Capítol 4.

Predicció de mutacions patològiques usant xarxes neurals.

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

I. INTRODUCCIÓ

Tal i com s'ha descrit en el capítol anterior, per algunes propietats estructurals, físico-químiques i evolutives de les mutacions puntuals el comportament de les patològiques i de les neutres és estadísticament diferent. Tanmateix, una limitació molt gran fins aleshores estava en la necessitat de mapar la mutació estudiada en una estructura tridimensional, ja sigui de la pròpia proteïna com d'una estructura homòloga. Aquesta limitació és més gran quan el número de seqüències creix molt més ràpid que el d'estructures conegudes. Tal com s'ha comentat en el capítol anterior el comportament del model de les mutacions patològiques en front certs paràmetres és significativament diferent del comportament del model neutre. Aquest fet obre la possibilitat de predir el caràcter patològic de les mutacions puntuals basant-nos en aquests paràmetres. Un mètode de predicció basat en l'ús de propietats estructurals té un abast limitat ja que encara són moltes les mutacions que es troben en proteïnes o en dominis de proteïnes de les quals encara no es coneix l'estructura. Per superar aquest problema, ens vam proposar el desenvolupament d'un protocol de predicció basat només en l'ús de propietats derivades a partir de la seqüència de les proteïnes involucrades.

El pas de seqüència a estructura es va fer usant el programa de predicció d'accessibilitat i estructura secundària desenvolupat per Rost et al. (Rost & Sander, 1993) PHD. Aquest mètode arriba fins al 70% d'encert en les millors condicions de predicció, per tant, com s'ha d'esperar aquesta pèrdua en la fiabilitat dels paràmetres estructurals ha de representar una pèrdua en la capacitat predictiva.

Es van calcular també nous paràmetres alguns relacionats amb els alineaments múltiples de seqüència d'altres relacionats amb anotacions en bases de dades i altres relacionats amb potencials estadístics pels residus implicats.
Era interessant també implementar un mètode de predicció que fos capaç d'usar d'una manera útil tots els paràmetres calculats. Tots els autors citats anteriorment usen mètodes molt simples. Nosaltres vam decidir fer ús de les xarxes neurals s'han usat amb èxit en diferents camps per la classificació de patrons (Bishop, 1995). Per poder classificar una mutació en neutre o patològica cal fer un entrenament de les xarxes amb els grups de dades coneguts. Així la xarxa aprèn les diferències entre mutacions neutres i patològiques segons els paràmetres que les defineixen. Les xarxes neurals solen usar models no lineals tot i que també poden usar models lineals i optimitzen la separació entre diferents grups de dades.

II. MATERIALS I MÈTODES

El protocol de predicció

L'objectiu d'aquest treball consistia en desenvolupar un mètode computacional per predir quan una mutació puntal en una proteïna serà o no patològica. Amb aquest objectiu les mutacions són caracteritzades segons un seguit de propietats (canvi de volum, localització en l'estructura etc.) que es poden relacionar amb el seu possible efecte lesiu a la proteïna. Amb el grup de propietats és aleshores usat en una xarxa neural per decidir si una mutació és patològica o neutra.

Mutacions associades a malaltia

Existeixen dos grans bases de dades de mutacions associades a malalties humanes: OMIM (McKusick-Nathans Institute for Genetic Medicine, 2000) i SwissProt (Boeckmann et al., 2003). Vam triar de treballar amb Swissprot ja que en un article recent (Goodstadt & Ponting, 2001) s'indica que aquesta base de dades dóna un mapat més acurat de les mutacions en les seqüències de les proteïnes. Aquest fet és especialment necessari en casos com els nostres on varis milers de mutacions són usades i es requereix un processat automàtic. Cal mencionar també, que Swissprot ha estat utilitzat per altres autors que treballen en predicció de mutacions patològiques (Ng & Henikoff, 2002; Saunders & Baker, 2002; Sunyaev et al., 2001).

Vam usar la versió 40 de la base de dades SwissProt (Boeckmann et al., 2003), obtenint el grup de mutacions patològiques seguint la metodologia proposada per Ng i Henikoff (Ng & Henikoff, 2001). Així vam cercar en la base de dades SwissProt usant les paraules DISEASE, VARIANT i HUMAN (veure figura 1). Es van descartar totes aquelles mutacions per les quals no hi havia una relació clara i directa amb malaltia. També es van descartar aquelles mutacions que no mapaven en cap alineament Pfam (Bateman et al., 2002) (cal més d'una seqüència per calcular algunes propietats basades en evolució usades en el procés de predicció). Finalment les mutacions associades a proteïna (DAMU) van ser 9334 en 811 proteïnes.

Com a referència també vam utilitzar un conjunt de mutacions que ocorren en zones d'estructura coneguda. Aquest grup basat en estructura es va construir usant les mutacions del grup anterior que van ser mapades en l'estructura de la proteïna. El conjunt final de mutacions associades a malaltia amb estructura coneguda va ser de 1319 en 90 proteïnes.

Mutacions neutres

Dos grups de mutacions neutres han estat usats en aquest treball. El primer va ser obtingut dels experiments fets de mutagènesi massiva sobre el repressor de l'operó Lac de *Escherichia coli* (Pace et al., 1997), quedant-nos amb aquelles mutacions que resulten en individus amb fenotip salvatge (Ng & Henikoff, 2001). Es va usar aquest grup de dades ja que havia estat usat per Ng i Henikoff (Ng & Henikoff, 2002) per assajar el seu mètode de predicció basat en la

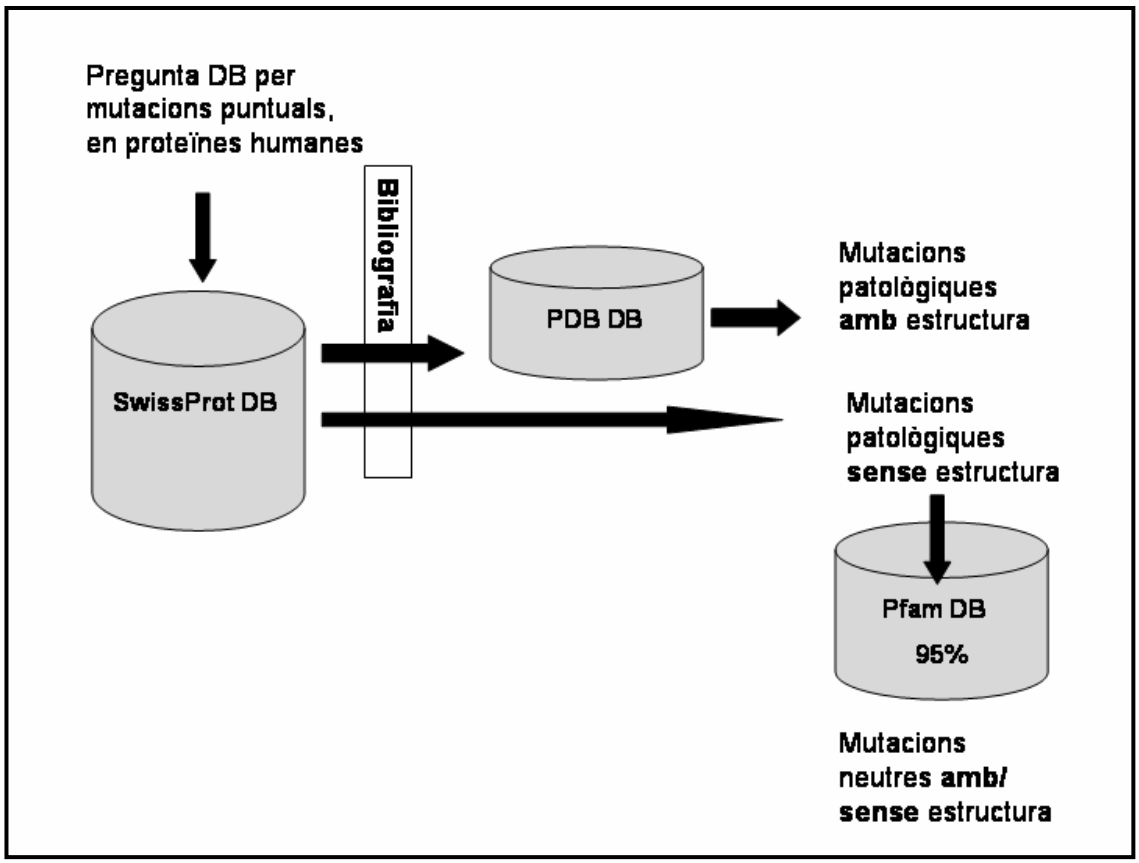

Fig 1. Esquema de selecció de mutacions neutres i patològiques.

seqüència, i per tant permet una comparació directa entre ambdos aproximacions.

Pel segon model de les mutacions neutres (NEMU), que anomenem model evolutiu, es van recollir mutacions neutres només per aquelles proteïnes per les quals es tenien mutacions associades a malaltia. Per aquestes proteïnes, les mutacions neutres corresponen a aquelles variants que ocorren en proteïnes de la mateixa família però en diferents espècies (Ferrer-Costa et al., 2002; Sunyaev et al., 2000). Amb aquesta idea es van prendre els alineaments múltiples de seqüència per la família de les proteïnes estudiades de la base de dades Pfam (Bateman et al., 2002). Per cada alineament se'n van eliminar totes les seqüències humanes diferents a l'estudiada en cada moment. També es van eliminar aquelles seqüències amb una identitat de seqüència amb l'estudiada inferior al 95%. El conjunt total de mutacions neutres va ser de 11372.
Pel mètode basat en estructura es van obtenir fins a 888 NEMUs derivades de les proteïnes que tenien l'estructura resolta.

Selecció d'estructures per les proteïnes

Les diferents estructures per les diferents proteïnes estudiades es van obtenir de la base de dades PDB (Berman et al., 2000), seguint els criteris exposats en el capítol anterior.

