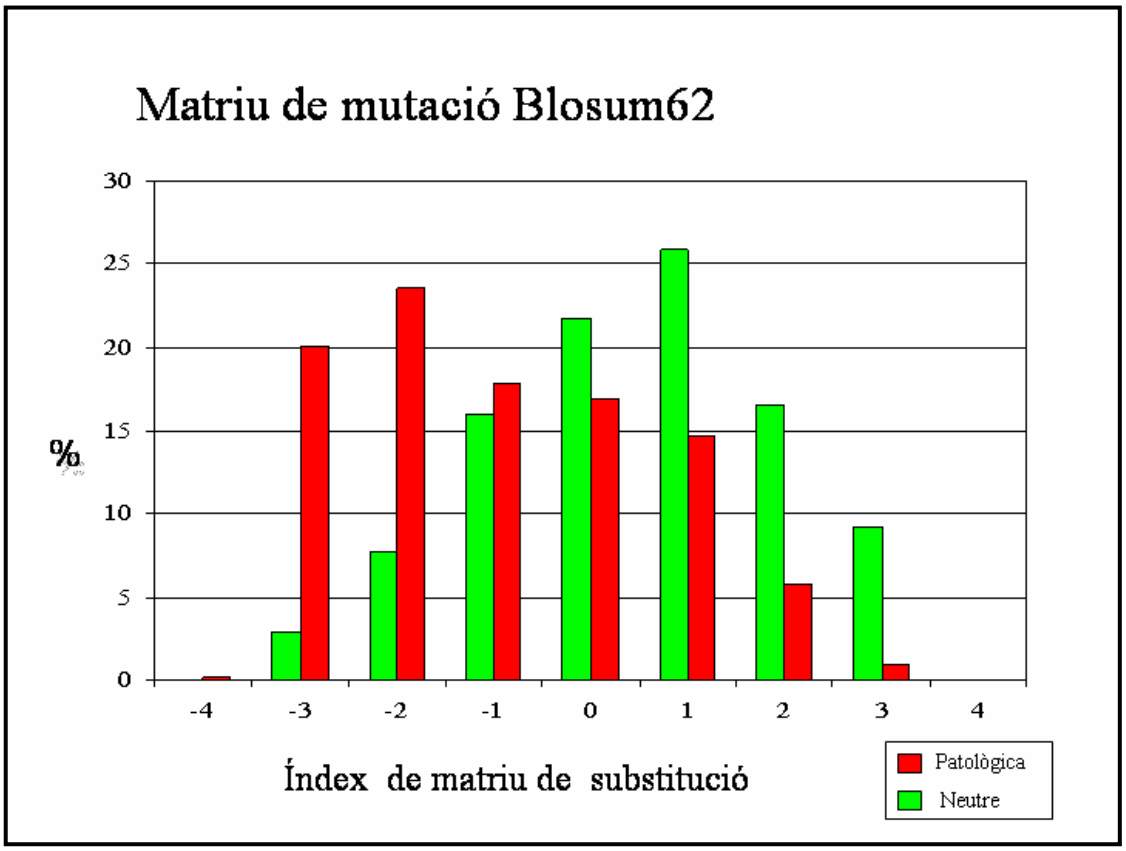


Fig1. Distribució dels valors de la matriu de mutació blosum62 de les mutacions patològiques i neutres. En verd les mutacions neutres en vermell les mutacions patològiques. En l'eix d'ordenades els valors de la matriu en l'eix d'absisses el percentatge de població per cada valor de la matriu.

acumular canvis que en general l'evolució no ha acceptat bé; per altra banda les mutacions neutres tendeixen a acumular canvis que evolutivament han estat més acceptats. Un altre exemple clar és que les mutacions patològiques tenen tendència a trobar-se en zones enterrades (veure figura 2) mentre que les mutacions neutres es distribueixen d'una manera més uniforme des del punt de vista de l'accessibilitat al solvent. Altres paràmetres tenen distribucions més complexes i la distinció entre patològiques i neutres no es tan clara, com per exemple volums o hidrofobicitats (veure figura 3).

