Capítol 7.

Variació en el caràcter patològic de les mutacions puntuals dependents de l'entorn estructural.

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

I. INTRODUCCIÓ

De l'estudi comparat de seqüències proteiques de diferents espècies es va observar que alguns aminoàcids causants de malalties en humans eren trobats com a aminoàcid neutre en altres espècies. Així, Zuckerkandl i Pauling (Zuckerkandl & Pauling, 1962) van descriure la presència en la hemoglobina salvatge d'orangutà d'un residu associat a una malaltia lleu quan apareixia en humans. Estudis més recents mostren que aquestes variants poden ser relativament freqüents (Huang et al., 2004; Kondrashov et al., 2002; Kulathinal et al., 2004; Pavlicek et al., 2004). En un cas semblant la substitució de l'arginina salvatge en la posició 142 per glutamina en l'adenosina desaminasa està relacionada amb l'immunodeficiència combinada greu, mentre que la glutamina apareix en posicions equivalents en l'ortòleg de ratolí sense efecte aparent (Gao & Zhang, 2003; Huang et al., 2004). Aquestes mutacions també s'han observat en estudis comparats de Drosophila melanogaster amb altres genomes de dípters (Kulathinal et al., 2004).

Seguint la nomenclatura de Kondrashov et al. (Kondrashov et al., 2002) ens referirem a aquestes mutacions com a CPDs (*Compensated Pathogenic Deviations*) Desviacions Patològiques Compensades tot i que cal remarcar que també són anomenats FDDAM (*Fixed Differences of Disease-Associated Mutations*, Diferències Fixades de Mutacions Associades a Malaltia) per Gao i Zhang (Gao & Zhang, 2003).

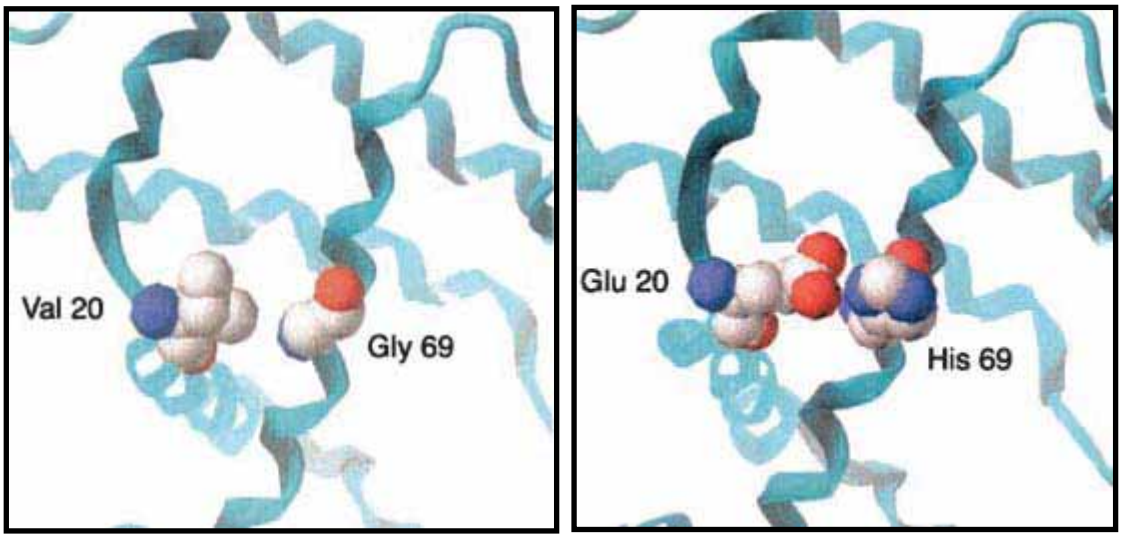
En aquest treball presentem l'estudi de fins a 70000 CPDs associades a mutacions puntuals patològiques derivades de la base de dades SwissProt (Apweiler et al., 2004) segons la metodologia desenvolupada en treballs previs (Ferrer-Costa et al., 2002; Ferrer-Costa et al., 2004; Ng & Henikoff, 2001; Sunyaev et al., 2001). L'estudi, basat en propietats estructurals i físico-químiques de les CPDs i la seva comparació amb les mutacions puntuals i neutres usades en treballs anteriors (Ferrer-Costa et al., 2002; Ferrer-Costa et al., 2004), intenta aportar llum a les raons que guien l'aparició d'aquestes CPDs.

