Capítol 3.

Caracterització estructural i evolutiva de mutacions puntuals patològiques.

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

I. INTRODUCCIÓ

A principis de l'any 2001 amb l'aparició del primer esborrany complet del genoma humà (Lander et al., 2001; Venter et al., 2001) apareix un gran interès pels SNPs. Aquest interès s'explica en part pel fet que del genoma se'n reforça la idea que els SNPs explicarien gran part de la variabilitat intraespecífica (Chakravarti, 2001; Collins et al., 1998; Collins et al., 1997). Aquesta diversitat intraespecífica explicaria les diferències morfològiques entre individus però també explicarien diferències entre individus pel que fa a la sensibilitat a patir malalties ja siguin monogèniques o complexes (Collins et al., 1997).

Aquests fets van portar a diferents grups de recerca a qüestionar-se la pregunta de si existeixen propietats dels SNPs que permeten classificar-los com a neutres o associats a malaltia. Canviant el punt de vista, la qüestió es reformula de la següent manera, és possible l'anotació de SNPs pel que fa al seu caràcter neutre/patològic a partir de les seves propietats de seqüència o estructura?

Com a model de treball per mutacions patològiques vam escollir el de les mutacions puntuals causants de malalties monogèniques. Aquest model consisteix en mutacions que afecten a una proteïna de manera prou greu com perquè aquesta alteri la seva funció de manera patològica. Aquestes dades estan molt contrastades des del punt de vista bioquímic i clínic. Així mateix, la recerca intensiva en diferents camps, com la cristal·lografia de proteïnes, la biologia estructural i la biologia evolutiva ens dóna un marc conceptual molt consistent per treballar sobre SNPs que afecten a proteïnes. Tal i com es mostrarà en les pàgines següents vam usar la base de dades SwissProt (Boeckmann et al., 2003) com a font primària de mutacions. En alguns casos va caldre contrastar algunes dades de SwissProt per tal de confirmar el caràcter patològic d'algunes mutacions. El model usat per les mutacions neutres és un model evolutiu que prové de l'anàlisi d'alineaments múltiples de seqüència. Aquest model considera com a mutacions neutres aquelles diferències puntuals que hi ha entre proteïnes

molt properes evolutivament. El raonament que hi ha darrera aquest model és que si dues proteïnes que són prou properes evolutivament tenen diferències que han sigut acceptades per l'evolució aquestes han de ser neutres si assumim que el context funcional i cel·lular és el mateix.

Per cada mutació es van determinar un seguit de propietats relacionades amb l'estructura tridimensional de la proteïna i la posició de la mutació, amb les propietats físico-químiques de les proteïnes i dels aminoàcids. Comparant les distribucions dels paràmetres per les mutacions neutre i patològiques podem intentar trobar quines propietats diferencien les mutacions neutres i patològiques.

II. MATERIALS I MÈTODES

Polimorfismes d'un sol aminoàcid associats a malaltia (daSAP *Disease-associated Single Aminoacid Polimorphisms*)

Les mutacions puntuals usats en aquest treball van ser obtinguts buscant a la base de dades SwissProt (Bairoch & Apweiler, 2000) en la seva versió 39. Es va fer una cerca usant les paraules claus: PDB, VARIANT, HUMAN per tal d'obtenir proteïnes humanes amb estructura tridimensional coneguda i amb variants de seqüència coneguda. D'aquesta selecció se'n va eliminar els residus relacionats amb centres actius o amb cisteïnes involucrades amb ponts disulfur. La raó d'aquesta selecció és que l'impacte patològic sobre la proteïna d'aquests residus resulta evident. Per la resta de variants es va comprovar manualment la seva associació amb malaltia consultant les anotacions de SwissProt o bé consultant les referències originals. Finalment es van mapar les mutacions sobre l'estructura eliminant aquelles que no queien en zones resoltes. Al final 1169 mutacions van ser seleccionades que es distribuïen sobre 73 proteïnes diferents.

És important notar que no totes les mutacions seleccionades tenen el mateix impacte patològic. Així mentre algunes mutacions provoquen síndromes lleus altres estaran associats a malalties serioses. Desafortunadament, excepte per les mutacions letals és molt difícil classificar les daSAPs d'acord amb el seu impacte ja que dependria en molts casos d'observacions subjectives o imprecises fetes per diferents autors segons l'estat del pacient. Addicionalment, dividint el grup de dades de daSAPs en diferents grups segons el seu impacte en la salut resultarien grups més petits que per tant disminuirien el valor estadístic dels resultats.

Variants neutres (nSAPs *Neutral SAPs*)

Al mapar i caracteritzar les mutacions puntuals patològiques sobre les estructures tridimensionals sembla clar que cal un estat de referència per tal de confirmar les tendències observades i el comportament diferenciat de les patològiques sobre les neutres. Una primera aproximació és usar un model a l'atzar en el qual una mutació neutre pot ocórrer en qualsevol posició dins del grup de proteïnes que contenen mutacions patològiques. Aquest model és usat com a referència en algunes anàlisis. Tanmateix una aproximació més realista es defineix considerant les mutacions neutres observades pel mateix grup de proteïnes com a model de referència. Com s'ha comentat anteriorment, el model de mutació neutre que usem aquí es basa en considerar com a mutació neutre tota variabilitat puntual observada en alineaments múltiples de seqüència (MSA *Multiple Sequence Alignment*) seguint l'aproximació usada per Sunyaev et al. (Sunyaev et al., 2000). Per fer-ho hem pres com a SAPs neutres aquelles variants naturals que apareixen en altres espècies després d'eliminar les seqüències humanes dels MSA. Per tal de simular al màxim l'ocurrència de les variants neutres hem eliminat de l'alineament múltiple totes les seqüències que tenen una identitat de seqüència inferior al 95% amb la proteïna humana estudiada (Sunyaev et al., 2000).

Tot el procediment va ser aplicat a les 73 proteïnes esmentades anteriorment. Un total de 741 variants neutres constitueix el grup de mutacions neutres o nSAPs.

Les estructures de les proteïnes

Les estructures de les 73 proteïnes van ser obtingudes de la base de dades PDB(Berman et al., 2000). Vam observar que en molts casos hi havia diverses estructures disponibles, en aquests casos es va actuar segons els següents criteris:

- Raigs X és preferit a RMN, descartant-ne els teòrics.
- Estructures natives preferides sobre les mutants.
- Estructures sense complex preferides sobre les unides a complex.
- Estructures d'alta resolució preferides sobres les de baixa.

Part del nostre anàlisi era basat en l'ús d'estructures quaternàries de les proteïnes. Les coordenades de les estructures quaternàries van ser obtingudes de la base de dades PQS de l'Institut Europeu de Bioinformàtica (EBI-EMBL *European Bioinformatics Institute-European Molecular Biology Laboratory*)

Càlculs d'estructura secundària i accessibilitat

L'estructura secundària va ser calculada usant el programari SSTRUC escrit per David Keith Smith, que implementa la metodologia desenvolupada per Kabsch i Sander (Kabsch & Sander, 1983).

Els càlculs d'accessibilitat a solvent van ser fets amb el programari NACCESS (Hubbard & Thornton, 1993). L'accessibilitat relativa, és a dir l'accessibilitat del residu estudiat en la proteïna dividida per l'accessibilitat del residu en un tripèptid extès de seqüència Ala-X-Ala, va ser calculada per NACCESS i mapada a la forma de tres estats (Enterrada=0-9%; Semienterrada=9-36%; Exposada=36-100%) segons la classificació establerta per Rost i Sander (Rost & Sander, 1993).

Alineaments Múltiples de Seqüències.

En tots els casos els alineaments múltiples van ser obtinguts de la bases de dades Pfam. Aquesta base de dades conté alineaments de molta qualitat i amb moltes seqüències.

Les dades de PFAM es van tractar per tal d'eliminar redundàncies segons l'esquema Henikoff i Henikoff (Henikoff & Henikoff, 2000) de ponderació de les seqüències. Tot i que òbviament els resultats varien numèricament al aplicar aquesta metodologia de ponderació per eliminar redundància dels alineaments, no s'ha observat cap variació important en les tendències generals per cap propietat estudiada. A tall d'exemple, els percentatges d'accessibilitat relativa en tres estats per residus exposats, semienterrats i enterrats abans d'aplicar la metodologia, són 70.8%, 8.9% i 20.2%. Després d'aplicar la metodologia de Henikoff i Henikoff obtenim 66,3%, 10.6% i 23% pels mateixos estats d'accessibilitat.

Mesures de la variabilitat en les posicions dels alineaments múltiples.

La variabilitat o el grau de conservació d'un residu en una posició dins un alineament múltiple de seqüències es va mesurar usant l' entropia de Shanon (Shannon, 1948) de la distribució dels aminoàcids en la posició estudiada. Hem usat la formulació següent que ja ha estat aplicada en altres treballs previs (Atchley et al., 2000; Sander & Schneider, 1991)

$$-\Sigma p_i \ln_2 p_i \qquad \text{(eq.1)}$$

On el subíndex i fa referència als diferents tipus d'aminoàcids presents en la posició estudiada de l'alineament múltiple i la p_i correspon a la freqüència relativa de cada aminoàcid.

La informació redundant dels alineaments múltiples pot resultar en una estimació esbiaixada de les freqüències d'aminoàcids p_i en una posició donada (Altschul et al., 1997). La probabilitat es recalcula segons l'equació 2.

$$q_i = \Sigma_j w_{ij} \qquad \text{(eq.2)}$$

on el subíndex j corre sobre totes les seqüències en l'alineament que tenen el residu i en aquesta posició, wij són els pesos de Henikoff i Henikoff per les seqüències corresponents. L'entropia va ser calculada usant la nova equació:

$$-\Sigma_i(\Sigma_j w_{ij})\ln_2(\Sigma_j w_{ij}) \qquad \text{(eq.3)}$$

El subíndex i representa els diferents aminoàcids presents en la posició estudiada de l'alineament múltiples incloent també els *gaps* (Altschul et al., 1997).
Els valors de l'entropia variaran de 0 a 4.39, corresponent als graus màxims i mínims de conservació dels residus en aquesta posició.

Mesures relacionades amb les propietats de seqüència i de residu.

Dos tipus d'informació es van usar: matrius de mutació, propietats d'aminoàcids.
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 (Miller et al., 1987) et al. a partir

d'informació estructural. 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 grandària es van obtenir dels volums de van der Waals (Bondi, 1964) i dels volums mitjos dels residus enterrats (Chothia, 1975).

Tests Estadístics

Per tal de comparar les distribucions de mutacions neutres i patològiques per una variable determinada vam utilitzar els següents procediments.