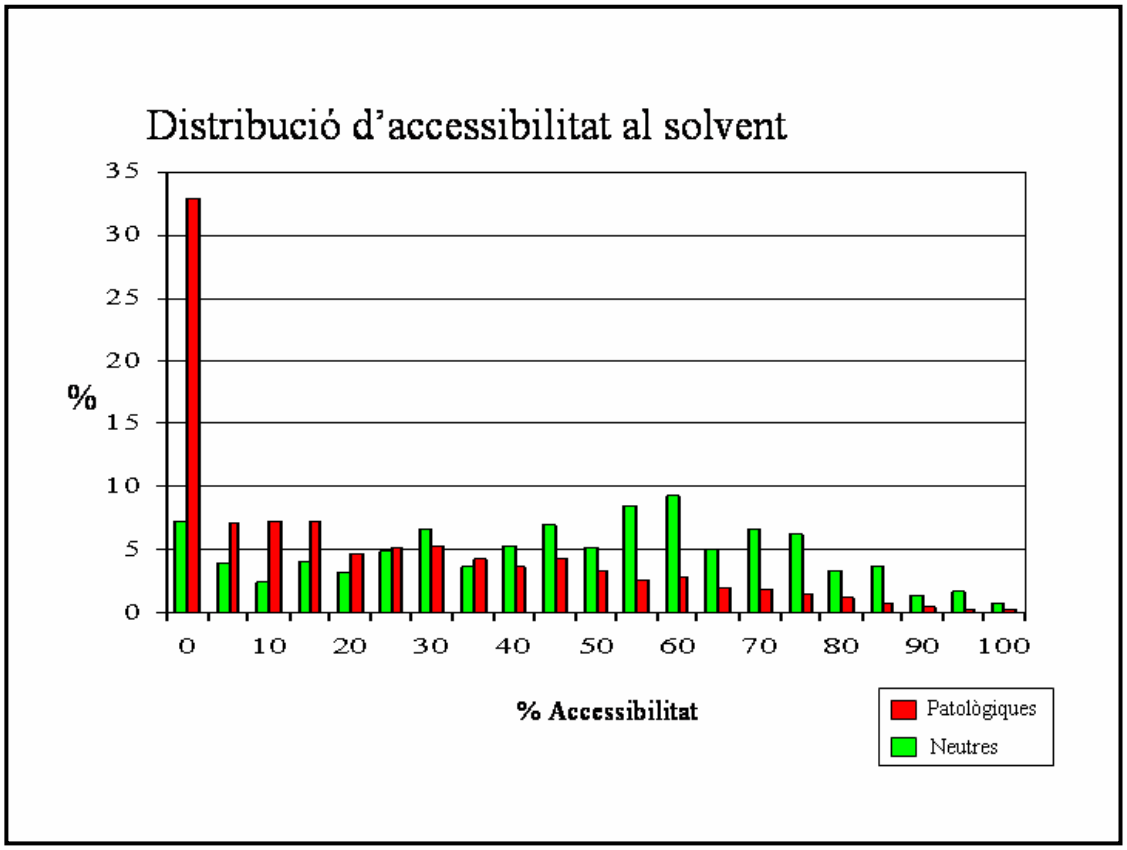


Fig2. Distribució dels valors de l'accessibilitat al solvent de les mutacions patològiques i neutres. En verd les mutacions neutres en vermell les mutacions patològiques. En l'eix d'ordenades els valors de l'accessibilitat en percentatge. En l'eix d'absisses el percentatge de població per cada valor d'accessibilitat.

A la vista d'aquests resultats, el següent pas a fer va ser provar l'ús d'aquests paràmetres per l'anotació o predicció de mutacions puntuals o SNPs no sinònims en regions codificants. Els primers indicis de que era una possibilitat palpable va ser que usant com a predicció la combinació adequada de valors de dos paràmetres, com són PAM40 i Blosum62, es pot encertar el 84% de les mutacions que complien la combinació de valors i que són un 40% de totes les mutacions.

Els diferents treballs publicats fins aleshores usaven només estratègies simples on s'usaven un o dos dels paràmetres, els que tenien més poder discriminant. El canvi que vam introduir va ser l'ús de xarxes neurals, per tal de combinar tots els