Propietats derivades de la seqüència i de l'estructura

En la versió del mètode basat en la seqüència es van usar 19 paràmetres per descriure les mutacions. Fins a 23 paràmetres es van usar per la versió basada en l'estructura (veure figura 2).

Cal deixar clar que alguns paràmetres estan clarament relacionats, per exemple les dues mesures de grandària. Mentre que la reducció de dimensionalitat pot ser d'interès per tant de guanyar en generalització del mètode (Mehrotra et al.,

1997a), mantenir mesures correlacionades interessa per mostrar el màxim poder predictiu del mètode.

Propietats derivades de l'estructura

Dos paràmetres estructurals són usats en la versió basada en la seqüència: l'estructura secundària i l'accessibilitat al solvent. Els valors s'obtenen a partir de metodologies ben establerts de predicció estructural. En la versió basada en l'estructura s'usen sis paràmetres: estructura secundària, accessibilitat (en tres estats i relativa) i potencials estadístics (tres valors).

Prediccions d'estructura secundària i accessibilitat. Es van obtenir de les seqüències de les proteïnes, usant el paquet de programari PHD (Rost & Sander, 1993) (www.embl-heidelberg.de/predictprotein) amb els paràmetres definits per defecte. Els tres estats d'estructura secundària hèlix, beta i *coil* que van ser codificats com a 0, 1, 2 respectivament. Els tres estats d'accessibilitat, enterrat, semi-enterrat i exposat van ser codificats com a 0, 1, 2 respectivament.

Accessibilitat i estructura secundària observada. Van ser calculats a partir de l'estructura experimental de la proteïna. L'estructura secundària a la posició de la mutació va ser obtinguda usant el programa SSTRUC que implementa el mètode de Kabsch i Sander (Kabsch & Sander, 1983) per David Keith Smith. Els tres estats resultants van ser codificats com en el paràgraf anterior. Els valors d'accessibilitat van ser obtinguts usant el programa NACCESS (© S.J.Hubbard & J.M.Thornton, Department of Biochemistry and Molecular Biology, University College London). Els valors relatius i en tres estats van ser usats. Els valors relatius s'obtenen de dividir l'accessibilitat del residu en la proteïna entre l'accessibilitat en un pèptid estès Ala-X-Ala. L'accessibilitat en tres estats es va obtenir seguint el model de Rost i Sander (Rost & Sander, 1994). Així les

accessibilitats entre 0-9% corresponen a enterrat, accessibilitats en l'interval 9-36% correspon a semi-enterrat i finalment l'interval 36-100% correspon a exposat. Aquests tres estats van ser codificats igualment segons el paràgraf anterior.

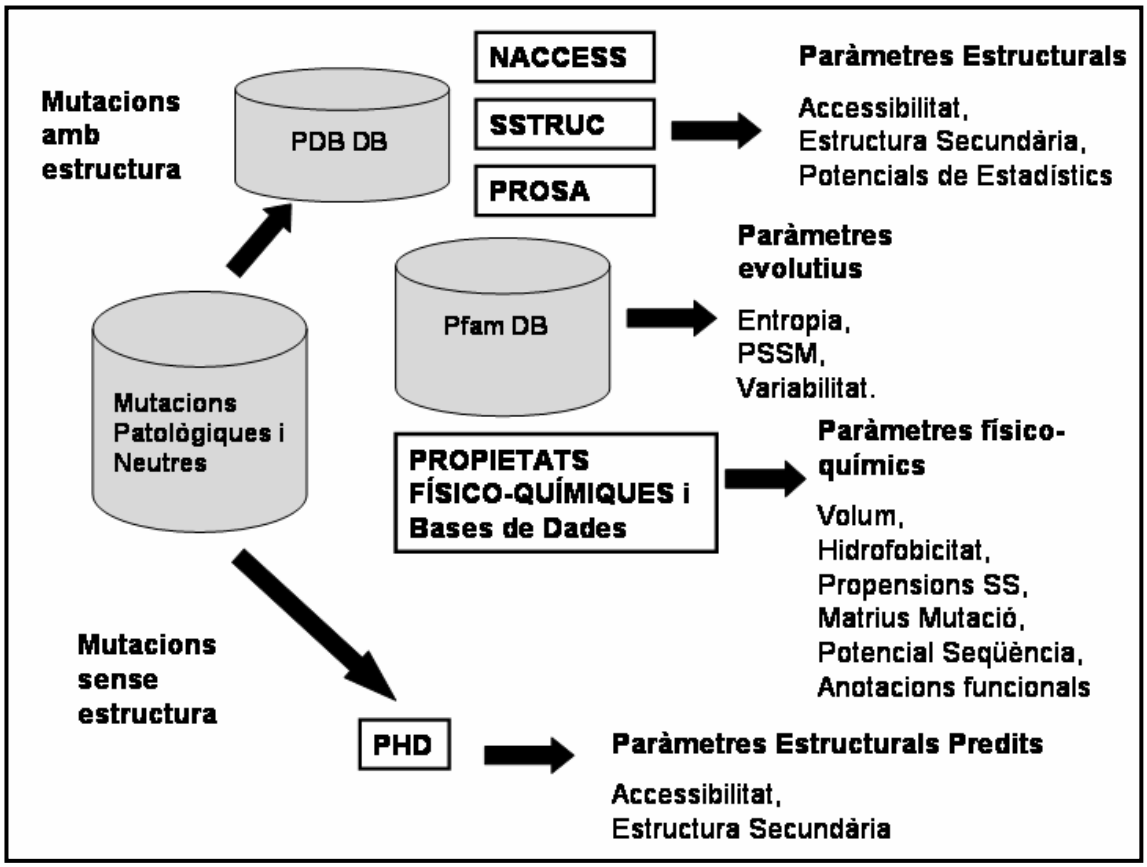

Fig2. Esquema de la determinació dels paràmetres de les mutacions.

Potencials estadístics: Van ser usats per establir efectes desestabilitzants dels mutants. Per cada residu en una proteïna d'estructura coneguda, el programa ProsaII (Sippl, 1993) dóna tres mesures indirectament relacionades amb la contribució del residu a l'estabilitat de la proteïna: són un potencial de superfície, un de contacte i la suma dels dos anteriors. Aquests tres valors es van calcular

pels residus natius i pels mutants. L'estructura dels mutants es van obtenir mapant el residu canviat sobre l'estructura sense fer cap mena de modelatge estructural. Els valors finals dels potencials es van obtenir de la diferència entre els valors dels mutants i dels salvatges.

Propietats de seqüència i de residu.

Tres tipus d'informació es van usar: matrius de mutació, propietats d'aminoàcids i potencials de seqüència.

Matrius de mutació. Per cada mutació s'obtenien els valors de les matrius Blosum62 (Dayhoff, 1978) i Pam 40 (Henikoff & Henikoff, 1992).

Canvis en propietats d'aminoàcids. Es van usar sis paràmetres basats en les propietats dels residus. Dos índexs hidrofòbics, dues propensions per estructura secundària i dos índexs de volums. Per cada índex es va calcular el valor associat a la mutació com la diferència entre els valor de l'aminoàcid mutant x_m i el valor de l'aminoàcid salvatge x_w. Els índex hidrofòbics es van obtenir de mesures de l'energia lliure de transferència entre aigua i octanol (Fauchere & Pliska, 1983), i de potencials estadístics derivats per Miller et al. a partir d'informació estructural (Miller et al., 1987). Les propensions d'estructura secundària es van obtenir de l'anàlisi estàndard de Chou i Fasman (Chou & Fasman, 1974) i també de l'anàlisi de Swindells et al. (Swindells et al., 1995). Els descriptors de volum es van obtenir dels volums de van der Waals (Bondi, 1964) i dels volums mitjos dels residus enterrats (Chothia, 1975).

Potencials de Seqüència. Per tal de tenir en compte els efectes de l'entorn de seqüència en l'efecte de la mutació (Yang et al., 1997) es van usar un potencial de seqüència simple, $Ptseq(r_j)$ (veure eq. 4), relacionat amb la probabilitat d'observar un residu r_j en la posició j en un donat entorn de seqüència.

$$Ptseq(r_j) = \ln\left[\prod_{i=-5}^{5} P(r_j / r_{j+i})\right] \qquad \text{(eq. 4)}$$

La $P(r_j / r_{j+i})$ es va obtenir de la manera següent:

$$P(r_j / r_{j+i}) = n(r_j, r_{j+i})/n(r_{j+i}) \qquad \text{(eq. 5)}$$

on $n(r_j,r_{j+i})$ és el número de parells d'aminoàcids de tipus r_j i r_{j+i} a la distància de seqüència i. $n(r_{j+i})$ es el número total de residus de tipus r_{j+i}. Aquests valors es van calcular usant el conjunt complet de seqüències humanes de la base de dades SwissProt (Boeckmann et al., 2003).

Les mutacions es van puntuar usant la diferència entre el valor de potencials de seqüència del mutant i del residu salvatge de la mateixa manera que les propietats relacionades amb els aminoàcids.

Paràmetres evolutius

La informació evolutiva es va explorar usant 4 paràmetres, dues mesures de la variabilitat aminoacídica i dos puntuacions relacionades amb l'especificitat de la posició. El càlcul d'aquests paràmetres necessiten l'ús d'alineaments múltiples de seqüència. Es van obtenir els alineaments de la base de dades Pfam (Bateman et al., 2002). Tot i que el mètode pot funcionar perfectament amb alineaments múltiples generats pel programari BLAST (Altschul et al., 1997).

Variabilitat en la posició de la mutació en un alineament múltiple de seqüència.