En la revisió de la bibliografia recent referent a aquest fenomen cal destacar en primer lloc el treball de Kondrashov et al. (Kondrashov et al., 2002). Aquí els autors analitzen un conjunt de 32 proteïnes per cada una de les quals es coneix més de 50 mutacions puntuals patològiques i més de tres ortòlegs amb una identitat de seqüència superior al 50%. Per cada una de les mutacions patològiques analitzades busquen CPDs, en detecten 608, que responen als següent criteri de qualitat. les CPDs no presenten cap gap en 10 residus per sobre i per sota de la mutació en l'alineament entre la proteïna humana i l'ortòloga d'una altra espècie. De les anàlisis que fan els autors en deriven que el 10 % de totes les desviacions de les proteïnes no humanes respecte els seus ortòlegs humans són CPDs, és a dir que 1 de cada 10 de les mutacions observades en les proteïnes durant l'evolució serien en principi patològiques però esdevenen benignes gràcies a l'existència d'una mutació compensatòria. D'aquesta manera hipotetitzen que totes les CPDs existeixen gràcies a la compensació generada per una altra mutació. Malauradament la confirmació d'aquesta hipòtesi passa per la detecció de les mutacions compensatòries i aquesta no és una tasca fàcil. De manera general el coneixement de les estructures de les proteïnes i l'anàlisi dels alineaments múltiples de seqüència pot mostrar alguns residus candidats a ser mutacions compensatòries, així i tot poden aparèixer diferents candidats o cap de clar. Els autors centren la seva cerca de mutacions compensatòries de CPDs a aquelles que interactuen de manera directa amb la CPD, i que trenquen per tant interaccions directes estabilitzadores. Amb tot es poden donar casos de mutacions compensatòries a llarga distància així com mutacions que es troben en altres proteïnes que interactuen amb la proteïna estudiada.

En el cas de les mutacions compensatòries que interactuen de manera directa amb la CPD el mecanisme d'acció seria el següent. L'aparició de la CPD trencaria una interacció directa i estabilitzadora amb l'entorn. la posterior aparició de la mutació compensatòria generaria una nova interacció estabilitzadora. Aquesta hipòtesi també s'aplicaria, segons els autors, a interaccions a llarga distància malgrat que reconeixen que sense treball experimental és impossible de

determinar quins són els residus que interactuen.

Fig 1a **Fig 1b**



Dels exemples presentats pels autors cal destacar la mutació patològica sobre la β-Hemoglobina humana a la posició 20 on una Valina és substituïda per àcid Glutàmic. En les anàlisis fetes sobre els alineaments múltiples de seqüència s'observa la presència de Glutàmic en la posició 20 en altres espècies i sempre acompanyada per Histidina en la posició 69, tal i com s'observa en la figura 1b i que correspon a la β-Hemoglobina de cavall. En canvi en la proteïna humana (fig 1a) que presenta Valina en la posició 20 presenta a la posició 69 Glicina. Els autors interpreten que en la proteïna de cavall la His-69 interacciona amb Glu-20 possiblement per la formació de pont d'hidrogen. En la proteïna humana la interacció de la Val-20 amb la Gly-69 possiblement és deguda a interacció de Van der Waals. D'aquesta manera els autors conclouen que la mutació Gly-69-His pot compensar l'efecte patològic de la mutació Val-20-Glu.