paràmetres usats en la caracterització de les mutacions, i millorar així les

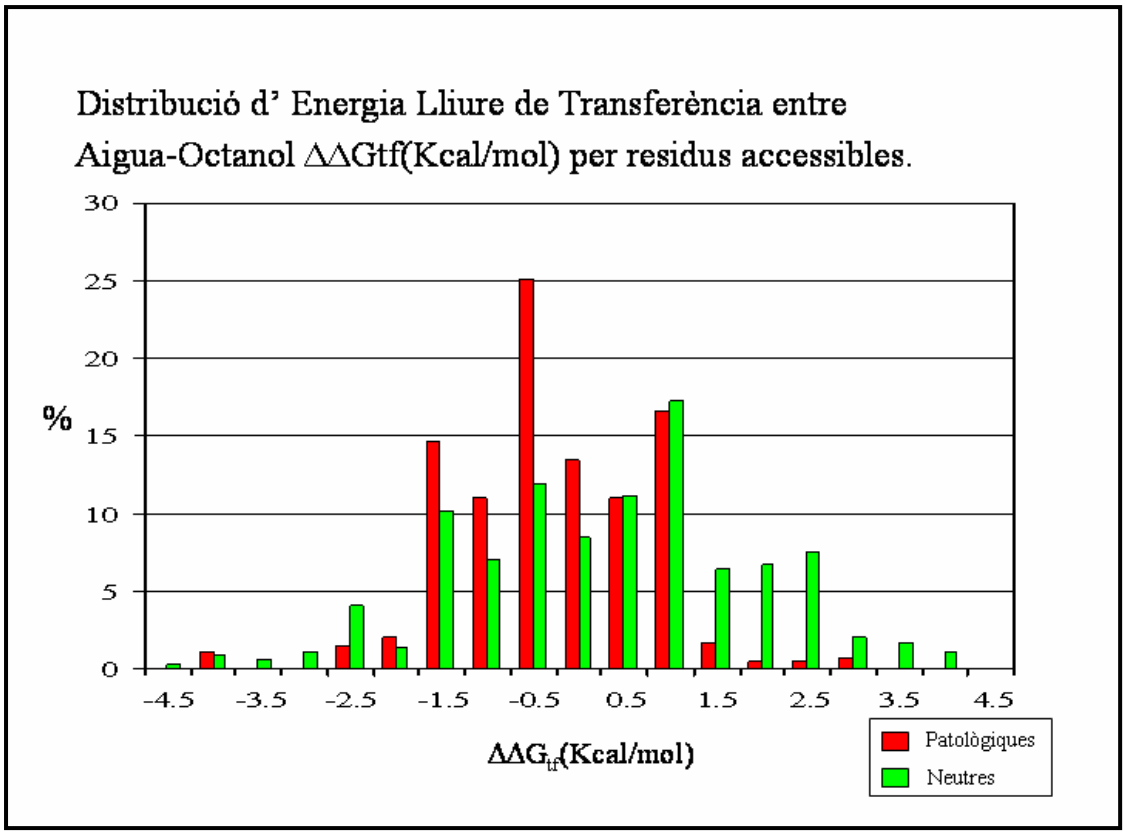


Fig3. Distribució dels valors d'hidrofobicitat de les mutacions patològiques i neutres representades pels valors de la variació d'energia lliure de transferècia d'octanol i aigua. En verd les mutacions neutres en vermell les mutacions patològiques. En l'eix d'ordenades els valors de la variació d'energia lliure de transferècia d'octanol i aigua en Kcal/Mol. En l'eix d'absisses el percentatge de població per cada valor d'hidrofobicitat.

resultats en la predicció del caràcter patològic de les mutacions puntuals.

Una altra preocupació era poder usar només dades derivades de la seqüència per tal de caracteritzar mutacions puntuals en el màxim nombre de proteïnes. El número de proteïnes amb estructura coneguda és molt reduït comparat amb el número de seqüències dipositades en les bases de dades. Això va portar a una reducció en el número de paràmetres especialment a aquells obtinguts a partir de l'estructura de la proteïna. La resta de paràmetres es van mantenir ja que no

depenen del coneixement de l'estructura.

Per establir la bondat de la predicció es va usar una estratègia de validació creuada de la predicció. Aquesta estratègia consisteix en mesclar les mutacions neutres i patològiques i dividir-les en 5 grups. Amb aquests grups de dades s'usen 4 grups per entrenar la xarxa i el grup restant s'usa per predir. Aquest procediment es fa 5 vegades canviant de manera iterativa el grup de predicció, obtenint 5 resultats diferents de predicció que són independents dels grups d'entrenament. El resultat final és la mitja ponderada de les 5 prediccions. Amb aquest procediment s'obté amb la xarxa més optimitzada (1 capa oculta i 20 nodes) un encert total del 87% i un enriquiment de la predicció sobre l'atzar del 73%. Aquestes dades són clarament millors que les corresponents als mètodes de predicció publicats fins al moment. Per exemple, usant la validació creuada amb les dades de Henikoff (Ng & Henikoff, 2001), s'obté un 70% d'encert total i un 37% d'enriquiment sobre una predicció feta a l'atzar. Al usar dades derivades només de seqüència s'obté un encert total del 84% amb un enriquiment sobre l'atzar del 67%. Així en el pas de l'estructura a la seqüència tenim una petita pèrdua de rendiment que és assumible pel guany en el número de mutacions. En comparació amb els altres mètodes es veu una millora en les prediccions atribuïble segurament al bon comportament de les xarxes neurals.

L'evolució natural d'aquest projecte dedicat a la caracterització i predicció de les mutacions puntuals patològiques en proteïnes passava, vam creure, en fer accessible l'ús d'aquesta metodologia. Es va fer mitjançant el disseny d'un servidor web que permetés la predicció de les mutacions puntuals per diferents proteïnes. El servidor s'anomena PMut i és accessible via internet a l'adreça http://mmb2.pcb.ub.es:8080/PMut. Consta de dues parts diferenciades, una que és el servidor pròpiament dit i una segona que s'organitza com a base de dades i conté totes les prediccions possibles de les estructures del cluster 90% del PDB

(totes les seqüències del PDB agrupades per identitats de seqüència, en aquest cas un 90%). Pel que fa al servidor, el protocol de predicció s'ha modificat lleugerament per tal de fer-lo funcional, així el número de paràmetres s'ha reduït a 15. S'han canviat els alineaments múltiples per la problemàtica de mapar els alineaments de PFAM a qualsevol possible seqüència. Per això es va optar per construir els alineaments a partir de Psiblast. Un cop els paràmetres estan calculats s'usa la xarxa entrenada amb les dades de SWP presentades anteriorment. Els resultats es presenten per cada mutació, amb tots els resultats intermitjos i el resultat final de la xarxa, i la fiabilitat de la predicció. En el cas que la mutació es localitzi en una zona de la proteïna per la qual existeix una estructura homòloga, es pot representar la mutació sobre aquesta estructura.