Va ser mesurada usant l'entropia de Shannon (Shannon, 1948) i una mitja de valors de matrius de mutació (Martin et al., 2002). L'entropia es va calcular segons la formulació $-\Sigma_i p_{ij}.\ln_2 p_{ij}$ on el subíndex recorre els diferents tipus

d'aminoàcids presents en la posició j, que correspon a la posició del residu mutat en l'alineament múltiple. p_{ij} correspon a la freqüència relativa de l'aminoàcid estudiat.

La segona mesura es calcula seguint la metodologia implementada per Martin et al. (Martin et al., 2002):

$$\left[\frac{\sum_{k=1}^{N} \sum_{l=k+1}^{N} s_{kl}}{\frac{N!}{2 \times (N-2)!}} \right] / S_{max} \qquad \text{(eq. 6)}$$

on s_{kl} correspon a l'element de Blosum62 (Henikoff & Henikoff, 1993) que correspon a la comparació de residus entre la seqüència k i la l en l'alineament múltiple de seqüència en la posició de la mutació estudiada. S_{max} correspon al valor màxim de s_{kl} en les seqüències de l'alineament.

Matrius de puntuació de posició específica (*Position-specific scoring matrices* PSSM) Es van usar dos tipus de paràmetres depenents de PSSM. El primer es basa en el logaritme de la relació de probabilitats $\log(p_{mj}/p_m)$, on p_{mj} correspon a la freqüència relativa de l'aminoàcid mutant de tipus m a la posició j de l'alineament múltiple de seqüències (j correspon aquí a la localització del residu mutant) i p_m correspon a la freqüència del mateix aminoàcid en totes les seqüències de SwissProt (Boeckmann et al., 2003). Per tal de minimitzar el problema de l'absència de dades es va usar una versió modificada del logaritme de la relació de probabilitats (Sippl, 1990):

$$\frac{N\sigma}{1+N\sigma} \log(p_{mj}/p_m) + \frac{1}{1+N\sigma} B62_{wm} \qquad \text{(eq. 7)}$$

on N correspon al número de seqüències en l'alineament múltiple de seqüències. σ correspon a un factor arbitrari que va ser assignat com 1/50(Sippl, 1990). w i m corresponen als residus salvatge i mutant respectivament. $B62_{wm}$ és l'element de la matriu de mutació Blosum62 (Henikoff & Henikoff, 1992) corresponent a la mutació que s'esdevé al passar del residu salvatge (w) al residu mutant (m). Quan N es petit, el valor de l'índex basat en PSSM s'apropa al valor $B62_{wm}$, però a mesura que N creix el valor de l'índex s'aproxima al valor del logaritme de la relació de probabilitats.

El segon paràmetre relacionat amb PSSM es calcula segons l'equació 8 i mostra el mateix comportament asimptòtic que l'índex anterior.

$$\frac{N\sigma}{1+N\sigma}\,[\log(p_{mj}/p_m) - \log(p_{wj}/p_w)] + \frac{1}{1+N\sigma}B62_{wm} \qquad \text{(eq. 8)}$$

on m, w, N, σ, p_{mj}, p_m i $B62_{wm}$ tenen el mateix significat que en l'índex anterior. p_{wj} és la freqüència relativa de l'aminoàcid salvatge de tipus w a la posició j en l'alineament múltiple de seqüències p_w és la freqüència del mateix aminoàcid en ls seqüències humanes presents a SwissProt (Boeckmann et al., 2003).

Informació de la base de dades

Quatre índexs diferents es van usar per tal d'incloure informació provinent d'anotacions funcionals presents en la base de dades SwissProt (Boeckmann et al., 2003). Aquests índexs indiquen si un residu mutat correspon a un residu que: (i) intervé en un pont disulfur, tiolester o tiolèter; (ii) cau dins una regió que sofreix *splicing* alternatiu; (iii) és un residu modificat (té un enllaç a carbohidrat o és una selenocisteïna); (iv) és un centre actiu, o uneix nucleòtids, Ca^{2+} or Zn^{2+}.

Per cada índex un valor de 0 indica absència d'anotació, un valor de 1 indica presència d'anotació.

La xarxa neural

Es va usar una xarxa neural de tipus *feed-forward* (Rumelhart et al., 1986) amb una capa d'entrada de dades i una o cap (model de perceptró) (Mehrotra et al., 1997b) capa oculta i una capa de sortida de dades. En total es van usar 19 paràmetres per la versió basada exclusivament en seqüència i 23 paràmetres per la versió que a més usava dades derivades de l'estructura tridimensional com a dades d'entrada de la xarxa. La sortida de dades de la xarxa era 0 per les prediccions neutres i 1 per les prediccions patològiques.

Es va seguir la metodologia d'entrenament descrita per Shepherd et al. (Shepherd et al., 1999) on es mostrava a la xarxa un seguit de dades amb la corresponent dada de sortida. Els pesos de la xarxa es van optimitzar usant la metodologia de gradients conjugats escalats durant 500 iteracions.

La validació creuada

El rendiment del nostre mètode es va avaluar usant una validació creuada heterogènia i estringent de 5 etapes (Krishnan & Westhead, 2003). En comptes de dividir tot el grup de dades de DAMUS i NEMUS en 5 subgrups i aleshores entrenar les xarxes neurals es va dividir el grup de mutacions a nivell de proteïna. D'aquesta manera es va dividir el grup de proteïnes que almenys tenien una mutació patològica en 5 grups. Els corresponents 5 grups de mutacions es van construir de manera que 5 úniques combinacions de 4 subgrups diferents es van usar per entrenar les xarxes. Per cada combinació dels 4 diferents subgrups es va assajar en el cinquè grup exclòs. Pels resultats de les prediccions pels 5 grups d'assaig se'n va obtenir la mitjana que es mostra en aquest grup.

Mesures del rendiment

Quatre paràmetres diferents es van usar per avaluar el rendiment de les xarxes neurals: Percentatge de prediccions correctes (també referit com taxa d'èxit global; veure l'equació 9); percentatge normalitzat d'enriquiment sobre les prediccions a l'atzar (S, (Shepherd et al., 1999), veure equació 10) i percentatge de falsos positius (FP) i de falsos negatius (FN) veure equacions 11 i 12. El primer paràmetre dóna una visió global de la capacitat del mètode per detectar mutacions patològiques i neutres. L'índex S garanteix una mesura de la capacitat del mètode en relació amb una predicció purament basada en l'atzar (Shepherd et al., 1999). Finalment les taxes sobre falsos positius i falsos negatius garanteix una idea dels límits de l'exactitud de la metodologia emprada.

$$Q_{tot} = 100\frac{cp}{(cp+ip)} \qquad \text{(eq. 9)}$$

on cp i ip són el número global de prediccions correctes i incorrectes respectivament.

$$S = \frac{(p+n)-R}{t-R} \times 100 \qquad \text{(eq. 10)}$$

on $t = p+n+o+u$ i $R = \left[(p+o)\times(p+u)\times(n+u)\times n+o)\right]/t$. p, n són el nombre de mutacions correctament predites com a DAMUs i NEMUs, respectivament. o i u són el número de mutacions incorrectament predites com a DAMUs i NEMUs respectivament.

$$FP = \frac{o}{p+o} \times 100 \qquad \text{(eq. 11)}$$

$$FN = \frac{u}{p+u} \times 100 \qquad \text{(eq. 12)}$$

on p, n, o i u tenen el mateix significat que anteriorment.

Índex de fiabilitat

L'índex de fiabilitat es calcula a partir del resultat de sortida de la xarxa neural segons la formulació següent (Shepherd et al., 1999): integer[abs(NN_{output} − 0.5) x 20)], on NN_{output} correspon a la sortida de la xarxa neural. L'índex resultant varia entre 0 i 9.

III. ARTICLE DE RECERCA

Sequence-based prediction of pathological mutations. Ferrer-Costa, C, Orozco, M, de la Cruz, X. Proteins **57**: 811-19. 2004

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

PROTEINS: Structure, Function, and Bioinformatics 57:811–819 (2004)

Sequence-Based Prediction of Pathological Mutations

C. Ferrer-Costa,[1] M. Orozco,[1,2*] and X. de la Cruz,[1,3*]
[1]*Molecular Modeling and Bioinformatics Unit, Institut de Recerca Biomédica, Parc Científic de Barcelona, Barcelona, Spain*
[2]*Departament de Bioquímica i Biologia Molecular, Facultat de Química, Universitat de Barcelona, Barcelona, Spain*
[3]*Institució Catalana per la Recerca i Estudis Avançats (ICREA), Barcelona, Spain*

ABSTRACT **The development of methods to assess the impact of amino acid mutations on human health has become an important goal in biomedical research, due to the growing number of nonsynonymous SNPs identified. Within this context, computational methods constitute a valuable tool, because they can easily process large amounts of mutations and give useful, almost cost-free, information on their pathological character. In this paper we present a computational approach to the prediction of disease-associated amino acid mutations, using only sequence-based information (amino acid properties, evolutionary information, secondary structure and accessibility predictions, and database annotations) and neural networks, as a model building tool. Mutations are predicted to be either pathological or neutral. Our results show that the method has a good overall success rate, 83%, that can reach 95% when trained for specific proteins. The methodology is fast and flexible enough to provide good estimates of the pathological character of large sets of nonsynonymous SNPs, but can also be easily adapted to give more precise predictions for proteins of special biomedical interest. Proteins 2004;57:811–819.** © 2004 Wiley-Liss, Inc.

Key words: SNPs; disease-associated mutations; protein sequence; bioinformatics; neural networks; sequence variability

INTRODUCTION

Identifying the pathological character of the vast amount of known nonsynonymous SNPs (single nucleotide polymorphisms) has become an important challenge for biomedical sciences.[1,2] This has lead to the development of different computer-based methods to predict the pathological character of amino acid mutations.[3–11] Many of these methods were created to provide an initial and rapid identification of disease-associated mutations from large sets of nonsynonymous SNPs, allowing their ranking for subsequent experimental study.[5,7] To this end, amino acid mutations are mapped to the protein sequence and then characterized in terms of structure and sequence properties.[3,4,6,7,9–14] These properties are subsequently used to predict whether a given mutation is likely to be pathological.