Si la variable era contínua, per exemple l'accessibilitat relativa, vam usar el test de Kolmogorov i Smirnov (KS) (Press et al., 1992), l'estadístic del qual s'anomena D.

Si la variable és per naturalesa dividida en intervals com és el cas de l'accessibilitat en tres estats vam usar el test de la Chi-quadrada (CS) (Press et al., 1992) amb l'estadístic associat χ^2.

Procediment de Predicció usant matrius de mutació.

Volíem calcular el rendiment de les matrius de substitució de Blosum62 (Henikoff & Henikoff, 1992) i de PAM40 (Dayhoff, 1972) com a eines predictives del caràcter patològic de les mutacions patològiques. Amb aquesta fi vam seguir un procediment molt simple en el qual dos llindars van ser definits, un per cada matriu. Els valors de cada llindar variaven des del valor mínim al màxim dels elements de cada matriu. Aleshores per cada parell de llindars una mutació era predita:

Neutre: si els valors de Blosum62 i de PAM40, de la mutació eren més grans que el corresponent llindar per cada matriu.

Patològica: si els valors de les dues matrius per el canvi eren menors que el corresponent llindar per cada matriu.

No es feia predicció per les mutacions que no complien alguna de les dues premies anteriors.

El rendiment d'aquest mètode va ser calculat seguint una validació creuada de 5 etapes, en què els dos grups de dades van ser dividides en 5 grups de mutacions de manera aleatòria. Aleshores es construïa un grup d'entrenament amb 4 dels subgrups i es buscava el valor òptim dels llindars de Blosum62 i de PAM40 usant com a funció objectiva el coeficient de Matthews (Matthews, 1975). Aquest coeficient té en compte el número de positius i negatius certs així com el falsos positius i negatius. El rendiment d'aquests llindars va ser calculat amb el cinquè subgrup restant, anomenat el grup d'avaluació. Aquest procediment va ser repetit 5 vegades deixant consecutivament fora del grup d'entrenament un dels 5 subgrups generats. El rendiment del mètode correspon al final a la mitja dels 5 rendiments calculats per cada subgrup de test.

Vam usar dues mesures diferents del rendiment. El percentatge de les prediccions correctes que correspon al número de mutacions neutres i patològiques correctament predites dividida entre el número total de prediccions fetes. També vam usar el coeficient de Matthews (Matthews, 1975).

III. ARTICLE DE RECERCA

Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties. Ferrer-Costa, C, Orozco, M, de la Cruz, X. J Mol Biol **315**:(4) 771-86. 2002

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

doi:10.1006/jmbi.2001.5255 available online at http://www.idealibrary.com on IDEAL® *J. Mol. Biol.* (2002) **315**, 771–786

JMB

Characterization of Disease-associated Single Amino Acid Polymorphisms in Terms of Sequence and Structure Properties

Carles Ferrer-Costa[1], Modesto Orozco[1]* and Xavier de la Cruz[2]*

[1]*Departament de Bioquímica i Biologia Molecular, Facultat de Química, and*

[2]*Institute Eataló De Recerai Estudies, Avançats (ICREA) Universitat de Barcelona C/Martí i Franquès, 1 08028 Barcelona, Spain*

**Corresponding authors*

In the present work, we use structural information to characterize a set of disease-associated single amino acid polymorphisms exhaustively. The analysis of different properties, such as substitution matrix elements, secondary structure, accessibility, free energies of transfer from water to octanol, amino acid volume, etc., suggests that many disease-causing mutations are associated with extreme changes in the value of parameters relating to protein stability. Overall, our results indicate that, while knowledge of protein structure clearly helps in understanding these mutations, a finer understanding can come only from a quantitative knowledge of protein stability and of the protein environment in the cell. Interestingly, use of evolutionary information from multiple sequence alignments can be used to increase our knowledge of disease-associated mutations.

Keywords: disease; protein structure; amino acid polymorphisms; protein stability; protein function

Introduction

The knowledge of protein structure constitutes one of the main approaches towards understanding the molecular basis of human disease.[1–2] In particular, protein structure can be used to rationalize the effects of many disease-causing mutations.[1,3–6] One of the best-known examples is sickle cell anemia,[1] a disease characterized by a single Glu to Val mutation at position 6 in the β-globin chain. The mutation results in the formation of hemoglobin pairs through the binding of Val6 to a hydrophobic cavity present in the β-globin subunits. This favors the polymerization of hemoglobin molecules, which is associated with the premature death of the red cells and the resulting anemia. In the case of Friedreich's ataxia, a neurodegenerative disorder associated with lesions in the frataxin gene, the solution structure of frataxin[4] allowed the clustering of ataxia-causing mutations into two classes; mutations affecting protein stability, and mutations affecting protein-protein interactions.

Experimentally derived structures and homology models have been used for understanding the structural basis of pathological mutations. Zhang *et al.*[5] studied cancer-predisposing mutations in the C-terminal BRCT domain of BRCA1, finding that mutations could be associated either with fold destabilization or with the possible disruption of protein-protein interactions. Similar results were obtained by Busquets *et al.*[3] when mapping a set of mutations found in 34 patients with the glutaric aciduria type I disorder to a homology model of glutaryl-CoA dehydrogenase.

The interest in relating protein structure and disease has been fuelled by the completion of the first draft of the human genome[7,8] and the increasing amount of information on human genetic variation.[9] In this field, the identification of disease-associated single nucleotide polymorphisms (SNP) is one of the most important goals in biomedical research in the post-genome era.[9,10] Knowledge of protein structure is going to be one of the main tools to fulfil this objective.[10,11]

Interestingly, there is a two way relationship between protein structure and disease, since structure can help to rationalize disease, and information on disease-causing mutations can be used to increase our knowledge of protein structure and

Abbreviations used: SNP, single nucleotide polymorphism; SAP, single amino acid polymorphism; daSAP, disease-associated SAP; nSAP, neutral SAP; MSA, multiple sequence alignment.

E-mail addresses of the corresponding authors: modesto@luz.bq.ub.es; xavier@husky.bq.ub.es

of the structural basis of protein function. For example, in the case of Friedreich's ataxia, Musco *et al.*[4] observed that conserved exposed residues cluster on one side of the frataxin protein, thus suggesting that these residues may constitute a functional surface. This was supported by the fact that some clinically important mutations happened in residues at that surface. Also, in Factor VII the mutation Arg304Gln leads to undetectable plasma FVII activity, suggesting that this residue and its structural neighbors may be important for protein function.[12] In glutaric aciduria type I, some of the known mutations in the associated protein, glutaryl-CoA dehydrogenase, point to those residues involved in tetramer formation, the active form of the enzyme.[3] Finally, Jameson & Hollenberg[13] suggest that a series of pathology-linked mutations found in different hormones could be used to define the binding surface of the latter. In summary, disease-causing mutations can be used as structure/function probes in the same way as engineered mutants are used to relate protein stability and structural features in site-directed mutagenesis studies.[14,15]

In the present work, we study the relationship between protein structure and disease using a set of 1169 disease-associated single amino acid polymorphisms (daSAPs) corresponding to 73 proteins. We chose SAPs because a large number of them are available in the carefully annotated SwissProt database.[16] In addition, their impact on protein structure is easier to interpret than that coming from more complex sources of genetic variation. Here, daSAPs are characterized in terms of sequence and structure-based properties. Our results show that daSAPs display a specific behavior when studied by means of mutation matrices. It is found also that daSAPs happen with very low frequency at highly conserved positions in multiple sequence alignments. Furthermore, we find that some structural and physico-chemical properties may be used successfully to characterize the disease-character of many mutations. Overall, our results show that a fruitful understanding of protein disease can come only from combining knowledge of protein structure with that of protein stability and of the protein cellular environment.

Results

In order to study the relationship between protein structure and human disease, we used a set of 1169 daSAPs distributed over 73 proteins (see Materials and Methods) and a set of 741 nSAPs derived for the same set of proteins (see Materials and Methods). Both daSAPs and neutral SAPs (nSAPs) were mapped onto the corresponding protein structure and a set of different properties of the mutation site are studied. The distributions resulting for both types of SAPs were compared. We analyzed, among others, the following properties: secondary structure, size and accessibility in tertiary and quaternary structure. We characterized the mutation sites in terms of the conservation degree as derived from multiple sequence alignments (MSA), frequency of the observed mutation in the MSA, and the characteristics of the local sequence environment.

Mutation matrices and daSAPs

For each daSAP we found the corresponding substitution matrix element in either a Blosum[17] or a PAM[18] substitution matrix. The results were plotted as a frequency histogram (Figure 1). The same procedure was repeated for the nSAPs, which constitute our model for neutral variability (see Materials and Methods).

The Blosum62 substitution matrix distributions for the daSAPs and the nSAPs (Figure 1(a)) are significantly different at the 1 % level ($\chi^2 = 267.4$, nine degrees of freedom). Both distributions overlap partially (Figure 1(a)) at values around zero. However, clear differences are seen at the tails of the distributions: for values less than -1, we find 43.7 % of daSAPS, and only 10.8 % of nSAPs. For values greater than 0, we find 21.5 % of daSAPs and 51.6 % of nSAPs. The PAM40 distribution is more complex (Figure 1(b)), but the same general trends of the Blosum62 distribution are found. Thus, for values of PAM40 greater than 0 there is a larger population of nSAPs, while for negative values, daSAPs are much more common. These results suggest that extreme values of the mutation matrices may be used to discriminate between neutral and disease-associated SAPs.

Secondary structure, accessibility, size and daSAPs

The frequencies with which daSAPs are found in different secondary structure elements are shown in Table 1A. We observe that they are found more frequently in coil, 42.6 %, than in helices, 36.7 %, or β-strands, 20.7 %, in accordance with the secondary structure composition of our protein sample ($\chi^2 = 0.327$, three degrees of freedom). Interestingly, the distribution for daSAPs and nSAPs are significantly different at the 1 % level ($\chi^2 = 36.6$, three degrees of freedom). Thus, the percentage of daSAPs in β-strands is higher, 9.1 percentile points, than that for nSAPs. On the contrary, daSAPs are less frequent than nSAPs in helices and coil, 11.3 and 2.2 percentile points, respectively. These results indicate that daSAPs may happen at almost any location, independently of its secondary structure, while this is not the case for nSAPs. Therefore, secondary structure may provide some degree of discrimination between daSAPs and nSAPs.