En un article posterior de Gao i Zhang (Gao & Zhang, 2003) es presenta l'estudi comparatiu de 7293 mutacions patològiques en 687 gens humans amb els seus ortòlegs de ratolí. En la gran majoria de casos les posicions estudiades contenen el mateix residu tant en la proteïna humana salvatge com en la de ratolí (6587

posicions) o bé un residu que no correspon ni amb el salvatge ni amb el mutant humà (547 posicions). En només un 2% dels casos el residu salvatge en ratolí correspon amb el residu patològic en humans. Aquest 2% correspon a 160 casos dels quals 20 han estat experimentalment demostrats la relació directa entre la mutació i la malaltia i comprovats a nivell de seqüència que les posicions en ratolí són fixades. Els autors mitjançant l'anàlisi d'aquestes 20 mutacions intenten donar explicació a la fixació en ratolí de les mutacions associades a malaltia en humans. Donen cert pes a un efecte fundador en ratolí o una reduïda selecció en fenotips de malaltia que apareixen en edats avançades, però malgrat tot assumeixen que calen altres explicacions per la gran majoria de mutacions, i accepten que probablement l'explicació més clara i simple seria l'existència de mutacions compensatòries.

En l'estudi a nivell de seqüència, a nivell físico-químic i estructural de la col·lecció de CPDs intentem entendre les raons evolutives que permeten a aquestes mutacions ser acceptades en les proteïnes. Tal i com es mostrarà, malgrat que les mutacions compensatòries semblen tenir un pes important en l'explicació d'aquest fenomen altres raons semblen tenir també certa importància. Entre aquestes possibles explicacions hi ha que les CPDs tendeixen a ser menys lesives que la resta de mutacions puntuals o que l'estructura s'adapta en certa manera per acceptar aquesta mutació.

II. MATERIALS I MÈTODES

Mutacions patològiques neutres i CPDs

Com a material de partida es van usar les 9334 mutacions puntuals patològiques obtingudes de la base de dades Swissprot (Apweiler et al., 2004) sobre un total de 811 proteïnes humanes tal i com es descriu en altres treballs propis(Ferrer-Costa et al., 2002; Ferrer-Costa et al., 2004). També es van usar les 11374

mutacions neutres derivades segons el model evolutiu(Ferrer-Costa et al., 2002; Ferrer-Costa et al., 2004; Ng & Henikoff, 2001; Sunyaev et al., 2001) sobre les 811 proteïnes humanes .

Pel que respecte als CPDs cal destacar que constitueixen, per definició, un subconjunt de les mutacions patològiques i per tant es van obtenir a partir d'aquestes. Per cada mutació patològica es va localitzar la seva posició en l'alineament PFAM (Bateman et al., 2002) corresponent. Una vegada localitzada la posició es va buscar en la columna corresponent de l'alineament totes les seqüències que mostraven com a residu salvatge el que apareixia com a mutant patològic per la seqüència humana. Es van eliminar tots aquells casos en que les dues seqüències alineades eren humanes. També es van eliminar aquelles seqüències que en l'alineament PFAM tenien menys d'un 10% d'identitat de seqüència per garantir una qualitat mínima en l'alineament. Com a resultat d'aquesta selecció es van obtenir un total de 71295 CPDs, que es deriven de 341 proteïnes humanes i que tenen lloc en 20986 proteïnes diferents que són en 2341 espècies.

Per cada cas es va comprovar si alguna de les dues seqüències, bé la humana, be la de l'altra espècie tenien part o tota la seva estructura resolta, consultant la base de dades PDB (Berman et al., 2000). Per totes les seqüències que tenien PDB associat es va mapar l'estructura sobre la seqüència amb la intenció de col·lectar totes les mutacions per les quals es coneixia alguna estructura. Per aquelles seqüències que es coneixia més d'una estructura es va seleccionar aquella que cobria més aminoàcid i una resolució més elevada. En total per 17282 d'aquests CPDs es coneix l'estructura d'una de les dues seqüències alineades per 185 CPDs es coneixen les estructures per les dues seqüències.

S'han dividit el tipus d'anàlisi en dues categories segons: (i) no hi hagi cap estructura coneguda, (ii) es conegui l'estructura d'almenys una de les proteïnes involucrades.