The rationale behind these methods is that the disease-causing effect of many mutations can be understood in terms of their effect on protein structure,[6,12,13,15–17] e.g., stability losses, disruption of protein interactions, etc. This is supported by the fact that some structure-based properties, like solvent accessibility, may have a high discriminating power between neutral and disease-associated mutations.[7,12,13] Unfortunately, the rapidly growing number of proteins for which no structure is available,[18] seriously limits the applicability of structure-based methods. One possible way to overcome this problem is the use of evolutionary information alone, derived from multiple sequence alignments.[4,11,14] Some authors have fruitfully tested this idea in small sets of specific systems: Miller and Kumar,[14] studying a set of six proteins and their mutations, have recently shown that disease mutations tend to be overabundant at highly conserved positions in multiple sequence alignments; Santibáñez-Koref et al.[11] have used evolutionary information to assess the significance of missense mutations in the case of P53. On the contrary, Ng and Henikoff[19] use many different proteins to test their prediction method, also based on the use of evolutionary information, with promising results. However, overall error rates are still high (see Table 4 in Ng and Henikoff[19]), leaving room for improvement.

An interesting alternative has been recently proposed[7] where mutations are characterized using evolutionary information together with structure properties obtained from de novo protein structure predictions. Preliminary results are encouraging,[7] showing that the recognition power of evolutionary information can be extended by the use of predicted structure properties.[7] However, this method is limited by the fact that, for the moment, computer-costly de novo predictions can only be obtained routinely for small protein domains[20] and cannot be applied to membrane proteins. Interestingly, in the case of mutations affecting viability in microbial systems, Krishnan and Westhead[9] have proposed that predicted structural properties could be used to complement evolutionary properties.

In the present article we extend the idea of combining evolutionary information together with structure informa-

Abbreviations: SNP, single nucleotide polymorphism; DAMU, disease-associated mutation; NEMU, neutral mutation; NN, neural network.

*Correspondence to: Xavier de la Cruz or Modesto Orozco. Parc Científic de Barcelona, C/Josep Samitier, 1-5, 08028 Barcelona, Spain. E-mail: xavier@mmb.pcb.ub.es or, modesto@mmb.pcb.ub.es.

Received 30 January 2004; Accepted 1 June 2004

Published online 10 August 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20252

tion for the prediction of human disease-associated mutations (DAMUs). We propose a prediction method in which mutations are labelled using only sequence-derived information. The structure properties utilized (secondary structure and accessibility) are obtained using standard, fast, sequence-based methods. In addition to evolutionary and structure information, residue properties and database annotations are also used. All this information is input to a neural network (NN) that provides the final prediction on whether the target mutation is pathological, or is a neutral mutation (NEMU). The resulting predictive model has a high overall success rate, 83.5%, which represents a 66.5% improvement over a random prediction approach. We find that evolutionary information provides a substantial amount of the discrimination power of the method, although the contribution of other parameters is non-negligible. Our results indicate that the method can be applied to the quick characterization and ranking of mutations, supporting its future use for the priorization of nonsynonymous SNPs, within the context of biomedical studies.

MATERIALS AND METHODS

Prediction Protocol

The goal of our work is to devise a computational method to predict whether amino acid mutations are pathological or not. To this end, mutations are labelled with a set of properties (size change, location in structure, etc.) that can be related to their possible damaging effect to the protein. The resulting set of properties is then used by a NN to decide if the mutation is pathological or neutral.

Sets of Mutations

Disease-associated mutations

There are two main repositories of human disease-associated mutations (DAMUs) at present:[21] OMIM[22] and SwissProt.[23] We chose to work with SwissProt, a manually curated database, because recent work[21] indicates that it provides a more accurate mapping of the mutations onto the protein sequence. This is particularly relevant in cases like ours, were several thousands of mutations are utilized, and automatic processing is required. In addition, DAMUs from SwissProt have been utilized by authors working in the study and prediction of pathological mutations.[5,7,13,19] Finally, DAMUs from SwissProt and those from OMIM have very similar properties at the protein structure and sequence levels, e.g., we find that 39.1% of DAMUs happen at buried residues, and Steward et al.,[17] who use OMIM data, find 39%; we also find that DAMUs are more frequent at conserved sites, in agreement with Steward et al.[17]

We utilized version 40 of SwissProt,[23] and obtained the set of DAMUs following Ng and Henikoff.[19] To this end, we searched SwissProt using as keywords DISEASE, VARIANT, and HUMAN. We discarded mutations for which no clear link to disease could be established. Finally, mutations for which only the human sequence was available in the Pfam[24] alignment for the protein family were also discarded (more than one sequence is required to compute some evolutionary properties used in the prediction process, see below). A total of 9334 DAMUs, happening in 811 human proteins were kept.

In the structure-based case, we used a set of mutations that could be mapped to the protein structure. This gave a total of 1319 DAMUs happening in 90 proteins.

Neutral mutations

Two models of neutral mutations (NEMUs) have been used in this work. The first was derived from a massive mutagenesis experiment done in *Escherichia coli* Lac repressor protein,[25] keeping those mutations that result in organisms phenotypically similar to the wild-type bacteria.[4] We have utilized this dataset because it has been used by Ng and Henikoff[19] to test their sequence-based method, thus allowing a direct comparison between both approaches.

In the second model for NEMUs, which we call the evolutionary model, mutations were collected only for the 811 human proteins for which DAMUs were available. For these proteins, NEMUs corresponded to those variants occurring in members of the protein family from other species.[12,13] To this end, we took the Pfam[24] multiple sequence alignment for the protein family, and eliminated: (1) all the human sequences but that of the human target protein; (2) nonhuman sequences with less than 95% identity to the human target sequence. Any amino acid change between the target human sequence and the remaining sequences was then considered as a neutral mutation. A total of 11372 NEMUs were obtained.

For the structure-based method we used a subset of 888 NEMUs from the evolutionary model that mapped into the protein structure.

Selection of Structural Data for Proteins

Protein structures were downloaded from the PDB.[26]

Sequence and Structure-Based Properties

In the sequence-based version of our procedure we used 19 parameters to label mutations; 23 were used in the structure-based version.

It will be noted that some of the parameters are clearly related, e.g., the two measures of residue size. While dimensionality reduction is advisable in order to gain generalization performance,[27] our results [Table I(A)] show that good generalization is already attained with our NN models. In addition, it may be advisable to keep a larger input in order to avoid losing useful parameters.[28] For this reason we decided to work on the full model and then analyze separately the contribution of the different parameters (see Results and Discussion), in order to provide a better understanding of what makes our model successful.

Structure properties

Two structure-related parameters, secondary structure and solvent accessibility, were utilized in the sequence-based version of our method, in which no experimental three-dimensional information whatsoever was used. They

TABLE I. Performance of Our Approach for Predicting Disease-Associated Mutations†

(A)	Qtot[a]	S[b]	FP[c]	FN[d]
Ng and Henikoff[e]	70.3	37.0	14.7	30.5
SWP-Lac[f]	87.3 (± 0.3)	54.2 (± 2.5)	10.4 (± 0.6)	4.6 (± 0.5)
SWP-Evol[g]	83.5 (± 0.3)	66.5 (± 0.5)	16.5 (± 0.6)	21.1 (± 0.6)
PDBst[h]	87.0 (± 1.6)	73.0 (± 3.1)	10.4 (± 2.1)	11.5 (± 1.3)

(B)	Qtot[a]	S[b]
Pssm[i]	77.6 (± 0.5)	49.1 (± 1.1)
ΔPssm[j]	73.1 (± 0.6)	39.7 (± 0.9)
B62[k]	70.0 (± 0.4)	35.7 (± 0.7)
PAM[l]	68.6 (± 0.6)	29.0 (± 0.7)
Var.[m]	68.0 (± 0.7)	25.4 (± 1.1)
Entr.[n]	66.5 (± 0.4)	24.6 (± 1.0)
VdW.Vol[o]	62.1 (± 0.3)	11.4 (± 0.3)
Vol.Bur.[p]	61.7 (± 0.4)	10.3 (± 0.5)
W/O trf.[q]	59.7 (± 0.5)	8.8 (± 2.1)

(C)	Qtot[a]	S[b]
Pssm[i]	77.7 (± 0.5)	49.9 (± 1.4)
ΔPssm[j]	73.2 (± 0.7)	38.9 (± 1.2)
B62[k]	70.0 (± 0.4)	35.7 (± 0.7)
VdW.Vol[o]	68.5 (± 0.1)	30.5 (± 0.3)
PAM[l]	68.3 (± 0.5)	61.9 (± 0.7)
Vol.Bur.[p]	68.0 (± 0.1)	31.7 (± 0.6)
I/O prot.[r]	67.9 (± 0.4)	28.7 (± 0.9)
Var.[m]	67.9 (± 0.7)	24.5 (± 1.1)
Entr.[n]	66.6 (± 0.5)	22.3 (± 0.8)
W/O trf.[g]	65.4 (± 1.7)	22.6 (± 5.6)
Acc.[s]	62.2 (± 0.6)	46.9 (± 2.0)
Chou and Fassman[t]	57.0 (± 0.9)	6.8 (± 1.0)

†(A) Performance using all available information. (B) and (C) Performance of the perceptron and the H12 neural networks, respectively, when trained using specific parameters only.
[a]Overall success rate (see Methods).
[b]Normalized performance relative to random (see Methods).
[c]False positive rate.
[d]False negative rate.
[e]Performance of the method by Ng and Henikoff, derived from data provided by Ng and Henikoff[19] (Table 4 in reference).
[f]Cross-validated performance of our method when tested with NEMUs from the Lac system (see Methods).
[g]Cross-validated performance of our method when tested with NEMUs from the evolutionary model (see Methods).
[h]Cross-validated performance of our method when tested with NEMUs from the evolutionary model using observed values for the structure-related properties (accessibility, secondary structure and statistical potentials; see Methods).
[i–t]Cross-validated performance for the NN when using as input only one of the following parameters: Pssm, position-specific scoring matrices (Equation 4 in Methods); ΔPssm, difference in position-specific scoring matrices (Equation 5 in Methods); B62, blosum62 matrix elements; PAM, PAM40 matrix elements; Var., average of mutation matrix scores (Equation 3 in Methods); Entr., Shannon entropy; VdW.Vol., van der Waals amino acid volumes; Vol.Bur., volume of buried residues; W/O trf., water/octanol transfer free energy; I/O prot., structure-derived hydrophobicity; Acc. PHD accessibility predictions; Chou and Fassman, Chou and Fassman secondary structure propensity. All these parameters are explained in the Methods section.

were obtained from sequence-based predictions (see below). In the structure-based version of our approach we used six structure-related parameters: observed secondary structure and solvent accessibility (three-state and relative), and statistical potentials (three values).