We next explored how daSAPs distribute relative to the solvent-accessibility of the mutated residue. We first used a standard three-state accessibility scheme utilized in accessibility prediction programs, in which residue relative accessibilities

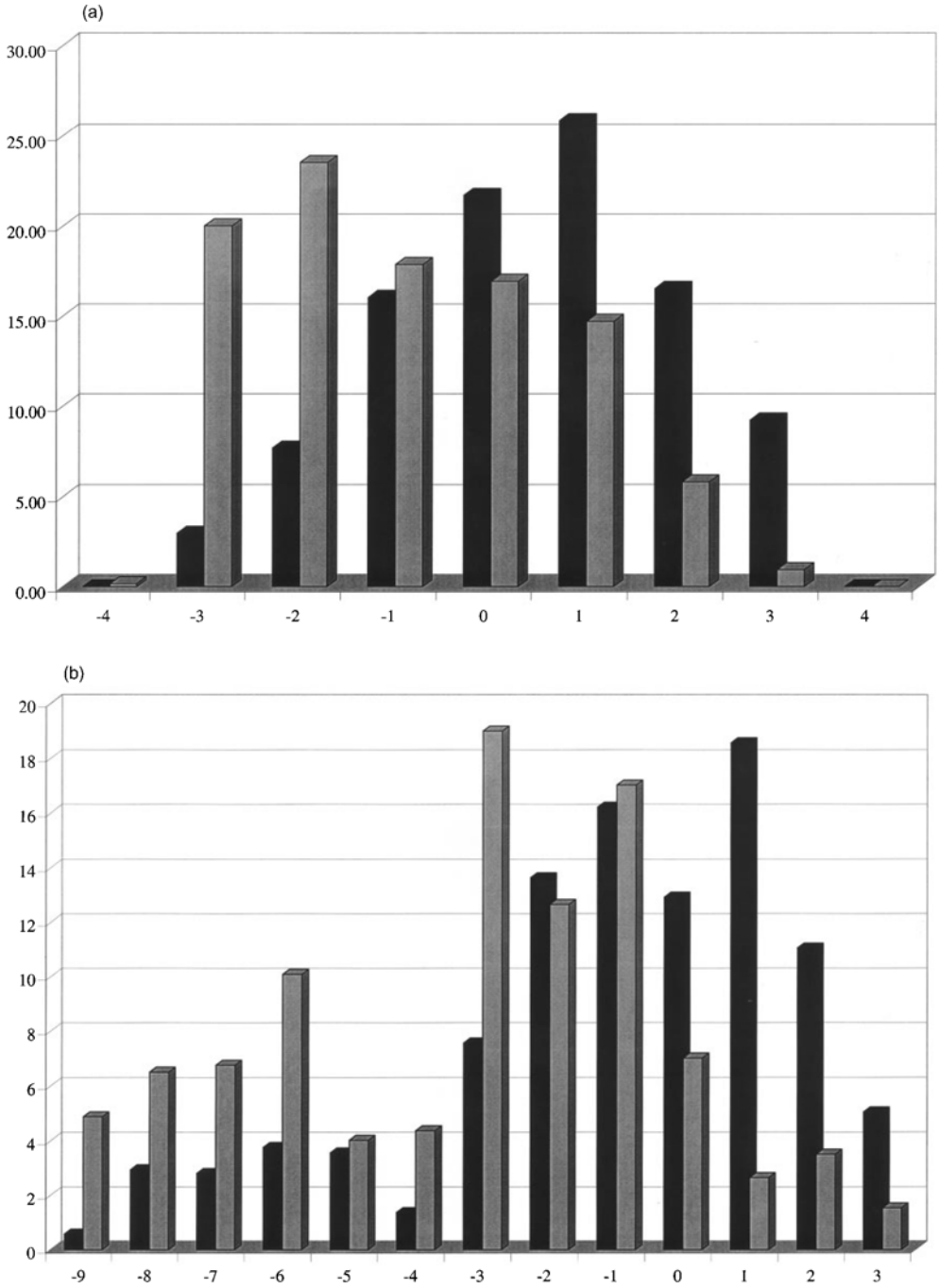

Figure 1. Distribution of daSAPs and nSAPs relative to the substitution matrices values. The (a) Blosum62 and (b) PAM40 matrices were used. daSAPS are shown in gray, nSAPs in black.

are clustered into three states (see Materials and Methods): buried, half-buried and exposed. The results given in Table 1B show that daSAPs tend to happen more frequently at buried (39.1 %) and half-buried sites (31.6 %) than at exposed sites (29.3 %). This distribution is significantly different at the 1 % level ($\chi^2 = 48.6$, three degrees of freedom) from the three-state accessibility composition

Table 1. Secondary structure and accessibility distributions (%) for daSAPs and nSAPs

	daSAPs	nSAPs
A. *Secondary structure*		
Helix	36.7	48.0
Beta	20.7	11.6
Coil	42.6	40.4
B. *Accessibility*		
Exposed	29.3	66.3
Buried	39.1	10.6
Half-buried	31.6	23.0
C. *Joint secondary structure and accessibility*		
Exp-Helix	9.6	30.6
Exp-Beta	2.1	5.4
Exp-Coil	17.7	30.3
Bur-Helix	15.3	6.6
Bur-Beta	12.7	1.3
Bur-Coil	11.1	2.7
Hlf-Helix	11.8	10.8
Hlf-Beta	6.0	4.9
Hlf-Coil	13.8	7.4

Exp, exposed; Bur, buried; Hlf, half-buried.

of our protein sample (i.e. in terms of accessibility, the daSAPs distribution is different from that corresponding to a random mutation model). More interesting, nSAPs and daSAPs distributions also are significantly different at the 1% level ($\chi^2 = 267.4$, three degrees of freedom). In particular, daSAPs are found more frequently at buried and half-buried sites: 28.5 and 8.6 percentile points, respectively. For exposed sites, nSAPs are about 37 percentile points more frequent than daSAPs.

When relative accessibilities were grouped into 5% bins (results not shown), we observed that daSAPs and nSAPs were different at the 1% level ($D = 0.430$). In particular, maximum differences (around 25 percentile points) between both distributions were found for highly buried residues (those with relative accessibility lower than 5%).

Overall, these results show that accessibility of the mutation site has a clear discrimination power between nSAPs and daSAPs, confirming that structure destabilization may be at the origin of many pathological mutations. In fact, combination of Blossum62 and accessibility analysis shows that around 78% of all daSAPs are located in buried regions (where even a small change can lead to important structural modifications), or imply changes in Blosum62 index equal to or less than -1 (i.e. non-conservative mutations). Remarkably, only 36% of the nSAPs fulfil these requirements, giving strong support to the theory that changes in the structure are related to the pathological properties of mutations.

Because there may be a certain degree of correlation between secondary structure and accessibility, we studied how daSAPs are distributed regarding the nine possible combinations of secondary structure and accessibility states. Our results (Table 1C) indicate that the distributions for daSAPs and nSAPs are significantly different at the 1% level ($\chi^2 = 318.7$, three degrees of freedom). The main difference between daSAPs and nSAPs happens for the exposed-helical state, for which nSAPs are 21 percentile points more frequent than daSAPs. This explains a large amount of the 37 percentile points difference between the number of daSAPs and nSAPs at exposed sites. The remaining can be attributed to differences in the exposed-coil sites.

Free energies of water to octanol transfer, size, secondary structure propensities and daSAPs

To further characterize pathological and neutral mutations, we studied the changes associated with them in the following physico-chemical properties at the residue level: free energies of transfer from water to octanol[19] ($\Delta\Delta G_{tf}$), residue size[20] (ΔVol) and secondary structure propensities[21] (ΔSS).

In Figure 2 we show the histograms for $\Delta\Delta G_{tf}$ associated with daSAPs and nSAPs at exposed (Figure 2(a)) and buried (Figure 2(b)) positions. In both cases, the nSAPs and daSAPs are found to be significantly different at the 1% level ($\chi^2 = 137.991$, 18 degrees of freedom, and $\chi^2 = 68.225$, 18 degrees of freedom, for the exposed and buried cases, respectively). Neutral mutations tend to have smaller $\Delta\Delta G_{tf}$ absolute values than pathological mutations, at both exposed and buried sites. In particular, for exposed sites (Figure 2(a)) the daSAPs distribution is shifted towards positive values when compared with the nSAPs distribution. This indicates that mutating exposed residues may be associated with disease when the mutation implies large increases in the hydrophobic character at the mutation site.

For buried sites (Figure 2(b)), the tails of the daSAPs distribution are more populated than the tails of the nSAPs distribution, in particular for those $\Delta\Delta G_{tf}$ values indicating an increase in the hydrophilic character of the mutated residue. This indicates that part of the pathogenic effect caused by mutations at buried sites can be explained by an increase in the hydrophilic character of the mutated residue.

The distribution of amino acid volume changes for accessible residues (Figure 3(a)) are significantly different for nSAPs and daSAPs at the 1% level ($\chi^2 = 109.332$, 16 degrees of freedom). The distribution for daSAPs is more populated at both tails than that for nSAPs. This indicates that extreme volume changes at accessible sites are slightly more frequent for daSAPs than for nSAPs. In the case of buried sites (Figure 3(b)), both distributions are significantly different at the 1% level ($\chi^2 = 39.774$, 15 degrees of freedom). The daSAPs distribution spreads over a broader range of values than the nSAPs distribution, again showing a trend of daSAPs to be associated with larger volume changes than nSAPs. Overall, these results suggests that large volume changes may be related to pathological effects when happening at either exposed or