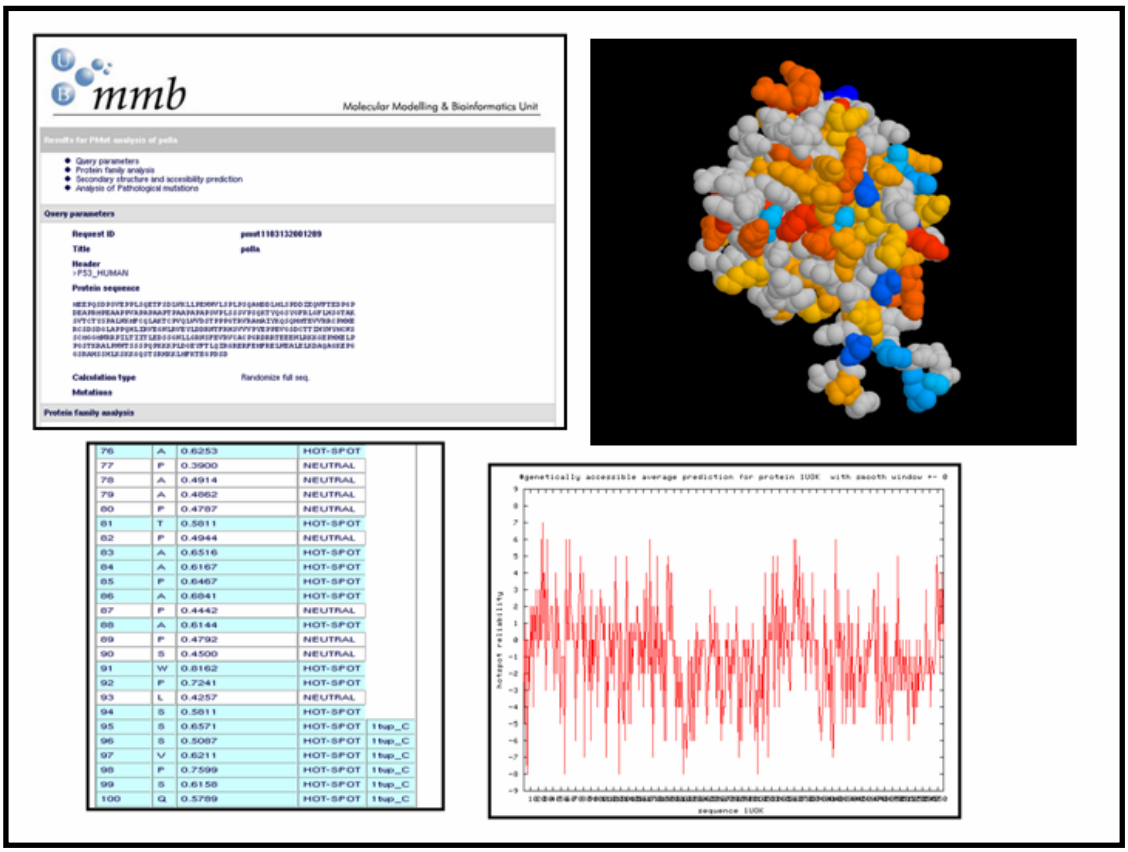
PMut és accessible a http://mmb2.pcb.ub.es:8080/PMut/. Una versió reduïda de PMut es troba al servidor PupasView del grup del Dr. Dopazo http://pupasview.bioinfo.ochoa.fib.es/ i una altra versió simplificada es troba dins dels workflows del Instituto Nacional de Bioinformática. Diversos exemples de la representació dels resultats del servidor es poden observar a la figura 4.

Al llarg d'aquestes pàgines s'ha descrit tota la metodologia desenvolupada per l'anotació funcional de les mutacions puntuals en proteïnes humanes. Tanmateix, altres projectes de seqüenciació han estat desenvolupats degut a l'interès que tenen els animals model per la biomedicina. En molts animals, especialment en ratolí, s'han intentat generar mutants que siguin models de malalties humanes. En aquesta línia cal destacar diferents projectes de mutació massiva en ratolí, mitjançant tècniques basades en mutagènesi generada per ENU (Hrabe de Angelis & Strivens, 2001; Hrabe de Angelis et al., 2000), on es generen milers

d'animals mutants que se sotmeten a l'estudi fenotípic per a la detecció de nous gens implicats en diferents malalties d'interès humà. Degut a aquests diferents

projectes molta informació sobre variabilitat puntual ha esdevingut accessible en aquestes espècies i això fomenta la necessitat de desenvolupar eines ràpides

Fig 4. Exemples dels diferents resultats que dóna el servidor Pmut



76	A	0.6253	HOT-SPOT	
77	P	0.3900	NEUTRAL	
78	A	0.4914	NEUTRAL	
79	A	0.4862	NEUTRAL	
80	P	0.4787	NEUTRAL	
81	T	0.5811	HOT-SPOT	
82	P	0.4944	NEUTRAL	
83	A	0.6516	HOT-SPOT	
84	A	0.6167	HOT-SPOT	
85	P	0.6467	HOT-SPOT	
86	A	0.6841	HOT-SPOT	
87	P	0.4442	NEUTRAL	
88	A	0.6144	HOT-SPOT	
89	P	0.4792	NEUTRAL	
90	S	0.4500	NEUTRAL	
91	W	0.8162	HOT-SPOT	
92	P	0.7241	HOT-SPOT	
93	L	0.4257	NEUTRAL	
94	S	0.5811	HOT-SPOT	
95	S	0.6571	HOT-SPOT	1 bup_C
96	S	0.5087	HOT-SPOT	1 bup_C
97	V	0.6211	HOT-SPOT	1 bup_C
98	P	0.7599	HOT-SPOT	1 bup_C
99	S	0.6158	HOT-SPOT	1 bup_C
100	Q	0.5789	HOT-SPOT	1 bup_C

accessibles i barates per l'anotació d'aquestes dades. Aquestes raons ens van empènyer en l'estudi de quan transferibles podien ser les eines derivades per l'anotació de mutacions puntuals humanes en l'anotació de mutacions en animals models.