Predicted secondary structure and accessibility. They were obtained from the protein sequence, using the PHD software package[29] (www.embl-heidelberg.de/predictprotein) with default parameters. The three secondary structure states, helix, beta, and coil, were encoded as 0, 1, and 2, respectively. The three accessibility states, buried, half-buried and exposed, were encoded as 0, 1, and 2, respectively.

Observed secondary structure and solvent accessibility. They were computed from the experimental structure of the protein. Secondary structure at the mutation site was obtained using the SSTRUC implementation of the Kabsch and Sander method,[30] by David Keith Smith. The resulting three-states were encoded as in the previous section. Accessibility values were obtained using the program NACCESS.[31] Both relative and three-state accessibilities were used. The former are equal to the residue accessibility in the protein divided by its accessibility in an extended Ala-X-Ala peptide. Relative accessibilities were mapped to three-state accessibilities: buried (0–9% relative accessibility), half-buried (9–36% relative accessibility) and exposed (36–100% relative accessibility). These three-states were encoded as in the previous section.

Statistical potentials. They were used to assess the destabilizing effect of the mutant. For each residue in a protein of known structure, Prosa II[32] calculations gave us three measures indirectly related to the residue contribution to protein stability: surface potential, contact potential, and an overall potential (weighted sum of the two previous potentials). The three terms were computed for the native residue, for the mutated one (the structure of the mutated protein was obtained by just mapping the residue change in the wild-type structure, no structural modeling was done). The mutation was then labeled with the three corresponding differences between mutant and native values.

Residue/sequence properties

Three information types were used: mutation matrices, amino acid properties and sequence potentials.

Mutation matrices. Each mutation was labelled with both the Blosum62[33] and PAM40[34] matrix values.

Changes in single amino acid properties. We used six residue-based parameters: two hydrophobic indexes, two secondary structure propensities, and two volume changes. For each of them, the value associated to the mutation was the difference between the mutant, x_m, and wild-type, x_w, values of this property: $x_m - x_w$. The hydrophobic parameters were taken from water/octanol free energy measurements,[35] and from statistical potentials derived by Miller et al.[36] from structural information. Secondary structure propensities were obtained from standard Chou and Fasman analysis,[37] as well as from the Swindells et al. analysis.[38] Size descriptors were van der Waals volumes,[39] and volume of buried residues.[40]

Sequence potential. To take into account the effect of sequence environment on the mutation effect[41] we used a

simple potential, Ptseq(r_j) (see equation 1), related to the probability of observing residue r_j at position j, in a given sequence environment:

$$Ptseq(r_j) = \ln\left[\prod_{i=-5}^{5} P(r_j/r_{j+i})\right] \quad (1)$$

The $P(r_j/r_{j+i})$ were obtained following:

$$P(r_j/r_{j+i}) = n(r_j,r_{j+i})/n(r_{j+i}) \quad (2)$$

where $n(r_j,r_{j+i})$ is the number of pairs of amino acids of types r_j and r_{j+i} at a sequence distance i. $n(r_{j+i})$ is the total number of residues of type r_{j+i}. These numbers are computed using the whole set of human sequences from the SwissProt database.[23]

Mutations were scored using the difference between the value of the sequence potential for the mutant and wild-type residues, as for the simple amino acid properties.

Evolutionary properties

Evolutionary information was exploited using four parameters: two measures of the amino acid variability, and two position-specific scores. Computation of these parameters required the use of multiple sequence alignments, that were obtained from the Pfam[24] database. Pfam alignments (27) were chosen for their quality, however, the method can be easily extended to other kinds of multiple sequence alignments, in particular to those generated by the program BLAST.[42]

Variability at the mutation position in the multiple sequence alignment. This was measured using the Shannon entropy[43] and an average of mutation matrix scores.[16] The former was computed following: $-\Sigma_i p_{ij}.\ln_2 p_{ij}$, where subindex i runs over the different amino acid types found at position j, location of the mutated residue in the multiple sequence alignment. p_{ij} are the relative frequencies of these amino acids.

The second measure was computed following Martin et al.[16]

$$\left[\frac{\sum_{k=1}^{N}\sum_{l=k+1}^{N} s_{kl}}{\frac{N!}{2 \times (N-2)!}}\right] \Big/ S_{\max} \quad (3)$$

where s_{kl} is the element of Blosum62[33] corresponding to the comparison between residues in sequences k and l in the multiple sequence alignment, at the position of the mutation under study. S_{max} is the larger s_{kl}. N is the number of sequences in the alignment.

Position-specific scoring matrices (PSSM). We utilized two kinds of PSSM parameters. The first is based on the log-odds ratio: $\log(p_{mj}/p_m)$, were p_{mj} is the relative frequency of the mutant amino acid type m at position j in the multiple sequence alignment (j is here the location of the mutated residue) and p_m is the frequency of the same amino acid type in all human sequences in SwissProt.[23] To alleviate the problem of missing data, we used a modified version of the log-odds ratio:[44]

$$\frac{N\sigma}{1+N\sigma}\log(p_{mj}/p_m) + \frac{1}{1+N\sigma}B62_{wm} \quad (4)$$

where N is the number of sequences in the multiple sequence alignment, σ is an arbitrary factor, and was taken[44] equal to 1/50. w and m stand for the normal and mutant amino acid, respectively. $B62_{wm}$ is the element of the mutation matrix Blosum62[33] corresponding to the mutation from the wild-type residue (w) to the mutant residue (m). When N is small, the value of the PSSM index approaches that of $B62_{wm}$; when it is large it approaches the value of the log-odds ratio.

The second PSSM parameter is determined following equation 5, and shows the same asymptotic behavior as the previous index.

$$\frac{N\sigma}{1+N\sigma}[\log(p_{mj}/p_m) - \log(p_{wj}/p_w)] + \frac{1}{1+N\sigma}B62_{wm} \quad (5)$$

where m, w, N, σ, p_{mj}, p_m, and $B62_{wm}$ have the same meaning as before. p_{wj} is the relative frequency of the original amino acid type w at position j in the multiple sequence alignment. p_w is the frequency of the same amino acid type in human sequences in SwissProt.[23]

It has to be noted that a certain amount of redundancy is to be expected in the multiple sequence alignments used, thus reducing the contribution of these variables to the success of the method. While this problem can be alleviated using a filtering procedure, previous results from our group suggest that differences between raw and filtered data are minor.[13]

Database information

Four indexes were used to include database functional annotations taken from the SwissProt[23] database. They indicate that the mutated residue is: (1) involved in disulfide bridge, thiolester, thiolether; (2) inside an alternative splicing region; (3) a modified residue (has a bound carbohydrate, is a selenocysteine); (4) an active site, or binds nucleotide, Ca^{2+} or Zn^{2+}. For each index, a value of 0 indicates absence of annotation, and a value of 1 indicates presence of annotation.

The Neural Network

A feed-forward neural network,[45] with one input layer one, or none (perceptron model[27]) hidden layer and one output layer was used. For the NN having a hidden layer, the latter was constituted by 20 units for the full version of the method, and by 2 units for the study of the individual contributions of the different parameters. A total of 19 (sequence-only version) or 23 (sequence plus structure version) parameters were input to the network—the mutation labels described above.

For a given mutation, the network output is a number comprised between 0 and 1, that is transformed into a discrete prediction as follows: for values above 0.5, the mutation is predicted as DAMU; for values below 0.5, the

mutation is predicted as NEMU. Strong predictions will be those for which the output is either very close to one, or very close to zero, e.g., 0.9 or 0.01, respectively. Weak predictions will be those with an output closer to 0.5, e.g. 0.44 or 0.58.

We followed the training procedure described in Shepherd et al.,[46] presenting the network with a number of inputs, together with their associated target outputs. The network weights were optimized using scaled conjugate gradients with 500 iterations.

Cross-Validation

The performance of our method was evaluated using a stringent heterogeneous 5-fold cross-validation procedure.[9] Rather than randomly dividing the whole mutations set (DAMUs plus NEMUs) into five subsets and subsequently train the NN, the mutation dataset was split at the protein level. Thus, we divided into five parts the set of proteins for which at least one DAMU was found. The corresponding five sets of mutations were then built and, subsequently, each of the five unique combinations of four different subsets was used to train the network. For each combination, the network was tested on the excluded subset. The results for the test sets were then averaged to provide the results shown in this work.

Performance Measures

We utilized four parameters to evaluate the performance of our prediction method: percentage of correct predictions, normalized percentage of improvement over random predictions, false positive and false negative rates. We describe them below.

Percentage of correct predictions (Q_{tot}, also referred to as overall success rate). It provides an overall view of the ability of the procedure to detect pathological/neutral mutations. It is computed as follows:

$$Q_{tot} = 100\,\frac{cp}{(cp + ip)} \tag{6}$$

where cp and ip are the overall number of correct and incorrect predictions respectively.