Descriptors de les mutacions

Es van usar dos tipus de descriptors per estudiar les CPDs: accessibilitat al solvent i propietats dels aminoàcids. El primer dóna una idea aproximada de com el context estructural d'una mutació està relacionat amb el dany que pot causar (Ferrer-Costa et al., 2002; Matthews, 1995; Sunyaev et al., 2000). Per altra banda les propietats del aminoàcids ens donen una idea de la natura del dany introduït pel canvi aminoacídic des del punt de vista fisicoquímic o de la naturalesa dels aminoàcids. D'aquesta manera mutacions que representen un canvi de volum gran resultaran més probablement en una desestabilització de la proteïna o una pèrdua de funció.
Molts d'aquests descriptors han estat usats i descrits llargament (Ferrer-Costa et al., 2004) en capítols anteriors..

Accessibilitat al solvent: Els valors d'accessibilitat es van obtenir usant el programa NACCESS.

Propietats de Residu/Seqüència: Els canvis físico-químics induïts per una mutació són mesurats usant dos paràmetres que corresponen a les següents propietats: hidrofobicitat i volum. Per cada un d'aquests paràmetres el valor associat a la CPD es descriuencom la diferència $\Delta x = x_m - x_w$, on x_m i x_w corresponen als valors de la propietat en la seqüència no humana i la humana respectivament. Canvis en la hidrofobicitat, $\Delta\Delta G_{trf}$ s'han descrit usant les mesures d' energia lliure de transferència entre aigua i octanol. Canvis en la grandària AV, s'han descrit usant els volums de van der Waals.
Finalment, els valors de la matriu de mutació Blosum62 s'han usat per descriure també les mutacions, ja que inclouen diferents contribucions que es poden relacionar amb el dany causat per una mutació.

Entorn espaial del residu: Per aquelles CPDs que ocorren en una proteïna amb estructura coneguda es van computar la llista dels veïns en l'entorn tridimensional de la següent manera. Primer es va obtenir l'accessibilitat al solvent per totes les proteïnes, usant el programa NACCESS. Seguidament es van eliminar tots els àtoms del residu de la posició de la CPD i es va computar de nou l'accessibilitat. Finalment es va considerar com a veïns del residu estudiat tots els residus on s'observa un canvi en l'accessibilitat relativa.

En els casos en què es coneix l'estructura tridimensional de les proteïnes alineades en la zona de la mutació estudiada es va calcular l'alineament estructural amb MAMMOTH (Ortiz et al., 2002), i es va determinar l'entorn local al voltant del residu per les dues estructures tal i com s'ha explicat anteriorment.

Es va calcular RMSd en carbonis alfa entre les dues estructures basant-se en l'alineament estructural determinat per MAMMOTH. El RMSd global es va fer per totes la proteïna i el RMSd local es va fer per l'entorn local al voltant de la mutació.

La identitat de seqüència estructural local es va calcular com la relació entre residus idèntics sobre tots els residus alineats en l'entorn local sobre l'alineament estructural.

Distribució de probabilitat d'identitat de seqüència local: Aquesta distribució s'usa per tal de poder detectar la presència de possibles mutacions compensatòries, de forma global. Si s'ignora l'existència de gaps, la distribució de probabilitat de la identitat local de seqüència al voltant de la CPD pot ser aproximada a la binomial Bi(n, p), on n és el número de veïns i p, l'anteriorment mencionada fracció, que és equivalent a la probabilitat de que dos residus alineats siguin idèntics (és la identitat local). Per aquelles CPDs que no es coneix informació estructural tridimensionals s'ha usat un valor n de 8, i correspon a tots els residus entre les posicions de seqüència i-4 i i+4, on i és la localització de la CPD en la proteïna humana (òbviament degut a que busquem mutacions

compensatòries, la posició de la CPD no és inclosa en aquest model). Aquest rang de seqüència no és arbitrari, ja que inclou les interaccions entre residus i-i+2 i i-i+3/i+4 associades als elements d'estructura secundària (cadena beta i hèlix alfa, respectivament). Per aquelles CPDs per les quals es coneix almenys l'estructura tridimensional d'una de les dues seqüències s'ha computat la binomial pel número de veïns corresponents, calculant al final la suma ponderada de binomials per obtenir el valor final corresponent al interval d'identitat global estudiat.