En la transferibilitat del mètode juguen diversos factors però cal destacar-ne dos especialment: la reducció de la mida de la base de dades i la distància evolutiva entre les proteïnes.

Per aquest estudi es van cercar a la base de dades Swissprot totes les mutacions puntuals patològiques per proteïnes no humanes. D'aquesta manera es va generar una base de dades de mutacions patològiques i d'aquestes es van derivar les neutres corresponents. D'aquestes mutacions, la majoria corresponen a ratolí (71% de les patològiques 42% de les neutres). Com a model humà es va usar la base de dades usada pel treball en proteïnes humanes. Un dels primers aspectes a estudiar era en quin grau de probabilitat una mutació que és patològica en humans serà neutre en ratolí. Per fer això, es va representar la distribució d'identitats entre proteïnes homòlogues d'humà i de ratolí en front de dues distribucions: la distribució d'identitats entre proteïnes humanes que tenen almenys una mutació puntual i les homòlogues d'altres espècies; i la distribució d'identitats de seqüència entre proteïnes humanes i de ratolí quan les dues tenen mutacions puntuals patològiques. Tal i com es pot veure en la figura 5 en la zona d'identitats de seqüència superior al 75% l'àrea superposada és realment petita, indicant que la proporció de mutacions patològiques en humans i neutres en ratolí serà petita. Per tant, esperem que la gran majoria de mutacions patològiques en humà seran també patològiques en ratolí i també al revés. La distribució d'identitats de seqüència per gens relacionats amb malaltia és molt similar que el global per ratolí i humà.

Una altra qüestió per resoldre estava en si el grup de mutacions humanes i de ratolí cobrien el mateix espectre funcional de proteïnes. Si això era així, la transferència entre espècies seria més efectiva, altrament existiria un biax. Una manera d'aproximar el problema consistia en representar la distribució dels diferents paràmetres per les mutacions patològiques humanes i d'altres espècies, i en cap cas cap diferència estadísticament significativa era observable.

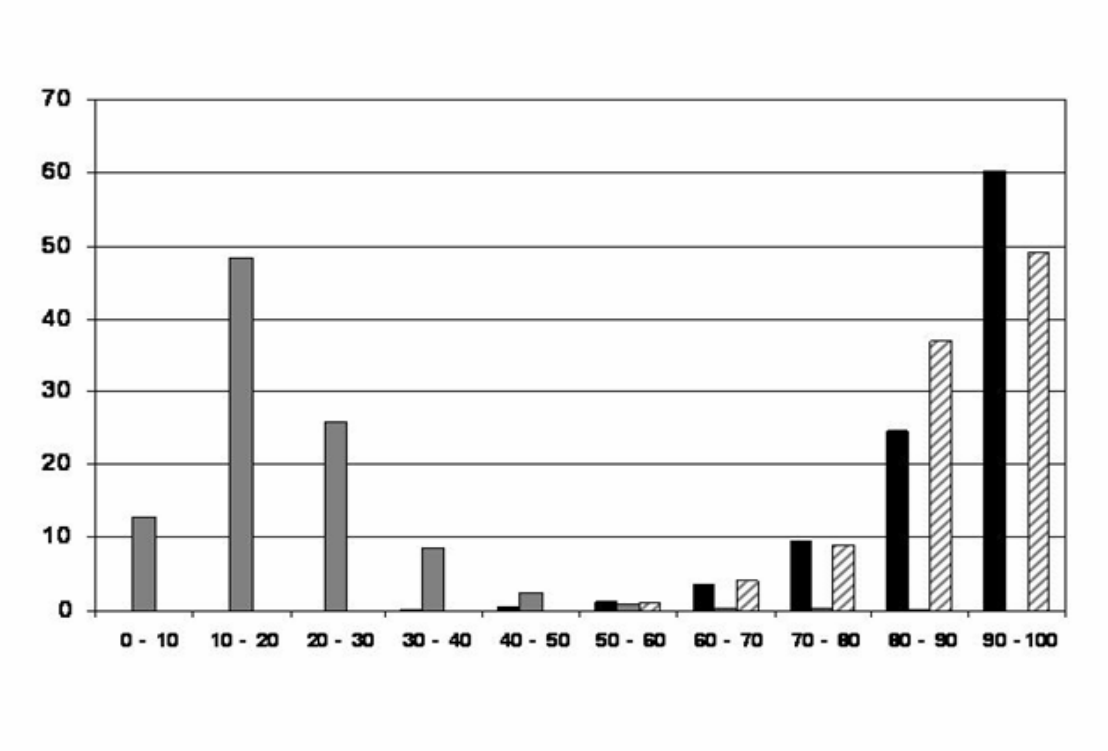


Fig 5. Distribució d'identitats de seqüència entre: (i) homòlogues (NEGRE); (ii) proteïnes humanes i de ratolí, en les quals les dues tenen mutacions associades a malaltia segons SwissProt (RATLLAT); i (iii) proteïnes humanes i les homòlogues no humanes, on les humanes tenen com a mínim una posició patològica. (GRIS).