Normalized percentage of improvement over random predictions (S). This parameter is a measure of how well the method is working relative to a random predictor with expected performance R, normalized to eliminate the scale effect from R and from the total number of observations.[46] It varies between 0% (no improvement relative to a random method) and 100% (perfect, non-random, predictions). It is computed as follows:

$$S = \frac{(p + n) - R}{t - R} \times 100 \tag{7}$$

where $t = p + n + o + u$. p, n are the number of mutations correctly predicted as DAMUs and NEMUs, respectively; o, u are the number of mutations incorrectly predicted as DAMUs and NEMUs, respectively. $R = [(p + o).(p + u).(n + o).(n + u)]/t$ is the expected number of correctly classified mutations generated by a random predictor that would take into account the proportion of DAMUs and NEMUs in our sample.

Percentage of false positives (FP) and percentage of false negatives (FN). They provide an idea of the accuracy limits of our approach. They are computed as follows:

$$FP = \frac{o}{p + o} \times 100 \tag{8}$$

$$FN = \frac{u}{p + u} \times 100 \tag{9}$$

where p, n, o and u have the same meaning as before.

Reliability Index

The reliability index is computed from the NN output following:[46]

$$\text{integer}[\text{abs}(\text{NN}_{\text{output}} - 0.5) \times 20)] \tag{10}$$

where $\text{NN}_{\text{output}}$ is is the NN output. The resulting index varies between 0 and 9, with low values of the index corresponding to poorer predictions, and high values corresponding to better predictions.

RESULTS AND DISCUSSION

As mentioned before, in this work we present a computational procedure to distinguish between disease-associated mutations and neutral mutations, utilizing sequence-based information and neural networks. The performance of our method has been tested on a large set of human DAMUs (9334 mutations). Some authors[3,9] have utilized sets of mutations affecting viability in microbial systems, as a model for human DAMUs. However, extrapolation to humans of the results obtained with microbial systems may be very complex and it is still an open issue.[9] In our case, we have used actual human DAMUs, because large amount of them are available in public databases and have been used for the characterization of human pathological mutations.[7,13,17,47] In particular, DAMUs used in this work were obtained from the SwissProt database,[23] following Ng and Henikoff[19] (see Methods).

For the NEMUs, we have utilized two different models (see Methods): (1) the Lac dataset, because it has been previously used by Ng and Henikoff[19] for their sequence-based method, thus allowing a direct comparison between methods, and (2) an evolutionary model, used by Bork and coworkers for their structure-based method,[5] and by Santibáñez-Koref et al.[11] for their sequence-based prediction method of deleterious mutations in p53.

Our method shows a high overall success rate, 87.3 (±0.4), when utilized to discriminate between DAMUs and NEMUs from the Lac dataset, with a substantial improvement over purely random predictions, 54.2 (±2.5). The false positive and false negative rates are 10.4 (±0.6) and 4.6 (±0.5), respectively. We can compare our results with those obtained by Ng and Henikoff,[19] who have derived the other sequence-based method described for the prediction of pathological mutations in large sets of different proteins. We can see that our method has a higher performance [Table I(A)]. Both overall success rate, 87.3 versus

70.3, and performance relative to random, 54.2 versus 37.0, are better in our case. The same is true for the false positive and false negative rates [see Table I(A)]. We believe that this improved performance of our method is probably due to the fact that we use more features to characterize mutations, the way we process evolutionary information, and the pattern recognition power of NNs.

When testing the method with the evolutionary model for NEMUs we also find a high performance rate, 83.5 (±0.3), with a clear improvement of 66.5 (±0.5) over purely random predictions. The false positive and false negative rates are 16.5 (±0.6) and 21.1 (±0.6), respectively. The overall performance of the method is similar to that obtained when utilizing NEMUs from the Lac system, with a better false positive and negative rate for the latter, but a significantly higher improvement over random predictions found with the evolutionary model. This is because the method's performance figures for neutral mutations are clearly better with the latter model than when using the Lac model for NEMUs: false positive rates of 16.6 (±0.5) and 26.6 (±1.2), respectively; false negative rates of 12.8 (±0.6) and 46.9 (±3.6), respectively. Compared with the method by Ng and Henikoff[19] the results of the second version of our method are also clearly better.

Taken together, the results from both tests indicate that DAMUs can be distinguished from NEMUs with a high overall success rate—between 83% and 87%, using only sequence-based information. The use of two different models for NEMUs does not affect the overall performance of the method, although utilizing NEMUs from the evolutionary model results in more balanced results for both DAMUs and NEMUs. For this reason, and because data from the Lac system are more likely to be biased towards this specific system, in the following we will only discuss performance figures corresponding to our method when trained in the set of evolutionary NEMUs (it has to be noted that the results obtained with the Lac dataset lead to similar conclusions).

Factors Contributing to the Recognition Rate of Our Approach

The fact that no structure information was required for the identification of DAMUs is very important from a practical point of view, because there is still a large number of proteins for which only sequence information is available.[48] To understand why this happens, we studied the discriminant power of the nineteen parameters used. To this end, for each of them we assessed the performance of two NNs: a simple perceptron and one with a two-units hidden layer (H12); both trained with the corresponding parameter as only input. We utilized these two NNs because the discriminant power of some variables may be substantially underestimated when using the perceptron.[28] Within each NN model, the results for the different parameters were ranked according to the overall success rate [Table I(B) for the perceptron and Table I(C) for H12].

First, we observe that none of the parameters alone shows a performance comparable to that of the full version of the method, indicating that several of them are required for the success of the latter. However, we also find that the best performance is always obtained with the evolutionary indexes—the two PSSM parameters, in accordance with previous studies,[7,19] with a success rate clearly above that from the other variables. Two reasons can explain the large predictive power of evolutionary information as measured in this work. First, residue conservation in protein families is directly related to its contribution to protein stability or function. Therefore, because a large number of DAMUs are likely to cause protein destabilization,[6] they can be modelled using this kind of information. The same applies for the most easy-to-understand DAMUs, those affecting the functional site of the protein.[5] The second reason is the fact that the DAMUs considered here are responsible of monogenic disorders, and thus more likely to have large damaging effects on protein function, or stability.[49] Because of this, they will tend to appear as clear departures from the multiple sequence alignment conservation pattern at the mutation site, making their identification easier.

The highest discrimination power observed for PSSM is paralleled, at the distribution level, by the smallest overlap between the distributions for NEMUs and DAMUs, relative to this parameter [Fig. 1(A)]. Positive and negative values of the latter tend to correspond to NEMUs and DAMUs, respectively. Interestingly, and in agreement with this picture, both NNs just use a single threshold to distinguish between NEMUs and DAMUs [Fig. 1(A)].

PSSM parameters are followed by mutation matrix elements, both Blossum 62 and PAM 40 matrix elements [Table I(B,C)]. However, for the H12 neural network van der Waals volumes give a slightly better success rate than PAM 40 matrix elements. This is in contradiction with the results obtained for the perceptron, where van der Waals volumes show clearly worse discrimination ability. This can be explained when looking at the distribution of mutations relative to volume changes [Fig. 1(B)]. We can see that NEMUs prevail in an interval around zero, that is, they are associated with volume changes smaller than those of DAMUs. This is consistent with the fact that large volume changes, whether positive or negative, are more likely to cause substantial structural damage.[13] In accordance with this picture, H12 defines an interval inside which mutations are predicted as NEMUs, and outside which they are predicted as DAMUs [Fig. 1(B)]. On the contrary, the volume-based perceptron follows a much poorer strategy, using a single threshold [Fig. 1(B)] to separate mutations in DAMUs or NEMUs, if their associated volume changes are above or below the threshold, respectively.

Comparing Table I(B) and (C) we can see that the parameters shown can be divided in two groups: (1) those for which the performance is almost the same for both NN models; (2) those that show a better performance with the H12 neural network. Analysis of the frequency histograms (Fig. 1, and results not shown) indicates that for members of the first group NEMUs and DAMUs tend to have values of opposite sign for the corresponding parameter. For this reason, a unique threshold value may be enough to provide

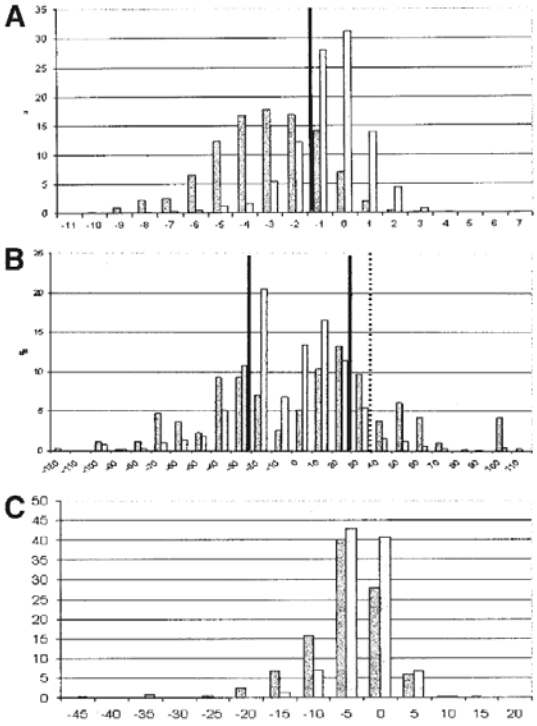

Fig. 1. Distribution of DAMUs (dark grey) and NEMUs (light grey) relative to different scoring parameters (see Methods for the definitions): (**A**) Position-specific scoring matrices (see equation 4, in Methods), (**B**) van der Waals volumes, (**C**) overall statistical potential. In (**A**), the thick bar corresponds to the threshold value below which mutations are predicted as pathological, and above which as neutral, by both the perceptron and the H12 neural networks (see text). In (**C**), the thick bars define the interval within which mutations are predicted as pathological, and outside which as neutral, by the H12 neural network. The dotted line corresponds to the threshold defined by the perceptron, above which mutations are predicted as pathological, and below which as neutral.