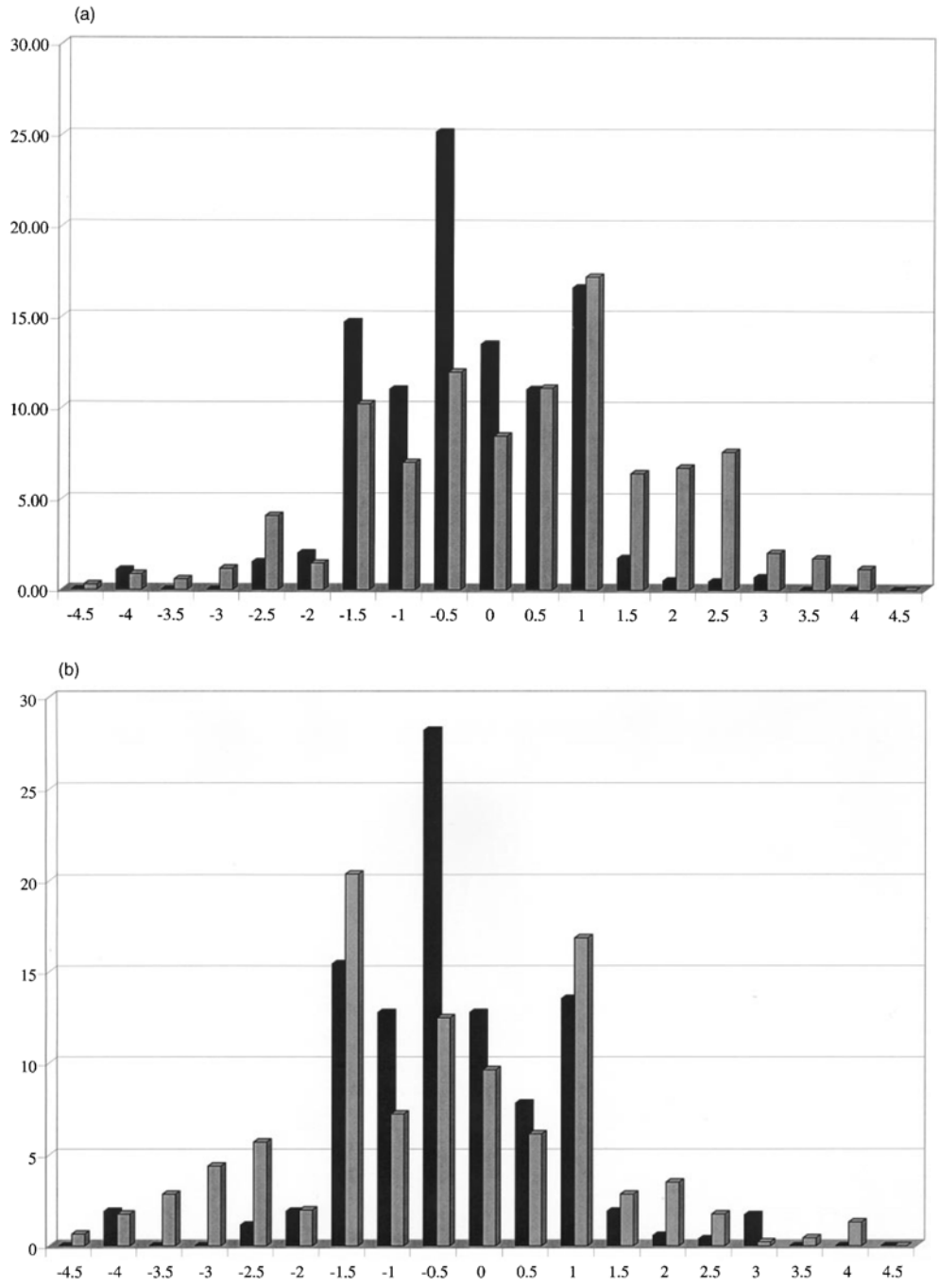

Figure 2. Distribution of daSAPs and nSAPs relative to the change in free energy of water to octanol transfer, $\Delta\Delta G_{tf}$ (kcal/mol), between native and mutated residues. (a) Accessible mutation sites; (b) buried mutation sites. daSAPS are shown in gray, nSAPs in black. Positive values of $\Delta\Delta G_{tf}$ indicate an increase in the hydrophobic character of the mutated residue.

buried residues. The latter may help to explain why mutations happening at buried residues are so pathogenic.

Finally, changes in secondary structure propensities associated with mutations were analyzed (Figure 4(a) and (b)). Again, the distributions for

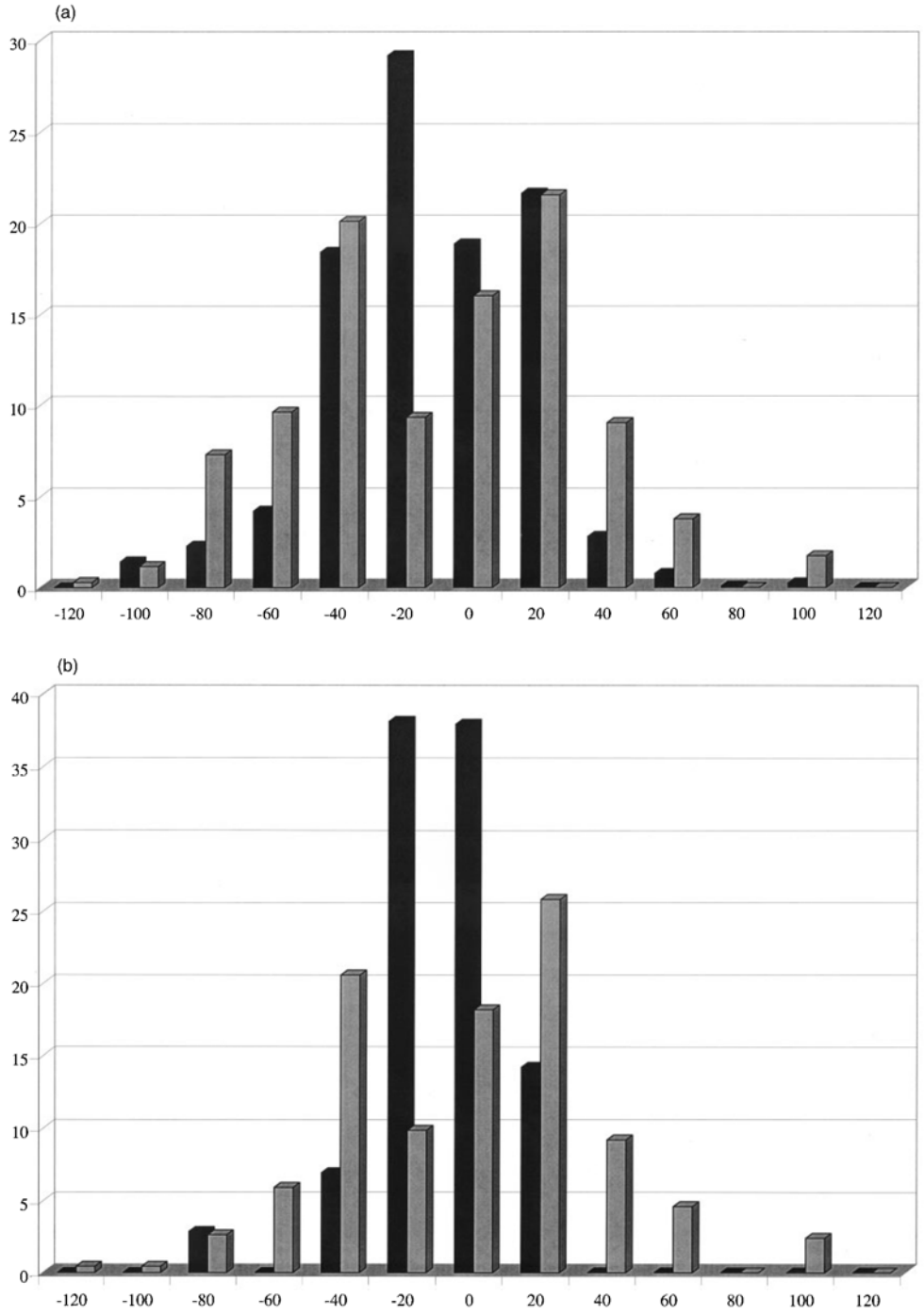

Figure 3. Distribution of daSAPs and nSAPs relative to size changes, ΔVol (Å^3). (a) Accessible mutation sites; (b) buried mutation sites. daSAPS are shown in gray, nSAPs in black. Positive values of ΔVol indicate an increase in the volume of the mutated residue.

nSAPs and daSAPs are significantly different at the 1% level for both the mutations happening in helices ($\chi^2 = 28.739$, 11 degrees of freedom) and those happening in β strands ($\chi^2 = 30.978$, 11 degrees of freedom). However, a detailed analysis of the data does not show any clear systematic

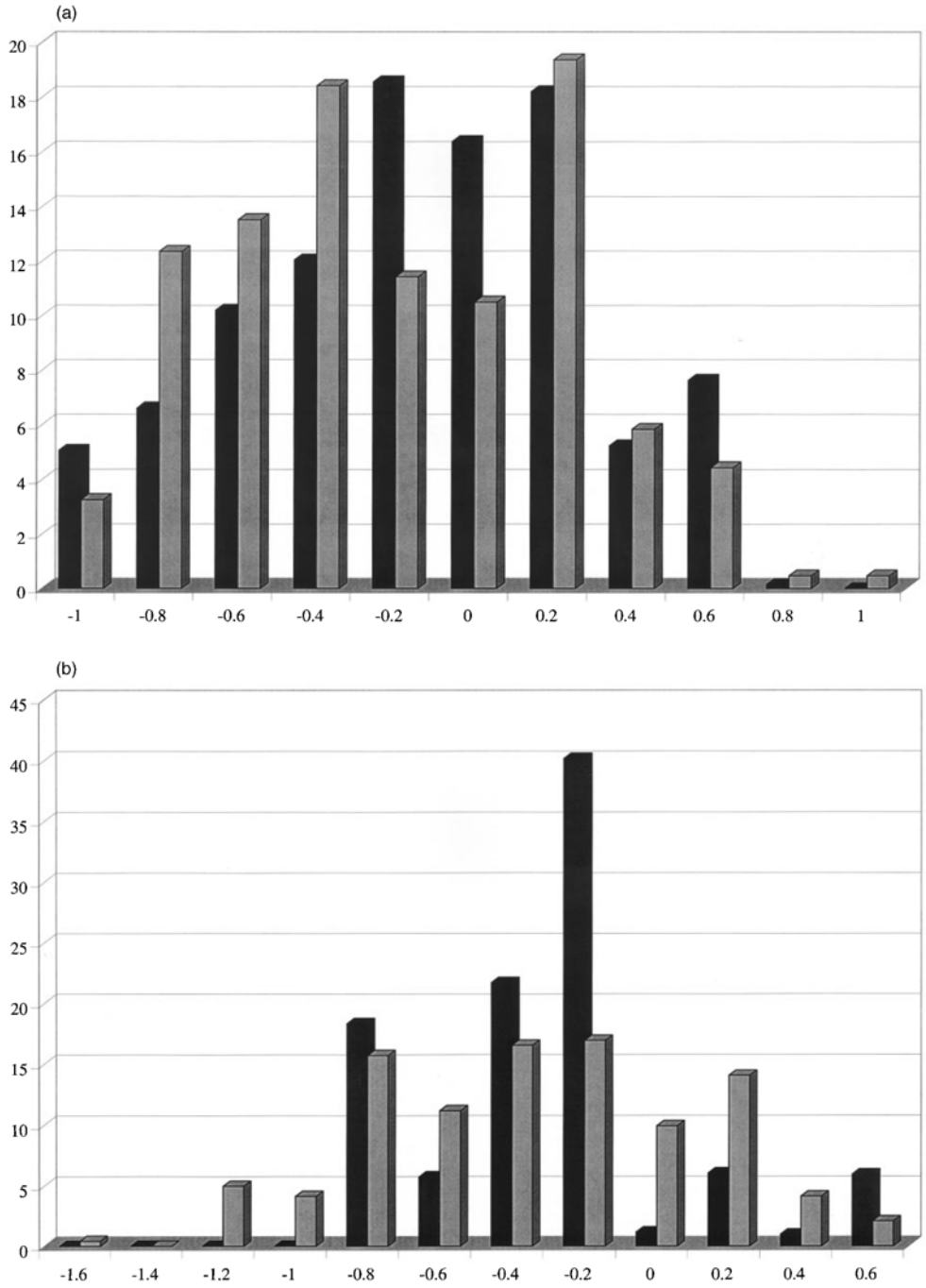

Figure 4. Distribution of daSAPs and nSAPs relative to changes in secondary structure propensities, ΔSS, measured with the Chou & Fassman[21] scale, between native and mutated residues. (a) Mutation sites in helices; (b) mutation sites in strands. daSAPS are shown in gray, nSAPS in black. Negative values of ΔSS indicate a destabilizing effect of the secondary structure element at the mutation site.