Havent assegurat aquests aspectes previs es va procedir a fer ús de les xarxes neurals per posar a punt procediments de predicció tant pel grup de dades de ratolí com per les prediccions creuades. Tal i com s'havia mostrat en treballs anteriors l'ús de xarxes en dades humanes s'observava un rendiment gran quan s'usaven tots els paràmetres, amb un 86% de prediccions correctes i un 65% d'enriquiment sobre una predicció a l'atzar (Ferrer-Costa et al., 2004). Amb aquests resultats com a referència es van usar les dades humanes per entrenar una xarxa per predir les dades de ratolí, i els resultats tot i que inferiors són bons, ja que tenim un encert total del 86% i un enriquiment sobre l'atzar del 53%. Al provar al revés aquestes xarxes, és a dir entrenant amb les dades de ratolí i predint sobre les dades humanes, els resultats es mantenen bastant, amb un encert total del 78% i un enriquiment respecte a l'atzar del 55%. La diferència de

comportament és atribuïble a la reducció en la mida de la base de dades.

Al llarg d'aquest treballs va sorgir una qüestió pel que fa a la transferibilitat de l'anotació de mutacions puntuals: quan una mutació que és patològica en una proteïna humana ho seguirà sent en proteïnes homòlogues? Aquesta, qüestió apareix amb més força quan en l'anàlisi dels alineaments múltiples apareix el residu patològic com a salvatge en altres proteïnes d'altres espècies, casos que es coneixen com CPDs (*compensated pathogenic deviations*) (Kondrashov et al., 2002). Davant aquest fet cal aclarir quines raons fan que un aminoàcid que provoca l'aparició d'una patologia en una espècie es manté com a residu salvatge en altres. En els pocs treballs apareguts fins ara els autors conclouen que les raons que expliquen les CPDs són les mutacions compensatòries que apareixen en un període de temps evolutivament proper i que fan que la proteïna pugui adaptar-se a aquest canvi. Aquestes mutacions compensatòries poden aparèixer en la mateixa proteïna o en proteïnes que interaccionen de manera directa amb aquesta. Malgrat que sol ser difícil de determinar quines són les mutacions compensatòries, l'ús conjunt de l'estructura i l'anàlisi dels alineaments múltiples pot ajudar a determinar quina és la mutació compensatòria (Gao & Zhang, 2003; Kondrashov et al., 2002), almenys en proteïnes evolutivament molt properes i amb entorns molt conservats. Davant aquests fets, ens vam decidir a explorar aquest fenomen amb el nostre grup de dades de mutacions patològiques humanes. De l'anàlisi dels alineaments de les 8337 mutacions patològiques es van derivar 71295 CPDs. D'aquestes CPDs es van derivar un seguit d'anàlisis a nivell de seqüència i d'estructura que semblen indicar-nos que altres raons podrien explicar també l'aparició d'aquests residus en altres espècies. El primer anàlisi consisteix en comparar la distribució de identitats locals entre la seqüència humana amb la homòloga no humana que conté la CPD, per intervals d'identitat global en front la distribució d'identitat local esperada, que es pot assimilar a una

binomial. El que esperem és que si existeixen les mutacions compensatòries s'haurien de mostrar amb un biaix entre la distribució d'identitat local observada i l'esperada. Fent l'anàlisi pels aparellaments de seqüències que tenen més d'un