III. ARTICLE DE RECERCA

How can non-human proteins accommodate as wild-type human disease-associated residues. Carles Ferrer-Costa, Modesto Orozco, Xavier de la Cruz. Submitted in TIG (Trends in Genetics)

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

HOW CAN NON-HUMAN PROTEINS ACCOMMODATE AS WILD-TYPE HUMAN DISEASE-ASSOCIATED RESIDUES

Carles Ferrer-Costa[1], Modesto Orozco, [1,2,3*], Xavier de la Cruz [1,3,4*]

[1] Molecular Modeling and Bioinformatics Unit. Institut de Recerca Biomédica. Parc Científic de Barcelona. Josep Samitier 1-5. 08028 Barcelona, Spain.

[2] Departament de Bioquímica i Biologia Molecular. Facultat de Química. Universitat de Barcelona. Martí i Franquès 1. 08028 Barcelona, Spain.

[3] Institut Nacional de Bioinformática. Fundación Genoma España.

[3] Institució Catalana per la Recerca i Estudis Avançats (ICREA), Passeig Lluís Companys, 23, 08018 Barcelona, Spain.

*** Send correspondence** to Xavier de la Cruz or Modesto Orozco. Parc Científic de Barcelona; C/Josep Samitier, 1-5; 08028 Barcelona; SPAIN. Fax: +34 93 403 71 57. Email: xavier@mmb.pcb.ub.es or modesto@mmb.pcb.ub.es.

KEYWORDS: disease-associated mutations, sequence variability, protein sequence, protein structure, site-directed mutagenesis, bioinformatics

ABSTRACT

Comparative studies between the human genome and the recently sequenced mouse and rat genomes has produced several instances of an interesting phenomenon: the presence in wild type non-human proteins of residues which induce pathology when present at an equivalent position in the human protein. Here we characterize a large set of cases in terms of both their structural environment and the intrinsic properties of the amino acid change. We find that apart from the existence of compensatory mutations, aminoacid substitutions that are pathological in humans but neutral in other related organisms are generally associated to smaller changes in physical/evolutionary properties than the average disease-associated mutations. Furthermore, the protein loci where these type of mutations occur tend to be more resilient.

The existence of single point aminoacid mutations leading to pathologies in humans, but that are present as wild type proteins is known since the sixties, were in a seminal work Zuckerkandl and Pauling (Zuckerkandl & Pauling, 1962) described the presence in orang-utan wild-type haemoglobin of a residue associated to a mild disease, when present in human haemoglobin. The analysis of the massive amount of genomic data currently available has shown that the situation described by Zuckerkandl and Pauling is not so rare (Huang et al., 2004; Kondrashov et al., 2002; Kulathinal et al., 2004; Pavlicek et al., 2004; Waterston et al., 2002). For example, the mutation Val198Met in human cryopyrin is associated to the familial cold autoinflammatory syndrome, while Met is found at an equivalent position in both mouse and rat wild-type sequences (Huang et al., 2004). Similarly, Arg142Gln mutation in human adenosine deaminase is associated with severe combined immunodeficiency disease, while the glutamine residue appears at an equivalent position in the mouse homolog with no apparent effect (Gao & Zhang, 2003; Huang et al., 2004). Following Kondrashov and co-workers (Kondrashov et al., 2002), we will refer to mutations that are pathological in humans but neutral in other organism as "compensated pathogenic deviations (CPD)".