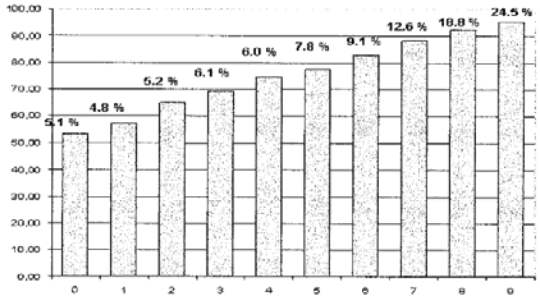

Fig. 2. Relationship between the percentage of correct predictions, Q_{tot}, (ordinates) and reliability (abscissae) of the predictions (see text). The percentage of pathological mutations predicted with a given reliability is shown above each bar.

an almost maximum discriminatory power, a strategy implicit in the perceptron. Members of this first group include: the two PSSM-related indexes, the two variability measures, and the Blossum 62 matrix. On the contrary, for members of the second group, it is the absolute value of the change rather than its sign, which determines the damage caused by the mutation and thus its association to disease. Therefore, good discrimination between mutations will require defining an interval within which mutations will be of one type, and outside which of the other type. This strategy can be obtained with more complex NN models, like H12. Members of this second group include the two volume and hydrophobicity properties, and predicted accessibility. Within this second group, we can also include the Chou and Fassman secondary structure propensities, although their predictive value is very poor. PAM 40 matrix shows features of both groups, with a clear improvement in the performance relative to random when using H12. Parameters not appearing in Table I(B) and (C) led to very small improvements relative to random, below 5%, for which reason they were excluded from this analysis (these parameters are: secondary-structure predictions, sequence potential, coil-derived secondary-structure propensities, and SwissProt annotations).

Reliability of the Predictions

Although the average success rate of our approach is reasonably good, it may be of interest to know how reliable are individual predictions, in order to prioritize mutations for further analysis. To this end we can use the NN output which is related to the nature of the damage caused by the mutation. For example, pathological mutations strongly disrupting the multiple sequence alignment conservation pattern at the mutation site will result in large, negative, PSSM values. The latter will, in turn, lead to strong DAMUs predictions by the NN.

The NN output can be easily transformed into a reliability index that will reflect the prediction strength, using equation 10 in the Methods section. The resulting index varies between 0 (unreliable predictions) and 9 (highly reliable predictions); it is directly related to the performance of the method, with better performances corresponding to higher values of the reliability index (Fig. 2). The latter confirms that the reliability index can be a useful tool for filtering sets of mutation data, ranking mutations according to the prediction reliability, and eliminating those with prediction reliabilities below a given threshold.

Improving the Performance of Our Approach

It has been shown previously that both predicted secondary structure and accessibility had a small contribution to the recognition power of our procedure. Therefore, a first option to improve the performance of the latter is to use observed, instead of predicted, structural properties derived from the three-dimensional structure of the protein affected by the mutations. In this case, in addition to observed secondary structure and accessibility, we also scored mutations utilizing statistical potentials,[32] which have been used for the prediction of mutant stability changes.[50]

TABLE II. Performance of Our Method When Trained and Tested in Individual Human Proteins (see Text)

	DAMUs[a]	NEMUs[b]	Qtot[c]	S[d]
P53[e]	172	104	96.7 ± 1.1	93.1 ± 2.2
ACTS	8	197	97.1 ± 0.0	47.2 ± 8.7
HBB	124	80	95.1 ± 1.0	89.8 ± 1.7

[a]Number of disease associated mutations.
[b]Number of neutral mutations.
[c]Overall success rate (see Methods).
[d]Normalized performance relative to random (see Methods).
[e]Proteins used: P53, p53; ACTS, Alpha-actin 1; HBB, Hemoglobin beta chain.

We find that the NN has a good overall performance rate, with 87.0 (±1.6) correct predictions. The performance relative to random is also good, 73.0 (±3.1). The false positive and false negative rates are 10.4 % (± 2.1) and 11.5% (±1.3), respectively. Except for the false negative rates, these figures represent a minor improvement relative to the sequence-based version of our procedure. This indicates that even when using observed structural properties, the first contributor to the method's success is evolutionary information. Accessibility and secondary structure improve their discrimination power, although it still is lower than that of evolutionary parameters (data not shown). Statistical potentials display a poor discrimination power, as can be seen by comparing Figure 1(A) and (C). Overall, these results suggest that further improvements to our method could come from a better treatment of the evolutionary information. At present our treatment of multiple sequence alignments is relatively crude, as it ignores the underlying structure of the data. A possible approach for a better use of evolutionary information would be to consider the underlying phylogenetic relationships, as has been described by Santibáñez-Koref et al.[11]

A second option would be to train our procedure for specific systems, concentrating in those proteins that may be of particular interest due to their relationship to diseases that have a high social cost and for which a vast amount of mutational data is available. This is the case for p53, a protein directly related to oncogenic processes,[51] for which more than 14,000 mutations are known.[52] In cases like this, predictive approaches derived for that specific system may provide better performance, as shown for p53 in recent studies.[11,16]

To test this idea with our method, we chose from our dataset those proteins for which the ratio between number of training mutations and parameters was between 5 and 10:[27] p53, alpha-actin 1, and hemoglobin beta chain. A simple NN with no hidden layer (perceptron) was then trained, and validated using a twofold cross-validation procedure. For two of the three proteins (p53 and hemoglobin) the cross-validated performance was very high (see Table II), clearly improving that of the general model [Table I(A)]. However, for alpha-actin 1 the performance was lower than that of the general model. This is probably due to the very low number of DAMUs, 8, relative to NEMUs, 197, resulting in poorly trained NNs. This was not the case for p53 and haemoglobin, for which the amounts of both types of mutations were more balanced, thus ensuring a better training of the network.

Clearly, our results support the idea that when interested in a given protein it is better to use a specific predictor, if enough mutation data are available and if the number of DAMUs and NEMUs is similar. This can be the case of proteins like p53, for which large mutations datasets are available and have been used to train specific prediction methods.[11,16] Otherwise, the general model is enough to obtain a reasonable prediction of the pathological character of SNPs.

CONCLUSIONS

We describe a simple method for the prediction of human disease-associated mutations, based on the use of sequence-based information and neural networks. The good success rate of the method, 83.5%, indicates that pathological mutations can be identified even in absence of structural information. This is particularly relevant when considering the increasing gap between known sequences and structures. In addition, use of the reliability index produced by the NN output provides a simple tool to assess the predictions, an option of particular interest when considering future applications of the method to large sets of nonsynonymous SNPs.

ACKNOWLEDGMENTS

The authors acknowledge the economic support provided by the Fundación Areces. They also acknowledge C. Saunders and D. Baker for kindly giving us their mutations sets. The work has been supported by the Spanish Ministry of Science and Technology (PM99-0046, GEN2001-4758-C07-07 and BIO2003-09327).

REFERENCES

1. Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Shaw N, Lane CR, Lim EP, Kalyanaraman N, et al. Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat Genet 1999;22:231–238.
2. Chakravarti A. To a future of genetic medicine. Nature 2001;409: 822–823.
3. Chasman D, Adams RM. Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. J Mol Biol 2001;307:683–706.
4. Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. Genome Res 2001;11:863–874.
5. Sunyaev S, Ramensky V, Koch I, Lathe W, 3rd, Kondrashov AS, Bork P. Prediction of deleterious human alleles. Hum Mol Genet 2001;10:591–597.
6. Wang Z, Moult J. SNPs, protein structure, and disease. Hum Mutat 2001;17:263–270.
7. Saunders CT, Baker D. Evaluation of structural and evolutionary contributions to deleterious mutation prediction. J Mol Biol 2002;322:891–901.
8. Ramensky V, Bork P, Sunyaev S. Human non-synonymous SNPs: server and survey. Nucleic Acids Res 2002;30:3894–3900.
9. Krishnan VG, Westhead DR. A comparative study of machine-learning methods to predict the effects of single nucleotide polymorphisms on protein function. Bioinformatics 2003;19:2199–2209.
10. Herrgard S, Cammer SA, Hoffman BT, Knutson S, Gallina M, Speir JA, Fetrow JS, Baxter SM. Prediction of deleterious functional effects of amino acid mutations using a library of structure-based function descriptors. Proteins 2003;53:806–816.
11. Santibáñez-Koref MF, Gangeswaran R, Santibíñez-Koref IP, Shanahan N, Hancock JM. A phylogenetic approach to assessing