difference in secondary structure propensity between nSAPs and daSAPs. Thus, mutations suggesting disruption of secondary structure are present in nSAPs, and a number of daSAPs associated with secondary structure-stabilizing values of ΔSS are found. Overall, these results suggest that

in many cases other disrupting effects, e.g. unfavorable volume changes, will prevail over the secondary structure-stabilizing/destabilizing contribution in defining the pathological character of a mutation.

Quaternary structure formation and daSAPs

The daSAPs located at accessible sites were analyzed to see if some of them could interfere with the formation of the quaternary structure of the protein. Analysis of residue accessibility at the mutation site in both tertiary and quaternary structures (from the PQS database[22]) shows clearly (Figure 5) that several mutations occuring at exposed sites in the tertiary structure are in fact located at the interface between monomers in the quaternary structure. This indicates that some daSAPs may lead to disruption or destabilization of the protein quaternary structure.

MSA variability and daSAP

Finally, we studied the relationship between daSAPs and the degree of residue conservation at the mutation site, as derived from the MSA of the corresponding protein. To this end, we computed the Shannon's entropy[23] (see Materials and Methods) at each position of the MSA for every protein in our dataset and compared the resulting distribution with that derived considering only those positions for which a daSAP was found.

The all-positions entropy distribution and that corresponding to the daSAP positions are significantly different at the 1% level ($D = 0.152$). The all-positions entropy distribution shows a sharp peak for very low entropy values, which corresponds to highly conserved regions, and a nearly flat distribution, up to entropy values below 3.0 bits. For the daSAPs distribution the peak for very low entropy values has disappeared, while the population of entropy values between 0.6 and 1.6 bits, and that of high entropy values (entropy values between 3.0 and 3.8 bits) has increased. It is then clear that daSAPs are rarely found at highly conserved regions. This indicates that mutations at these positions will generally result in lethal phenotypes, while mild diseases will, in general, correspond to changes in more variable regions.

Discussion

Protein structure provides a solid framework for the integration of the information available on a given disease, thus allowing a better interpretation of known disease-causing mutations.[1,2] Hence, knowledge of protein structure provides a more rational approach to the fight against disease.[24]

Here, we report a systematic study of the relationship between structure and diseases caused by SAPs. The final goal is to gain insight into the subtle reasons that make a mutation pathological, when it is not simply related to clear disruptions of the covalent structure (i.e. loss of disulphide

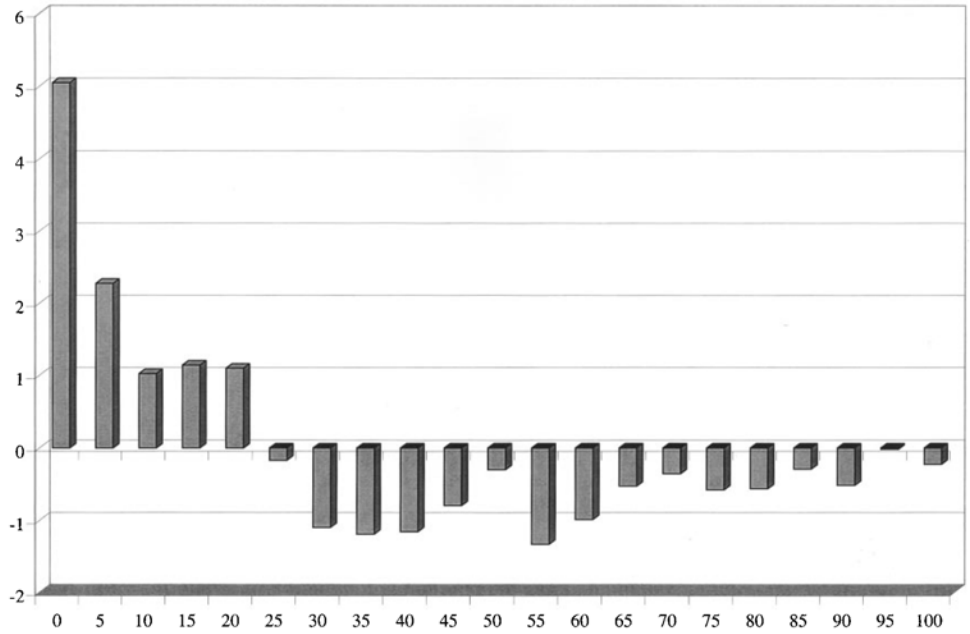

Figure 5. Histogram representing residue accessibility (Å^2) differences at the daSAP sites when going from the tertiary to the quaternary structure. The difference plotted is, for a given relative accessibility bin, frequency in quaternary structure minus frequency in tertiary structure.

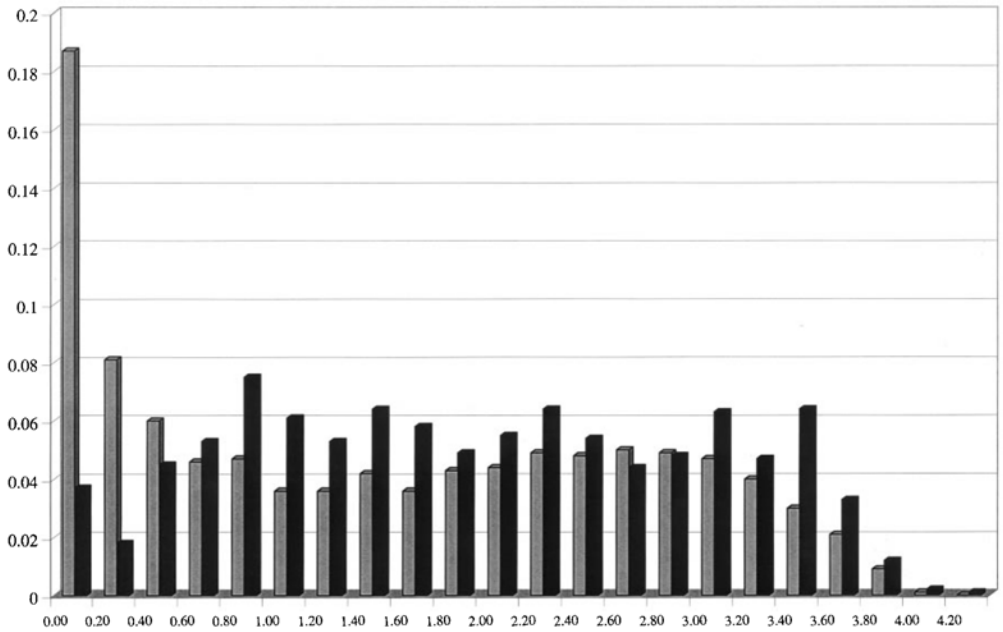

Figure 6. Distribution of daSAPs relative to degree of conservation in the multiple sequence alignment for the corresponding protein family. Conservation was measured using the Shannon[23] entropy: low and high entropy values correspond to low and high sequence variability at the mutation site, respectively. daSAPs are represented in black; all protein residues are represented in gray.

bridges), or to changes in the active-site residues. In our analysis, we have concentrated on the properties that can be derived directly from protein sequence, e.g. volume changes, etc., or can be predicted from it, e.g. secondary structure, accessibility. This is relevant when considering the possibility of using these properties to predict the pathological character of a mutation from only the knowledge of protein sequence.

daSAPs and protein stability

A simple and natural way of characterizing single amino acid mutations is the use of substitution matrices (e.g. Blosum[17] or PAM[18]). These matrices are at the core of the most-utilized sequence comparison programs,[25,26] and are used to detect both close and distant relationships between protein sequences.[27] Because sequence changes along the evolutionary pathway must be consistent with the underlying protein structure, the elements in the mutation matrices must reflect how the corresponding changes in the amino acid physico-chemical properties can affect the stability of protein structure. Indeed, substitution matrices have been used to predict protein stability changes arising from single amino acid mutations.[28] In our case, analysis of nSAPs and daSAPs in terms of substitution matrices showed that nSAPs (Figure 1(a)) are associated more frequently with positive values, and daSAPs with negative values, of the substitution matrix elements. Because the latter can be related to amino acid physico-chemical properties relevant to protein stability, this immediately suggests that protein destabilization could be a major cause of disease, in agreement with Wang & Moult.[29]

Inspection of Blosum62 and PAM40 results (Figure 1) indicates that extreme values of these matrix elements could be used to identify some of the daSAPs and nSAPs. A simple implementation of this idea was done by defining a single threshold for each matrix. Mutations with Blosum62 and PAM40 values below both thresholds were predicted to be daSAPs. Those with values above both thresholds were predicted as nSAPs. For the remaining mutations, no prediction was made. A fivefold cross-validation procedure was used to find the values of the thresholds (see Materials and Methods). We found that the best performance of this simple approach (84 % correct predictions; Matthews coefficient: 0.38) is found for values of −2 and 1, for the Blosum62 and PAM40 thresholds, respectively, with predictions made for a subset of 51 % of the total number of SAPs. This shows that mutation matrices can be utilized to predict the pathological character of SAPs, in contrast with suggestions by Ng & Henikoff.[30] In fact,

our simple model shows a predictive ability similar or even superior to that of other more complex models.[30]

As mentioned before, the results obtained for the mutation matrices indicate that some daSAPs involve changes in physico-chemical properties at the mutation site, which may result in protein destabilization. In order to dissect which could be the main contributing factors, we studied how daSAPs relate to the following properties: location in protein structure both in terms of accessibility and secondary structure, as well as with changes in hydrophobic character, amino acid size and secondary structure propensity. Many of the results found can be rationalized in terms of our previous knowledge from stability studies done in different sets of protein mutants (see below). However, an important difference between our work and conventional protein stability studies is that we provide a quantitative relationship between different parameters, either sequence or structure-based, and the pathological character of a mutation.