CPD have been recently studied within an evolutionary context by Kondrashov et al. (Kondrashov et al., 2002) and Gao & Zhang (Gao & Zhang, 2003). The first authors reported a survey of the CPD presence in different mammalian proteins. By combining information from multiple sequence alignments and three-dimensional structures they identified a series of compensatory mutations that would justify the existence of CPDs. In that analysis Kondrashov et al. found that the mutation Val20Glu associated to disease in humans, while Glu appears at an equivalent position in wild-type horse haemoglobin. Analysis of sequences show that this is related to the

compensatory mutation present in horse haemoglobin, Gly69His, that reduces the damage caused by the presence of a Glu residue at position 20. Gao & Zhang (Gao & Zhang, 2003), focusing in a small number -20- of well-characterised human-mouse CPD, concluded that the presence of compensatory mutations is the most likely explanation for the existence of CPD. It is then clear that compensatory mutations are largely responsible for the existence of CPDs, but are other hidden reasons?. For example: do mutations of the wild-type human residue to a classical DAMU involve larger physicochemical changes than in CPD.?, or do CPDs occur at protein loci with some special properties different to those of normal DAMUs.

Here address the problem of how non-human proteins can accommodate CPD, without apparent damage. To this end, we briefly address the issue of compensatory mutations but mainly focus in the study of the structure context of CPD, and in their characterization in terms of a series of sequence properties. The resulting distributions are compared with those for DAMUs and neutral mutations (NEMUs). Interestingly, we find that CPD display an intermediate behaviour between NEMUs and DAMUs, supporting the idea that a substantial proportion of CPD may correspond to slightly deleterious mutations. We also find that CPD tend to happen either at more resilient positions, or at positions that have undergone a certain degree of structural divergence at the CPD location, or close to it.

MATERIALS AND METHODS

SELECTION OF DAMUs, NEMUs and CPD

As in previous works (Ferrer-Costa et al., 2002; Ferrer-Costa et al., 2004)

disease-associated mutations (DAMUs) were extracted by queering SWISSPROT(Apweiler et al., 2004), version 40, with the keywords: DISEASE, VARIANT, and HUMAN/MOUSE. The hits found were filtered by human inspection to the literature. This yields to a total of 9334 pathological mutations in 811 human proteins. Neutral mutations (NEMUs) were defined using an evolutionary protocol (Ferrer-Costa et al., 2004; Santibañez-Koref et al., 2003; Sunyaev et al., 2000). Accordingly, a mutation is defined as NEMU when it appears in closely related sequences from other organisms. The alignments needed to detect NEMUs were taken by eliminating from the Pfam (Bateman et al., 2000) alignments all the human proteins and all sequences with less than 95 % identity to the target human sequence. This procedure yielded 11374 human NEMUs.

CPD are defined as those human DAMUs for which the disease-associated residue appears as wild-type in another species (Kondrashov et al., 2002). We then analyzed Pfam multiple alignments for the human protein near positions where DAMU are found. To reduce noise arising from the poor quality of the alignments for very remote sequences we eliminated from our alignments those non-human sequences with less than 10 % sequence identity to the human protein (we define this filtering procedure after considering other schemes like that suggested by Kondrashov et al. (Kondrashov et al., 2002)). The rest of the alignment is the scanned to find cases where human DAMUs are present in aligned non-human sequences. This procedure leads to a total of 71295 CPD happening in 20986 non-human proteins (corresponding to 341 human proteins) and 2341 species, 17282 of these CPDs mapped on proteins of known structure. In some cases the same CPD can be observed in different species and in fact the number of CPD locations in human proteins is only 2278.

MUTATION DESCRIPTORS

Solvent accessibility. It was measured for proteins with known three dimensional structure (in any species). Accessibility values were obtained using the program NACCESS (Hubbard & Thornton, 1993) with its default parameters.

Residue/Sequence properties. The physico-chemical changes induced by a mutation are measured by the changes in hydrophobicity (taken from water/octanol measurements; ΔΔGtr Fauchere & Pliska (Fauchere & Pliska, 1983)) and in molecular molume (ΔV; REF) induced by the aminoacid change.