the significance of missense mutations in disease genes. Hum Mutat 2003;22:51–58.
12. Sunyaev S, Ramensky V, Bork P. Towards a structural basis of human non-synonymous single nucleotide polymorphisms. Trends Genet 2000;16:198–200.
13. Ferrer-Costa C, Orozco M, de la Cruz X. Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties. J Mol Biol 2002;315:771–786.
14. Miller MP, Parker JD, Rissing SW, Kumar S. Quantifying the intragenic distribution of human disease mutations. Ann Hum Genet 2003;67(Pt 6):567–579.
15. Perutz MF. Protein structure: new approaches to disease and therapy. New York: W.H. Freeman and Co.; 1992.
16. Martin AC, Facchiano AM, Cuff AL, Hernandez-Boussard T, Olivier M, Hainaut P, Thornton JM. Integrating mutation data and structural analysis of the TP53 tumor-suppressor protein. Hum Mutat 2002;19:149–164.
17. Steward RE, MacArthur MW, Laskowski RA, Thornton JM. Molecular basis of inherited diseases: a structural perspective. Trends Genet 2003;19:505–513.
18. Guex N, Diemand A, Peitsch MC. Protein modelling for all. Trends Biochem Sci 1999;24:364–367.
19. Ng PC, Henikoff S. Accounting for human polymorphisms predicted to affect protein function. Genome Res 2002;12:436–446.
20. Bonneau R, Strauss CE, Rohl CA, Chivian D, Bradley P, Malmstrom L, Robertson T, Baker D. De novo prediction of three-dimensional structures for major protein families. J Mol Biol 2002;322:65–78.
21. Goodstadt L, Ponting CP. Sequence variation and disease in the wake of the draft human genome. Hum Mol Genet 2001;10:2209–2214.
22. McKusick VA. Mendelian inheritance in man: a catalog of human genes and genetic disorders. Baltimore: Johns Hopkins University Press; 1998.
23. Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res 2000;28:45–48.
24. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL. The Pfam protein families database. Nucleic Acids Res 2002;30:276–280.
25. Pace HC, Kercher MA, Lu P, Markiewicz P, Miller JH, Chang G, Lewis M. Lac repressor genetic map in real space. Trends Biochem Sci 1997;22:334–339.
26. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H,Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
27. Mehrotra K, Mohan CK, Ranka S, NetLibrary Inc. Elements of artificial neural networks. Cambridge, MA: MIT Press; 1997. xiv, 344 p.
28. Bishop CM. Neural networks for pattern recognition. Oxford, New York: Clarendon Press; Oxford University Press; 1995. xvii, 482 p.
29. Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy. J Mol Biol 1993;232:584–599.
30. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
31. Hubbard SJ, Thornton JM. 'NACCESS', computer program, Department of Biochemistry and Molecular Biology, University College London, 1993.
32. Sippl MJ. Recognition of errors in three-dimensional structures of proteins. Proteins 1993;17:355–362.
33. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 1992;89:10915–10919.
34. Dayhoff MO, Schwartz RM, Orcutt BC. A model of evolutionary change in proteins. In: Dayhoff MO, editor. Atlas of protein sequence and structure. Washington D. C.: National Biomedical Research Foundation; 1978. p 345–352.
35. Fauchere JL, Pliska V. Hydrophobic parameters of amino acid side chains from the partitioning of N-acetyl-amino acid amides. Eur J Med Chem 1983;18:369–375.
36. Miller S, Janin J, Lesk AM, Chothia C. Interior and surface of monomeric proteins. J Mol Biol 1987;196:641–656.
37. Chou PY, Fasman GD. Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. Biochemistry 1974;13:211–222.
38. Swindells MB, MacArthur MW, Thornton JM. Intrinsic phi, psi propensities of amino acids, derived from the coil regions of known structures. Nat Struct Biol 1995;2:596–603.
39. Bondi A. Van der Waals volumes and radii. J Phys Chem 1964;68:441–451.
40. Chothia C. Structural invariants in protein folding. Nature 1975; 254:304–308.
41. Yang W, Battineni ML, Brodsky B. Amino acid sequence environment modulates the disruption by osteogenesis imperfecta glycine substitutions in collagen-like peptides. Biochemistry 1997;36: 6930–6935.
42. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25: 3389–3402.
43. Shannon CE. A mathematical theory of communication. Bell System Tech J 1948;27:379–423.
44. Sippl MJ. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. J Mol Biol 1990;213: 859–883.
45. Rumelhart DE, McClelland JL. Parallel distributed processing: explorations in the microstructure of cognition. Cambridge, MA: MIT Press; 1986.
46. Shepherd AJ, Gorse D, Thornton JM. Prediction of the location and type of beta-turns in proteins using neural networks. Protein Sci 1999;8:1045–1055.
47. Sunyaev S, Hanke J, Brett D, Aydin A, Zastrow I, Lathe W, Bork P, Reich J. Individual variation in protein-coding sequences of human genome. Adv Protein Chem 2000;54:409–437.
48. Peitsch MC. About the use of protein models. Bioinformatics 2002;18:934–938.
49. Dipple KM, McCabe ER. Phenotypes of patients with "simple" Mendelian disorders are complex traits: thresholds, modifiers, and systems dynamics. Am J Hum Genet 2000;66:1729–1735.
50. Gilis D, Rooman M. PoPMuSiC, an algorithm for predicting protein mutant stability changes: application to prion proteins. Protein Eng 2000;13:849–856.
51. Vousden KH, Lu X. Live or let die: the cell's response to p53. Nat Rev Cancer 2002;2:594–604.
52. Hainaut P, Hernandez T, Robinson A, Rodriguez-Tome P, Flores T, Hollstein M, Harris CC, Montesano R. IARC Database of p53 gene mutations in human tumors and cell lines: updated compilation, revised formats and new visualisation tools. Nucleic Acids Res 1998;26:205–213.

IV. BIBLIOGRAFIA DEL CAPÍTOL

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25, 3389-402.

Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S. R., Griffiths-Jones, S., Howe, K. L., Marshall, M. & Sonnhammer, E. L. (2002). The Pfam protein families database. *Nucleic Acids Res* 30, 276-80.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Res* 28, 235-42.

Bishop, C. M. (1995). *Neural networks for pattern recognition*, Clarendon Press; Oxford University Press, Oxford, New York.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S. & Schneider, M. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* 31, 365-70.

Bondi, A. (1964). Van der Waals volumes and radii. *J. Phys. Chem* 68, 441-51.

Chothia, C. (1975). Structural invariants in protein folding. *Nature* 254, 304-8.

Chou, P. Y. & Fasman, G. D. (1974). Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. *Biochemistry* 13, 211-22.

Dayhoff, M. O., Schwartz, R. M. & Orcutt, B. C. (1978). A model of evolutionary change in proteins. In *Atlas of protein sequence and structure.* (Dayhoff, M. O., ed.), pp. 345-352. National Biomedical Research Foundation, Washington D. C.

Fauchere, J. L. & Pliska, V. (1983). Hydrophobic parameters of amino acid side chains from the partitioning of N-acetyl-amino acid amides. *European Journal of Medicinal Chemistry* 18, 369-75.

Ferrer-Costa, C., Orozco, M. & de la Cruz, X. (2002). Characterization of Disease-associated Single Amino Acid Polymorphisms in Terms of Sequence and Structure Properties. *J Mol Biol* 315, 771-86.

Goodstadt, L. & Ponting, C. P. (2001). Sequence variation and disease in the wake of the draft human genome. *Hum Mol Genet* 10, 2209-14.

Henikoff, S. & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* 89, 10915-9.

Henikoff, S. & Henikoff, J. G. (1993). Performance evaluation of amino acid substitution matrices. *Proteins* 17, 49-61.

Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22, 2577-637.

Krishnan, V. G. & Westhead, D. R. (2003). A comparative study of machine-learning methods to predict the effects of single nucleotide polymorphisms on protein function. *Bioinformatics* 19, 2199-209.

Martin, A. C., Facchiano, A. M., Cuff, A. L., Hernandez-Boussard, T., Olivier, M., Hainaut, P. & Thornton, J. M. (2002). Integrating mutation data and structural analysis of the TP53 tumor-suppressor protein. *Hum Mutat* 19, 149-64.

McKusick-Nathans Institute for Genetic Medicine, J. H. U. B., MD) and National Center for Biotechnology Information, National Library of Medicine, (Bethesda, MD). (2000). Online Mendelian Inheritance in Man, OMIM (TM).

Mehrotra, K., Mohan, C. K. & Ranka, S. (1997a). *Elements of artificial neural networks*. Complex adaptive systems, MIT Press, Cambridge, Mass.

Mehrotra, K., Mohan, C. K., Ranka, S. & NetLibrary Inc. (1997b). *Elements of artificial neural networks*. Complex adaptive systems., MIT Press, Cambridge, Mass.

Miller, S., Janin, J., Lesk, A. M. & Chothia, C. (1987). Interior and surface of monomeric proteins. *J Mol Biol* 196, 641-56.

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

Ng, P. C. & Henikoff, S. (2002). Accounting for human polymorphisms predicted to affect protein function. *Genome Res* 12, 436-46.

Pace, H. C., Kercher, M. A., Lu, P., Markiewicz, P., Miller, J. H., Chang, G. & Lewis, M. (1997). Lac repressor genetic map in real space. *Trends Biochem Sci* 22, 334-9.

Rost, B. & Sander, C. (1993). Prediction of protein secondary structure at better than 70% accuracy. *J Mol Biol* 232, 584-99.

Rost, B. & Sander, C. (1994). Conservation and prediction of solvent accessibility in protein families. *Proteins* 20, 216-26.

Rumelhart, D. E., McClelland, J. L. & University of California San Diego. PDP Research Group. (1986). *Parallel distributed processing : explorations in the microstructure of cognition*, MIT Press, Cambridge, Mass.

Saunders, C. T. & Baker, D. (2002). Evaluation of structural and evolutionary contributions to deleterious mutation prediction. *J Mol Biol* 322, 891-901.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Tech. J.* 27, 379-423.

Shepherd, A. J., Gorse, D. & Thornton, J. M. (1999). Prediction of the location and type of beta-turns in proteins using neural networks. *Protein Sci* 8, 1045-55.

Sippl, M. J. (1990). Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J Mol Biol* 213, 859-83.

Sippl, M. J. (1993). Recognition of errors in three-dimensional structures of proteins. *Proteins* 17, 355-62.

Sunyaev, S., Ramensky, V. & Bork, P. (2000). Towards a structural basis of human non-synonymous single nucleotide polymorphisms. *Trends Genet* 16, 198-200.

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.

Swindells, M. B., MacArthur, M. W. & Thornton, J. M. (1995). Intrinsic phi, psi propensities of amino acids, derived from the coil regions of known structures. *Nat Struct Biol* 2, 596-603.

Yang, W., Battineni, M. L. & Brodsky, B. (1997). Amino acid sequence environment modulates the disruption by osteogenesis imperfecta glycine substitutions in collagen-like peptides. *Biochemistry* 36, 6930-5.