For instance, we find that around 32% of the daSAPs are located at highly buried sites (relative accessibility <5%), while only around 7% of nSAPs are found in these regions. These results are in agreement with previous estimates by Sunyaev *et al.*[11] (35%) obtained using a different data set. While these results are consistent with our previous knowledge on mutational analysis of protein stability,[14,31,32] they go beyond the latter, since they establish a quantitative relationship between degree of burial and pathological character of a mutation.

Interestingly, the rare neutral mutations occurring in buried regions of the protein are very conservative, as noted by Blossum62 analysis. Thus, only 34% of the nSAPs show changes in Blosum62 index equal to or less than −1, while 64% of daSAPs occurring in buried regions fulfil this requirement. This finding, together with the fact that 78% of the daSAPS occur in buried regions or imply changes in Blosum62 index equal to or less than −1 (only 36% of the nSAPs fulfil these requirements) strongly support the connection between protein destabilization and disease.

Secondary structure location also seems to be associated with the pathological character of mutations, as daSAPs tend to be more frequent at β-strands than nSAPs, again in accordance with results reported by Sunyaev *et al.*,[11] who suggest that this could be related to protein function loss rather than to protein destabilization.[11]

The location and the nature of the mutation are important determinants of its pathological character. For example, mutations at the protein core involving a change in the hydrophobic character of the buried residue, e.g. Leu to Asp, are likely to be worse than non-disruptive mutations, e.g. Val to Ala. Moreover, mutations at accessible locations may result in different degrees of protein destabilization.[33,34] In summary, location and changes in physico-chemical properties may help to explain the pathological effect of some daSAPs, through changes in protein stability.

In order to gain additional insight into the possible structure-destabilizing effect of disease-causing mutations, we explored the changes in hydrophobic character for the different daSAPs, as measured using differences in free energies of transfer from water to octanol.[19] Because the impact of a given mutation will depend on the accessibility at the mutation site, daSAPs and nSAPs were partitioned into buried and exposed mutations.

For mutations at exposed locations we find (Figure 2(a)) that $\Delta\Delta G_{tf}$ values for nSAPs tend to cluster around zero, showing the tendency of neutral mutations to preserve the hydrophobic/hydrophilic character of the mutated residue. We also find that, despite an overlap between the nSAP and daSAP distributions, there is a clear difference between nSAPs and daSAPs for $\Delta\Delta G_{tf}$ values greater than 1.5 kcal/mol (1 cal = 4.184 J). That is, mutations involving large increases in residue hydrophobicities in exposed residues are likely to produce pathological effects. This can be related to three different effects: (i) protein destabilization;[14,33] (ii) non-specific aggregation of the protein; and/or (iii) disruption of biologically relevant intermolecular interactions.

The $\Delta\Delta G_{tf}$ curves for buried locations show some interesting features (Figure 2(b)): $\Delta\Delta G_{tf}$ values for nSAPs tightly cluster around zero, while those for daSAPs spread over a broader range. In particular, values of $\Delta\Delta G_{tf}$ less than −2 kcal/mol correspond mainly to daSAPs. This behavior can be rationalized when considering some basic results from mutational analysis of protein stability. First, it is known that mutations at the protein core involving decreases in hydrophobicity lead to destabilization of the structure either through a decrease in hydrophobic interactions or through unfavorable electrostatic effects.[14,31] Second, residue changes involving increases in hydrophobicity may also destabilize protein structure, since they can disrupt native interactions and introduce steric hindrance.[35] Following these results, the fact that daSAPs lead in many cases to dramatic alterations of hydrophobicity in the core region (both increases and decreases) can be understood as a reinforcement of the connection between protein structure destabilization and pathogenicity of the mutation. Again, an important feature of our analysis is the fact that it provides a quantitative relationship between extreme values of $\Delta\Delta G_{tf}$ and accessibility and the pathological character of the mutations.

To investigate the connection between daSAPs and volume changes, and following the previous protocol, we analyzed volume changes in nSAPs and daSAPs for both accessible and buried residues. Analysis of the results for accessible sites (Figure 3(a)) shows that daSAPs are more frequent at both tails of the distribution than nSAPs; i.e. mutations implying large volume changes (less

than -60 Å^3 or greater than 60 Å^3) tend to be more aggressive, normally leading to disease. In terms of protein stability, this is somewhat surprising, since volume changes at accessible positions are unlikely to destabilize the protein by themselves. The most probable explanation is that either these changes are accompanied by changes in the hydrophobic character of the daSAP, or they affect the formation of intermolecular interactions.

For buried sites (Figure 3(b)), the central part of the histogram, i.e. that corresponding to small volume changes, is populated mainly by the nSAP distribution. This agrees with the fact that large size changes would require compensating mutations to avoid dramatic disruptions of the protein core.[36,37] On the contrary, daSAPs show a nearly bimodal distribution: one maximum corresponds to large-to-small volume changes, the other to small-to-large volume changes. This is consistent with what we know about mutations in the protein core: both large-to-small[38] as well as small-to-large[35] mutations may have destabilizing effects on protein structure which, in turn, can trigger disease. In particular, our study indicates that changes of less than -20 Å^3 or greater than 60 Å^3 in the mutated residue are likely to be associated with disease.

A third aspect we analyzed was the secondary structure at the mutation sites (Table 1A). We found no dramatic dependence between secondary structure and the pathogenic character of the mutation. However, differences between nSAPs and daSAPs suggest that, while mutations in β-strands are less frequent than those in helices or coil regions, they are more likely to produce pathologies. To further analyze the possible relationship between daSAPs and secondary structure we studied the changes in secondary structure propensities induced by the mutations. In general (Figure 4(a) and (b)), no clear differences between daSAPs and nSAPs are found, except for the case of large decreases in β-strand propensity (<-0.6), which seem to be associated preferentially with daSAPs (Figure 4(b)). This suggests that, except for those few cases for which the secondary structure element is dramatically disrupted (e.g. for Ala to Pro mutations in helices) other factors will prevail over secondary structure at the mutation site in determining the pathological character of a mutation. For example, the accessibility state of the mutated residue. This would explain the fact that daSAPs are found at large positive values of β-strand propensities.

Overall, we can see that both sequence and structure-based properties can be used to provide a quantitative understanding of the pathological character of SAPs. In this sense, combination of both types of properties can be fruitfully used to examine the differences between nSAPs and daSAPs. It is worth noting both the work by Sunyaev *et al.*[39] and Wang & Moult,[29] who utilize different sets of structure-based rules to predict the pathological character of SNPs, together with additional information from SwissProt records.[16] Some of the properties they use are similar to those analyzed in this work (e.g. accessibility, secondary structure location). Although no breakdown of their recognition power is provided, their overall results suggest that these properties have a predictive power of the pathological character of SAPs. This would be in accordance with the results of our study. Interestingly, neither Sunyaev *et al.*[39] nor Wang & Moult[29] use mutation matrices, or other sequence-based properties that may be related directly to the pathological character of daSAPs, as shown here. At this level, it is also worth noting that the structural properties utilized in this study (accessibility and secondary structure) can be predicted from sequence, opening the possibility of using them in the prediction of daSAPs from only knowledge of protein sequence.

daSAPs and intermolecular interactions

As noted in the previous sections, in many cases daSAPs can be associated with some structural features and physico-chemical properties related to tertiary structure stability. However, it has been suggested that some diseases may be caused by disruption of intermolecular interactions, like protein-protein or protein-DNA interactions.[2] The daSAPs present at solvent-accessible sites can lead to disease through, at least, three different mechanisms: destabilization of the protein monomer, destabilization of protein interactions and aggregation. Unfortunately, the nature of the mutation cannot be used to distinguish between these cases, e.g. changes from hydrophilic to hydrophobic may either destabilize the protein monomer,[35] disrupt electrostatically driven protein-protein interactions[40] or lead to aggregation.[1,41]

A first step to address this problem is to study accessibility changes at the mutation site, when going from tertiary to quaternary structure. It is known that soluble proteins are usually found in oligomeric states,[42] thus suggesting that some daSAPs may lead to disease through disruption of the protein quaternary structure. Our results (Figure 5) confirm that this may be the case for about 10% of daSAPs, which are exposed in the tertiary structure, but are found at the interface between protein subunits when considering the quaternary structure. Assuming that 80% of daSAPs cause disease through protein destabilization,[29] and 10% may cause it through quaternary structure destabilization, the remaining 10% of daSAPs may be explained by other mechanisms involving the cellular environment, like aggregation, etc. These cases, which probably correspond to surface mutations, may be very difficult to identify. In a simple approach to this problem, the degree of conservation of the mutated residue in the protein family and accessibility at the mutation site can be used to provide indirect evidence on whether daSAPs may alter intermolecular inter-

actions.[4,43] Future research will be focused on this issue.

Residue conservation and daSAPs

Finally, to explore the relationship between residue conservation and daSAPs, we studied the degree of residue conservation in MSAs at the daSAP positions. It is found (Figure 6) that the peak corresponding to highly conserved residues (entropy values close to zero) is almost completely missing from in the daSAP distribution. This is because mutations at these positions will probably "knock-out" protein function or dramatically disrupt protein structure, thus resulting in very severe disorders leading to lethal phenotypes. Interestingly, a noticeable subset of daSAPs are found in regions of large variability according to Pfam MSA. It is clear that mutations in these regions might decrease protein stability, but two other mechanisms can be noted: (i) disruption of protein interactions specific in human beings; (ii) protein aggregation due to environmental conditions specific for human cells.