Evolutionary index. Blosum62 mutation matrix(Henikoff & Henikoff, 1993) elements were used to determine the evolutionary impact of a given mutation. This index can be used to have an estimate of the predicted damage in protein function caused by a mutation (Ferrer-Costa et al., 2004)(Nishikawa**).

SPATIAL RESIDUE NEIGHBOURS

For those CPD happening in a protein with known structure, we computed the list of three-dimensional neighbours as follows. First, we obtained the solvent accessibility for all the protein, using the NACCESS package (Hubbard & Thornton, 1993). Second, we manually eliminated all the atoms of the residue happening at the CPD position and computed the protein accessibility again. We then considered neighbours of the target residue as those residues undergoing a change in solvent accessibility.

PROBABILITY DISTRIBUTION OF LOCAL SEQUENCE IDENTITY

Let us consider a CPD happening between a human sequence and a non-human homolog, with a fraction p of identical residues between both sequences, after sequence alignment. If we ignore the existence of gaps, the probability distribution of the local sequence identity around the CPD can be approximated by a binomial Bi(n, p), where n is the number of neighbours and p, the above mentioned fraction which is equivalent to the probability that a pair of aligned residues are identical. We utilise two different n values, 8 and 10, corresponding to two types of CPD neighbours: sequence and three-dimensional. The former is utilised for those CPD for which no 3D-structure information is available, and corresponds to all the residues between sequence positions i-4 and i+4, where i is the CPD location in the human protein sequence on (obviously, because we are looking for compensatory mutations, the CPD position is not included in this model). This sequence range is not arbitrary, as it will include the important i-i+2 and i-i+3/i+4 residue-residue interactions associated to secondary structure elements (beta strand and alpha helix, respectively). The second is defined on the basis of accessibility computations as described in the previous section.

RESULTS AND DISCUSSION

Distribution of CPD among species

First, we looked at the distribution of sequence identities between human and other species, relative to the identity distribution between the human and non-human proteins carrying at least one CPD (Figure 1). We observe that as the non-human

sequence diverges from that of the human homolog, there is an increase in the chance that CPD appear. This would be in agreement either with the existence of specific compensatory mutations able to rescue the damaging effect of DAMUs (Gao & Zhang, 2003; Kondrashov et al., 2002), or with the existence of large compensatory changes in the structural environment at the CPD location (Russell & Barton, 1994).

Presence of compensatory mutations

We first decided to explore the existence of compensatory mutations following an approach slightly different from that of Kondrashov and co-workers (Kondrashov et al., 2002). These authors combine structural analysis together with the use of multiple sequence alignments to identify possible compensatory mutations of a given CPD (Kondrashov et al., 2002). The simultaneous presence in different species of both mutations, the CPD and its putative compensatory mutation, supports the role of the latter (Kondrashov et al., 2002). Here, we have tried to find sequence identity biasses in the neighbourhood of the CPD that would point to the existence of compensatory mutations. The rationale behind is that in some cases we probably can talk of a compensatory environment -constituted by several mutations, rather than of a single compensatory mutation.

For those CPD for which only sequence information is available, we looked at the distribution of sequence neighbours around the CPD for homolog pairs with high sequence identities, bigger than 90 %. While there seems to be a slight trend towards lower sequence identities (Figure 2A), suggesting the presence of compensatory mutations among sequence neighbours, the difference among distributions is not significant (p-value=0.** for a chi-square test).

A different situation arises when we consider the three-dimensional environment of the CPD (Figure 2B). If we concentrate in the case of homolog pairs for which sequence identity is large, between 90 % and 100 %, we find a significant biass in the identity distribution towards lower sequence identities (p-value=0.**, for a chi-square test). This is in accordance with the existence of compensatory mutations, although the distribution shape points to a very small number of them. This trend disappears as the sequence identity among homologs decreases (results not shown), indicating that the structure changes associated with sequence divergence (Russell & Barton, 1994) provide a friendly environment for the CPD.