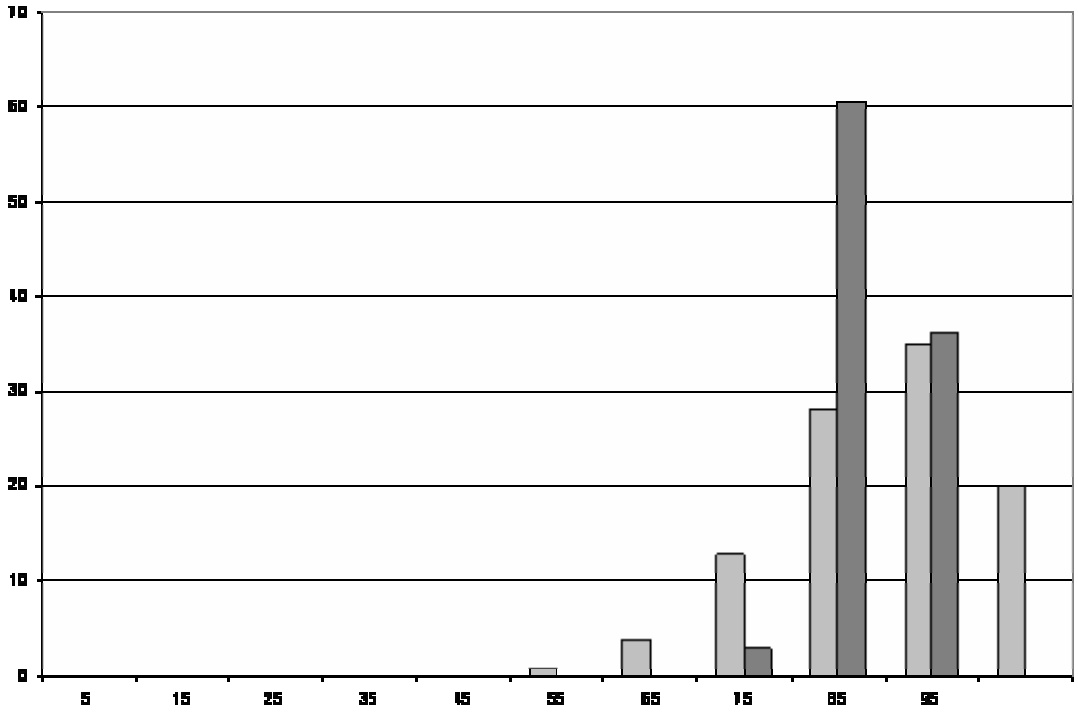


Fig. 6. Distribució d'identitat de seqüència a nivell estructural pels veïns de les CPDs. Només es consideren aquells parells de seqüències entre humanes i no-humanes amb una identitat global de seqüència superior al 90%. Per comparació es representa la distribució obsevada (gris fosc) i la distribució esperada (gris clar).

90% d'identitat global de seqüència, a nivell de seqüència no apareix aquest biaix que si que apareix quan l'anàlisi la fem a nivell estructural i la identitat de seqüència és local, estructuralment parlant. Així tal i com es pot veure en la figura 6 s'observa que la identitat de seqüència local és menor que la esperada per la finestra d'identitat de seqüència global que va de 90-100%.

Sembla doncs que per identitats de seqüència elevada seria la presència de

mutacions compensatòries les que explicarien les CPDs. Tanmateix per identitats globals més baixes aquesta tendència desapareix i res ens permet aventurar la presència de mutacions compensatòries.

Per tal d'explorar altres explicacions es va analitzar la distribució de les accessibilitats de les CPDs, en comparació amb les de les mutacions patològiques, es veu que hi ha un biaix de les CPDs cap a regions més exposades per identitats de seqüència elevada. Això indicaria que les CPDs tendeixen a situar-se en zones més elàstiques com són regions més exposades que permetrien acceptar aquest canvi millor. Aquesta tendència desapareix per identitats de seqüència baixes, per sota del 60%, això s'explicaria pel fet que en aquestes identitats de seqüència, l'entorn local ha pogut canviar prou com per acomodar el nou residu.

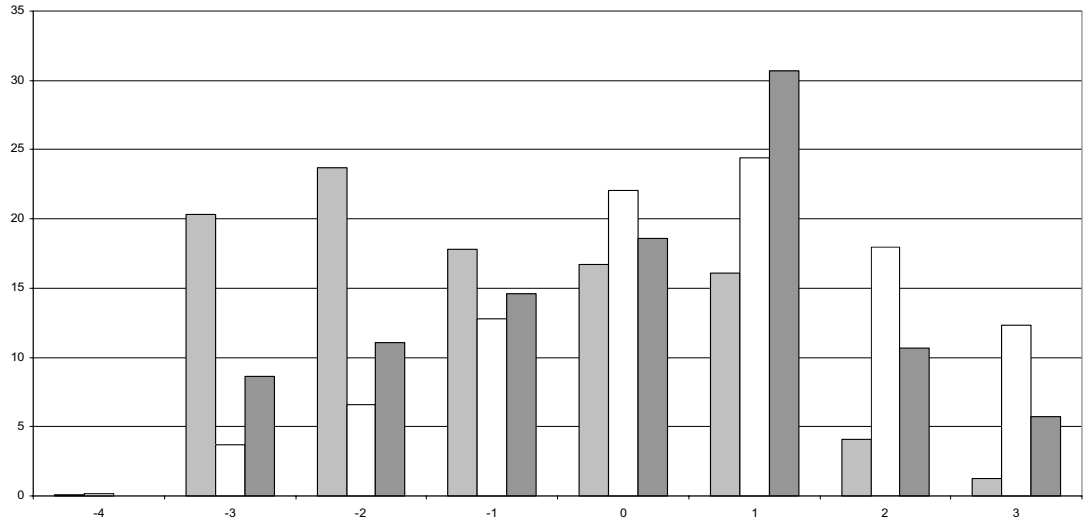


Fig 7. Distribució de valors de matriu Blosum62. Tres distribucions diferents es mostren: CPD (gris fosc), Patològiques (gris clar) I neutres (blanc).