An interesting issue when considering disease-causing mutations is the fact that they may appear in other species without leading to disease.[2] For instance, Zuckerkandl & Pauling[2] mention the case of the Norfolk mutation present in healthy orangutans but leading to disease in humans. Interestingly, in our dataset there are 418 daSAPs for which the disease-causing residue is found in at least another species, with a total of 5848 cases. This indicates that in some instances protein structures may tolerate the disease-causing residue at that position, probably linked to the existence of compensating mutations. A first approach to analyze the existence of compensating mutations is the study of sequence environment at the mutation site in those cases for which the mutation corresponds to a daSAP in humans. By comparing the sequence identity around the daSAP mutation site *versus* the global sequence identity, we can see whether regions around daSAPs store more variability than expected. Analysis of the data shows that for about 70 % of the cases the local identity around the mutation site is less than the global sequence identity (50 % is expected using a conservative random estimate). This suggests that in some species sequence-local compensating mutations took place to help accommodate without harm the disease-causing mutation. These results are in agreement with data reported by Yang *et al.*,[6] which showed that phenotypes in osteogenesis imperfecta, which range from mild to lethal, are related to the sequence environment at the mutation site. Overall, the bulk of the results suggests that the presence or absence of a disease-causing mutation in a protein family could be used to assess *a priori* the seriousness of the associated disease in humans. In particular, the number of compensating mutations required to accommodate disease-causing mutations could be used as a useful index for this purpose. Although this may be difficult to implement, the study of sequence-local environments may provide a first approach.

Concluding Remarks

Our results confirm that knowledge of protein structure is a first step towards providing a rationale of disease-causing mutations. They also show that some sequence and structural properties can be used to find a quantitative explanation of the pathological character of SAPs. However, our results show also that a better understanding of the molecular basis of disease can come only from the combined knowledge of the principles of protein stability, of the delicate networks of protein interactions in the cell and of the protein physiological environment. Fortunately, in some cases the use of evolutionary relationships may help to circumvent our present limited knowledge in these fields.

The differences between nSAPs and daSAPs observed in the present work were found using structure properties that are easily predictable from only knowledge of sequence. This opens the possibility of predicting the potential pathogenicity of SAPs from only knowledge of the protein sequence, thus allowing us to go beyond the limits of structure-based prediction methods.[29,39] In particular, preliminary results using a very simple procedure, based on the use of mutation matrices, indicate that reasonable predictive power can be attained without the knowledge of protein structure. Active research is pursued in our group in this direction.

Materials and Methods

Disease-associated SAPs (daSAP)

The SAPs used in the present study were obtained after searching the SwissProt database[16] version 39, using the keywords PDB, VARIANT, and HUMAN. Variants corresponding to residues involved in the protein active site, or cysteine residues involved in disulphide bridges were discarded, since the reasons for their pathological impact are evident. The remaining variants were checked manually for their association with disease, either using the SwissProt annotations, or by consulting the original references. Finally, those variants located in a sequence fragment for which there was no structure available were eliminated. The distribution of the resulting variants relative to the proteins in the dataset is shown in Figure 7. A total of 1169 variants distributed between 73 proteins were kept.

It is important to note that not all the selected daSAPs have the same pathological impact; that is, while some of them may lead to mild syndromes, others will be associated with serious diseases. Unfortunately, except for lethal diseases, classifying daSAPs according to their impact is very difficult, as it will depend in many cases on subjective or imprecise observations made by different authors on the state of the patient. In addition, divid-

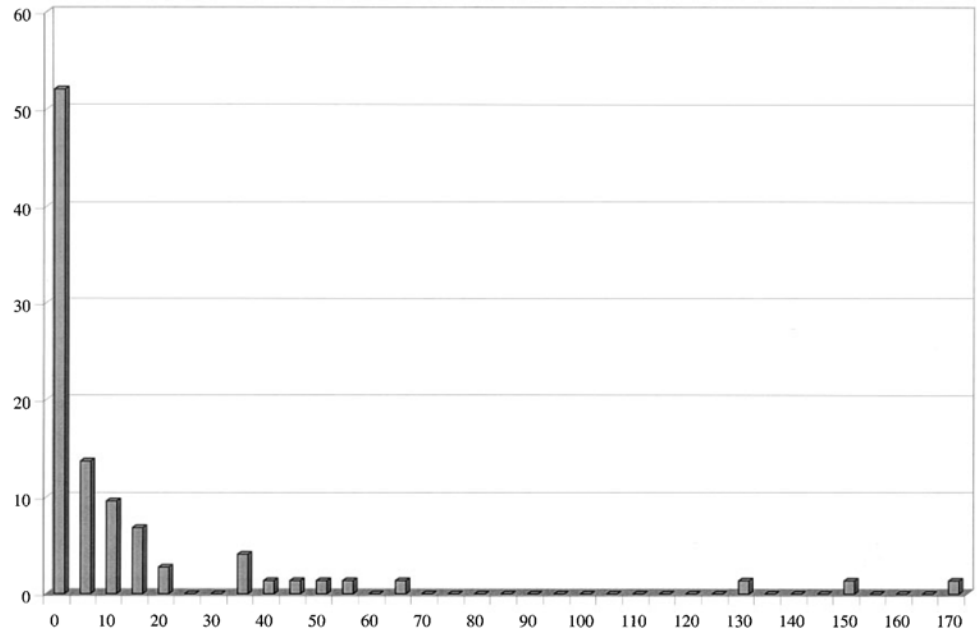

Figure 7. Distribution of daSAPs relative to the proteins in the dataset (see Materials and Methods). Abcissae, #daSAP/protein; ordinates, percentage of proteins.

ing our daSAPs dataset into several subsets according to their impact would result in smaller amounts of data per subset, thus decreasing the statistical strength of our results.

Neutral SAPs (nSAP)

When mapping the disease-causing SAPs to protein structure, a reference state should be used to confirm the trends observed. A first option is to use a simple random model, in which neutral mutations could happen at any position on the set of proteins used for the daSAPs. This model is used as a reference for some selected analysis. However, a more realistic reference model is defined by considering observed neutral mutations for the same set of proteins. That is, those SAPs not associated with any disease. Here we modelled neutral SAPs using the variability observed in multiple sequence alignments (MSA), following the approach described by Sunyaev *et al.*[11] To this end, we took as neutral SAPs those natural variants occuring in different species, after eliminating any human sequence from the MSA. To mimic as closely as possible the occurence of neutral variants, we filtered out those sequences in the MSA having a sequence identity of less than 95% with the human protein.[11]

The whole procedure was applied to the 73 proteins in the above section, giving a total of 741 variants, which constituted our set of nSAPs.

† S.J. Hubband and J.M. Thornton, Department of Biochemistry and Molecular Biology, University College London.

The protein structures

The structures for the 73 proteins were downloaded from the PDB.[44] We found that in many cases there were several structures available for a given protein. To select one of them we utilized the following criteria: X-ray structures were selected over NMR structures; theoretical models were discarded. When possible, we chose native forms over mutants, uncomplexed over complexed states, and high-resolution over low-resolution structures.

Part of our analysis was based on the use of the quaternary structure of the proteins. Coordinates for the quaternary structures were obtained from the PQS database[22] at the European Bioinformatics Institute (EBI-EMBL).

Secondary structure and accessibility computations

Secondary structure was computed using the SSTRUC implementation of the Kabsch & Sander[45] method, by David Keith Smith. Accessibility computations were done using the program NACCESS†. Relative accessibilities, equal to the residue accessibility in the protein divided by its accessibility in an extended Ala-X-Ala peptide, were computed by NACCESS and mapped to three-state accessibilities (buried 0-9%; half-buried 9-36%; exposed 36-100%), following Rost & Sander.[46]

Multiple sequence alignments

MSAs are used at diferent places in this work. In all the cases, the required MSAs were obtained from the Pfam[47] database of high-quality MSA.

It is well known that MSAs may contain a certain amount of redundant information.[48] To eliminate this redundancy from our computations, we used the Henikoff & Henikoff[48] weighting scheme. Obviously, numerical results change after applying the Henikoff & Henikoff[48] procedure to the raw distribution data (data not shown, but available upon request to the authors). However, no significant variations are observed in the overall trend for a given variable, e.g. accessibility data, etc. For example, the percentages of exposed, half-buried and buried residues before applying the Henikoff & Henikoff procedure are 70.8%, 8.9% and 20.2%. After applying the Henikoff & Henikoff procedure we obtain 66.3%, 10.6% and 23.0%, for the same accessibility states.

Measuring the variability at a given MSA position

The variability, or degree of residue conservation, at a given position in an MSA was measured using the Shannon[23] entropy of the amino acid distribution at that position, as has been done in previous studies:[25,29,30,49–52]

$$-\Sigma p_i \ln_2 p_i$$

where subindex i runs over the different amino acid types present in the MSA position under study, the p_i corresponds to the relative frequency of each amino acid type.

As mentioned before, MSAs contain a certain amount of redundant information, which may result in biased estimates of the amino acid frequencies, p_i, at a given position.[25] To address this problem, we followed Altschul *et al.*,[25] and new estimates (q_i) for the amino acid frequencies at a given MSA position were computed as:

$$q_i = \sum_i w_{ij}$$

where subindex j runs over those sequences in the MSA that have a residue of type i at that position, and w_{ij} are the Henikoff & and Henikoff[48] weights for the corresponding sequences. Therefore, the entropy was actually computed using the following equation:

$$-\sum_i \left(\sum_j w_{ij}\right) \ln_2 \left(\sum_j w_{ij}\right)$$

In our case, subindex i runs over the different amino acid types present in the MSA position under study, also including the gaps.[25]

The entropy values will vary from 0. to 4.39, corresponding to high and low degrees of residue conservation at that position.

Statistical tests

To compare the nSAP and daSAP distributions for a given variable we utilized the following procedures. (1) If the variable was continuous, e.g. relative accessibility, we used the Kolmogorov & Smirnov (KS) test, for which the statistic is called D. (2) If the variable was naturally binned, e.g. three-state accessibility, we used the chi-square (CS) test, for which the statistic is called χ^2.

Prediction procedure using mutation matrices

We wanted to assess the performance of the Blosum62 and PAM40 substitution matrices as predictive tools of the pathological character of SAPs. To this end, we followed a very simple procedure in which two thresholds were defined, one for each matrix. The values of each threshold were varied from the minimum value to the maximum value of the matrix elements. Then for each value of the thresholds a given SAP was predicted as: nSAP, if the value of both the Blosum62 and PAM40 elements for this SAP were higher than the corresponding thresholds for each matrix; daSAP, if the value of both the Blosum62 and PAM40 elements for this SAP were lower than the corresponding thresholds for each matrix. No prediction was made for those SAPs not fulfilling any of the previous conditions.