Finalment com a últim exemple, al analitzar alguna de les propietats aminoacídiques com pot ser el valor de la matriu Blosum62 i al representar-los

en front de les distribucions de les mutacions patològiques i neutres s'observa que la de les CPDs es distribueixen diferentment en una posició intermitja. Aquesta observació ens referiria al fet que lesCPDs serien canvis menys dràstics o lesius que les mutacions patològiques tal i com es reflexa a la figura 7. Així doncs en general podem concloure que les mutacions compensatòries explicarien certes CPDs, especialment aquelles que han tingut lloc entre proteïnes amb alta identitat de seqüència. Altres raons per l'existència de les CPDs podrien ser que aquesta ocórrer en zones menys lesives, com poden ser zones en superfície o bé que el canvi sigui de naturalesa més conservativa. Per proteïnes evolutivament més distants es podrien explicar pel fet que l'entorn estructural ha evolucionat prou com per adaptar aquest canvi.

II. BIBLIOGRAFIA DEL CAPÍTOL

Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Natale, D. A., O'Donovan, C., Redaschi, N. & Yeh, L. S. (2004). UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res* 32 Database issue, D115-9.

Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., Shaw, N., Lane, C. R., Lim, E. P., Kalyanaraman, N., Nemesh, J., Ziaugra, L., Friedland, L., Rolfe, A., Warrington, J., Lipshutz, R., Daley, G. Q. & Lander, E. S. (1999). Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat Genet* 22, 231-8.

Chakravarti, A. (2001). To a future of genetic medicine. *Nature* 409, 822-3.

Collins, F. S., Brooks, L. D. & Chakravarti, A. (1998). A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res* 8, 1229-31.

Collins, F. S., Guyer, M. S. & Charkravarti, A. (1997). Variations on a theme: cataloging human DNA sequence variation. *Science* 278, 1580-1.

Ferrer-Costa, C., Orozco, M. & De La Cruz, X. (2004). Sequence-based prediction of pathological mutations. *Proteins* 57, 811-819.

Fersht, A. R. & Serrano, L. (1993). Principles of protein stability derived from protein engineering experiments. *Current opinion in structural biology* 3, 75-83.

Gao, L. & Zhang, J. (2003). Why are some human disease-associated mutations fixed in mice? *Trends Genet* 19, 678-81.

Hrabe de Angelis, M. & Strivens, M. (2001). Large-scale production of mouse phenotypes: the search for animal models for inherited diseases in humans. *Brief Bioinform* 2, 170-80.

Hrabe de Angelis, M. H., Flaswinkel, H., Fuchs, H., Rathkolb, B., Soewarto, D., Marschall, S., Heffner, S., Pargent, W., Wuensch, K., Jung, M., Reis, A., Richter, T., Alessandrini, F., Jakob, T., Fuchs, E., Kolb, H., Kremmer, E., Schaeble, K., Rollinski, B., Roscher, A., Peters, C., Meitinger, T., Strom, T., Steckler, T., Holsboer, F., Klopstock, T., Gekeler, F., Schindewolf, C., Jung, T., Avraham, K., Behrendt, H., Ring, J., Zimmer, A., Schughart, K., Pfeffer, K., Wolf, E. & Balling, R. (2000). Genome-wide, large-scale production of mutant mice by ENU mutagenesis. *Nat Genet* 25, 444-7.

Kondrashov, A. S., Sunyaev, S. & Kondrashov, F. A. (2002). Dobzhansky-Muller incompatibilities in protein evolution. *Proc Natl Acad Sci U S A* 99, 14878-83.

Matthews, B. W. (1987). Genetic and structural analysis of the protein stability problem. *Biochemistry* 26, 6885-8.

Matthews, B. W. (1993). Structural and genetic analysis of protein stability. *Annu Rev Biochem* 62, 139-60.

Matthews, B. W. (1995). Studies on protein stability with T4 lysozyme. *Adv Protein Chem* 46, 249-78.

Ng, P. C. & Henikoff, S. (2001). Predicting deleterious amino acid substitutions. *Genome Res* 11, 863-74.

Sachidanandam, R., Weissman, D., Schmidt, S. C., Kakol, J. M., Stein, L. D., Marth, G., Sherry, S., Mullikin, J. C., Mortimore, B. J., Willey, D. L., Hunt, S. E., Cole, C. G., Coggill, P. C., Rice, C. M., Ning, Z., Rogers, J., Bentley, D. R., Kwok, P. Y., Mardis, E. R., Yeh, R. T., Schultz, B., Cook, L., Davenport, R., Dante, M., Fulton, L., Hillier, L., Waterston, R. H., McPherson, J. D., Gilman, B., Schaffner, S., Van Etten, W. J., Reich, D., Higgins, J., Daly, M. J., Blumenstiel, B., Baldwin, J., Stange-Thomann, N., Zody, M. C., Linton, L., Lander, E. S. & Altshuler, D. (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409, 928-33.

Santibañez-Koref, M. F., Gangeswaran, R., Santibanez-Koref, I. P., Shanahan, N. & Hancock, J. M. (2003). A phylogenetic approach to assessing the significance of missense mutations in disease genes. *Hum Mutat* 22, 51-8.

Sunyaev, S., Ramensky, V., Koch, I., Lathe, W., 3rd, Kondrashov, A. S. & Bork, P. (2001). Prediction of deleterious human alleles. *Hum Mol Genet* 10, 591-7.