The performance of this method was assessed following a fivefold cross-validation procedure, in which both the daSAP and the nSAP data were divided randomly into five SAP subsets. Then, we built a training set using four subsets, and found the optimal value of the Blosum62 and PAM40 thresholds, using as an objective function the Matthews coefficient.[51] The latter takes into account the number of true positives and negatives, as well as false positives and negatives. The performance of these thresholds was then assessed in the remaining subset, called the test dataset. This procedure was repeated five times, leaving each of the subsets out of the training set in turn. The performance of the method was then evaluated as the average performance over the five tests sets. We used two performance measures: (1) percentage of correct predictions, which is the number of correctly predicted daSAPs and nSAPs, divided by the total number of predictions made; (2) the Matthews coefficient.[51]

References

1. Perutz, M. (1992). *Protein Structure: New Approaches to Disease and Therapy*, W.H. Freeman and Company, New York.
2. Zuckerkandl, E. & Pauling, L. (1962). Molecular disease, evolution and genetic heterogeneity. In *Horizons in Biochemistry* (Marsha, M. & Pullman, B., eds), pp. 189-225, Academic Press, London.
3. Busquets, C., Merinero, B., Christensen, E., Gelpi, J. L., Campistol, J., Pineda, M. *et al.* (2000). Glutaryl-CoA dehydrogenase deficiency in Spain: evidence of two groups of patients, genetically, and biochemically distinct. *Pediatr. Res.* **48**, 315-322.
4. Musco, G., Stier, G., Kolmerer, B., Adinolfi, S., Martin, S., Frenkiel, T. *et al.* (2000). Towards a structural understanding of Friedreich's ataxia: the solution structure of frataxin. *Structure*, **8**, 695-707.
5. Zhang, X., Moréra, S., Bates, P. A., Whitehead, P. C., Coffer, A. I., Hainbucher, K. *et al.* (1998). Structure of an XRCC1 BRCT domain: a new protein-protein interaction module. *EMBO J.* **17**, 6404-6411.
6. Yang, W., Battineni, M. L. & Brodsky, B. (1997). Amino acid sequence sequence environment modulates the disruption by osteogenesis imperfecta glycine substitutions in collagen-like peptides. *Biochemistry*, **36**, 6930-6935.
7. Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G. *et al.* (2001). The sequence of the human genome. *Science*, **291**, 1304-1351.

8. International human sequencing consortium. (2001). Initial sequencing and analysis of the human genome. *Nature,* **409**, 860-921.
9. Chakravarti, A. (2001). Single nucleoide polymorphisms...to a future of genetic medicine. *Nature,* **409**, 822-823.
10. Sunyaev, S., Lathe, W., III & Bork, P. (2001). Integration of genome data and protein structures: prediction of protein folds, protein interactions and "molecular phenotypes" of single nucleotide polymorphisms. *Curr. Opin. Struct. Biol.* **11**, 125-130.
11. Sunyaev, S., Ramensky, V. & Bork, P. (2000). Towards a structural basis of human non-synonymous single nucleotide polymorphisms. *Trends Genet.* **16**, 198-200.
12. O'Brien, D. P., Gale, K. M., Anderson, J. S., McVey, J. H., Miller, G. J., Meade, T. W. & Tuddenham, E. G. (1991). Purification and characterization of factor VII 304-Gln: a variant molecule with reduced activity isolated from a clinically unaffected male. *Blood,* **78**, 132-140.
13. Jameson, J. L. & Hollenberg, A. N. (1992). Recent advances in studies of the molecular basis of endocrine disease. *Horm. Metab. Res.* **24**, 201-209.
14. Matthews, B. W. (1995). Studies on protein stability with T4 lysozyme. *Advan. Protein Chem.* **46**, 249-278.
15. Fersht, A., Matouschek, A. & Serrano, L. (1992). The folding of an enzyme. I. Theory of protein engineering analysis of stability and pathway of protein folding. *J. Mol. Biol.* **224**, 783-804.
16. Bairoch, A. & Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucl. Acids Res.* **28**, 45-48.
17. Henikoff, S. & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA,* **89**, 10915-10919.
18. 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, DC.
19. Fauchère, J.-L. & Pliska, V. (1983). Hydrophobic parameters of amino acid side-chains from the partitioning of *N*-acetyl-amino-acid amides. *Eur. J. Med. Chem.-Chim. Ther.* **18**, 369-375.
20. Chothia, C. (1975). Structural invariants in protein folding. *Nature,* **254**, 304-308.
21. Chou, P. Y. & Fasman, G. D. (1974). Conformational parameters for amino acids in helica, beta-sheet and coil regions calculated from proteins. *Biochemistry,* **13**, 211-222.
22. Henrick, K. & Thornton, J. M. (1998). PQS: a protein quaternary structure file server. *Trends Biochem. Sci.* **23**, 358-361.
23. Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Tech. J.* **27**, 379-423.
24. Blundell, T. L. (1996). Structure-based drug design. *Nature,* **384**, 23-26.
25. Altschul, S. F., Madden, T. L., Schäffer, 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. *Nucl. Acids Res.* **25**, 3389-3402.
26. Pearson, W. R. (2000). Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol. Biol.* **132**, 185-219.
27. Attwood, T. K. & Parry-Smith, D. J. (1999). *Introduction to Bioinformatics*, Longman, Leeds, UK.
28. Topham, C. M., Srinivasan, N. & Blundell, T. L. (1997). Prediction of the stability of protein mutants based on structural environment-dependent amino acid substitution and propensity tables. *Protein Eng.* **10**, 7-21.
29. Wang, Z. & Moult, J. (2001). SNPs, protein structure, and disease. *Hum. Mutat.* **7**, 263-270.
30. Ng, P. C. & Henikoff, S. (2001). Predicting deleterious amino acid substitutions. *Genome Res.* **11**, 863-874.
31. Goldenberg, D. P. (1992). Mutational analysis of protein folding and stability. In *Protein Folding* (Creighton, T. E., ed.), pp. 353-404, W.H. Freeman and Company, New York.
32. Alber, T. & Matthews, B. W. (1987). Structure and thermal stability of phage T4 lysozyme. *Methods Enzymol.* **154**, 511-533.
33. Funahashi, J., Takano, K., Yamagata, Y. & Yutani, K. (2000). Role of surface hydrophobic residues in the conformational stability of human lysozyme at three different positions. *Biochemistry,* **39**, 14448-14456.
34. Pakula, A. A. & Sauer, R. T. (1990). Reverse hydrophobic effects relieved by amino-acid substitutions at a protein surface. *Nature,* **344**, 363-364.
35. Liu, R., Baase, W. A. & Matthews, B. W. (2000). The introduction of strain and its effects on the structure and stability of T4 lysozyme. *J. Mol. Biol.* **295**, 127-145.
36. Wray, J. W., Baase, W. A., Lindstrom, J. D., Weaver, L. H., Poteete, A. R. & Matthews, B. W. (1999). Structural analysis of a non-contiguous second-site revertant in T4 lysozyme shows that increasing the rigidity of a protein can enhance its stability. *J. Mol. Biol.* **292**, 1111-1120.
37. Baldwin, E., Xu, J., Hajiseyedjavadi, O., Baase, W. A. & Matthews, B. W. (1996). Thermodynamic and structural compensation in "size-switch" core repacking variants of bacteriophage T4 lysozyme. *J. Mol. Biol.* **259**, 542-559.
38. Xu, J., Baase, W. A., Baldwin, E. & Matthews, B. W. (1998). The response of T4 lysozyme to large-to-small substitutions within the core and its relation to the hydrophobic effect. *Protein Sci.* **7**, 158-177.
39. Sunyaev, S., Ramensky, V., Koch, I., Lathe, W., III, Kondrashov, A. S. & Bork, P. (2001). Prediction of deleterious human alleles. *Hum. Mol. Genet.* **10**, 591-597.
40. Sheinerman, F. B., Norel, R. & Honig, B. (2000). Electrostatic aspects of protein-protein interactions. *Curr. Opin. Struct. Biol.* **10**, 153-159.
41. Dobson, C. M. (2001). The structural basis of protein folding and its links with human disease. *Phil. Trans. Roy. Soc. ser. B,* **356**, 133-145.
42. Larsen, T. A., Olson, A. J. & Goodsell, D. S. (1998). Morphology of protein-protein interfaces. *Structure,* **6**, 421-427.
43. Baraldi, E., Carugo, K. D., Hyvonen, M., Surdo, P. L., Riley, A. M., Potter, B. V. *et al.* (1999). Structure of the PH domain from Bruton's tyrosine kinase in complex with inositol 1,3,4,5-tetrakisphosphate. *Structure Fold. Des.* **7**, 449-460.
44. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H. *et al.* (2000). The Protein Data Bank. *Nucl. Acids Res.* **28**, 235-242.
45. Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers,* **22**, 2577-2637.

46. Rost, B. & Sander, C. (1994). Conservation and prediction of solvent accessibility in protein families. *Proteins: Struct. Funct. Genet.* **20**, 216-226.
47. Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Howe, K. L. & Sonnhammer, E. L. (2000). The Pfam Protein Families Database. *Nucl. Acids Res.* **28**, 263-266.
48. Henikoff, S. & Henikoff, J. G. (1994). Position-based sequence weights. *J. Mol. Biol.* **243**, 574-578.
49. Atchley, W. R., Wollenberg, K. R., Fitch, W. M., Terhalle, W. & Dress, A. W. (2000). Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis. *Mol. Biol. Evol,* **17**, 164-178.
50. Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. (1992). *Numerical Recipes in C*, Cambridge University Press, Cambridge.
51. Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta,* **405**, 442-451.
52. Sander, C. & Schneider, R. (1991). Database of homology-derived protein structures. *Proteins: Struct. Funct. Genet.* **9**, 56-68.

Edited by J. Thornton

(Received 2 April 2001; received in revised form 8 November 2001; accepted 11 November 2001)

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.

Atchley, W. R., Wollenberg, K. R., Fitch, W. M., Terhalle, W. & Dress, A. W. (2000). Correlations Among Amino Acid Sites in bHLH Protein Domains: An Information Theoretic Analysis. *Mol. Biol. Evol.* 17, 164-178.

Bairoch, A. & Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res* 28, 45-8.

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.

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.

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

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.

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.

Dayhoff, M. O., Ed. (1972). Atlas of Protein Sequence and Structure. Vol. 5.

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.

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. (2000). Amino acid substitution matrices. *Adv Protein Chem* 54, 73-97.

Hubbard, S. J. & Thornton, J. M. (1993). 'NACCESS', Computer Program,. Department of Biochemistry and Molecular Biology, University College London, London.

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

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., et al. (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860-921.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta* 405, 442-51.

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

Press, W. H., Flannery, B. P., Teukolsky, S. A. & Vetterling, W. T. (1992). *Numerical recipes in C : the art of scientific computing*. 2nd edit, Cambridge University Press, Cambridge ; New York.

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

Sander, C. & Schneider, R. (1991). Database of homology_derived protein structures. *Proteins* 9, 56-68.

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

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

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.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., et al. (2001). The sequence of the human genome. *Science* 291, 1304-51.

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