Location of CPD in the three-dimensional structure

Our previous results (Figures 2A-2B) are in accordance with the idea that compensatory mutations allow the existence of CPD (Gao & Zhang, 2003; Kondrashov et al., 2002). However, it is unclear whether compensatory mutations are always required to accomodate CPD, and in fact Gao & Zhang (Gao & Zhang, 2003) consider the possibility that CPD may be only slightly deleterious. This can happen in two different ways: (i) CPD occur at less damaging locations in the protein structure than DAMUs; (ii) the nature of the amino acid change associated with the CPD is less extreme, an option that will be discussed in the next section.

Site-directed mutagenesis studies show that the impact of a given mutation clearly depends on its structural context (Matthews, 1993). While the latter can be described with great detail (Fersht & Serrano, 1993), a simple variable like residue accessibility to the solvent provides a coarse-grained description with a surprinsingly

large explanatory power on the effect of single-point mutations (Matthews, 1993): it has been found that, in general, mutations happening at exposed locations in the protein structure are less likely to be destabilizing than those occuring at the protein core (Matthews, 1993). Actually, when applied to the study/prediction of DAMUs, accessibility has proven quite fruitful (Chasman & Adams, 2001; Ferrer-Costa et al., 2002; Ferrer-Costa et al., 2004; Sunyaev et al., 2000; Sunyaev et al., 2001). On that basis, we utilised residue accessibility to characterise and compare CPD to DAMUs. One would expect that if CPD are less deleterious than DAMUs they will show a trend towards happening at more exposed positions. However, this trend may be obscured if sequence divergence between human and non-human proteins is large enough. This is due to the fact that accessibility states, as well as other environment-dependent properties, may vary substantially below 60 % identity between homologs (Russell & Barton, 1994). We thus utilised the latter value as a threshold to divide our CPD dataset in two, according to the sequence identity between the human/non-human proteins.

Our results show (Figure 3) that when identity between homologs is above the 60 % threshold, CPD are biassed towards more exposed locations. This would be in accordance with the idea that preservation of the local environment in both the human and non-human proteins will introduce similar restraints on the nature of “acceptable” amino acids at that position. In this situation, either a compensatory mutation happens that will recover the damage caused by the CPD, or CPD will tend to happen at more resilient locations than DAMUs. Conversely, for sequence identities below 60 % CPD have a distribution closer to that of DAMUs. This would be in accordance with the fact that the local environment in the non-human structure has diverged substantially and the above mentioned restraints become weaker. A simple instance of the latter corresponds

to the case when the accessibility at the CPD location in the non-human protein is larger than for the human protein, as a result of sequence divergence. This accessibility increase will help accommodating more easily damaging mutations, in accordance with site-directed mutagenesis data (Matthews, 1993). In Table 1 we provide a list of these cases found in our dataset. An interesting case corresponds to the mutation H->R happening at position 43 in human superoxide dismutase, a mutation associated with disease. The arginine residue is also found at an equivalent position in yeast superoxide dismutase, as shown by the Pfam (Bateman et al., 2002) multiple sequence alignment and by the structural alignment of both structures (confirmed utilising three different structure comparison methods: MAMMOTH (Ortiz et al., 2002), CE (Shindyalov & Bourne, 1998) and LGA (Zemla, 2003)). Because position 43 in the human protein is essentially buried, with only ** Å of accessible surface area, destabilization by charge burial constitutes a plausible explanation for the damage caused by the mutation to arginine (Matthews, 1993; Matthews, 1995). Stability computations performed with the program PopMusic (Gilis & Rooman, 2000), indicate that this may be the case, as they show that the H43R mutation is associated to a stability decrease above 2.0 Kcal/mol. What has happened then in the yeast protein, to allow the presence of the arginine residue ? In the first place, solvent accessibility at this location has increased, ** Å, this is confirmed by the arginine binding of a phosphate from the crystallization solution. While this could be attributed to structure rearrangements induced by one or more compensatory amino acid substitutions, a simple visual inspection of the structure suggest that the most probable explanation is the presence of an amino acid insertion in the yeast sequence (Figure 4). This insertion, that happens in the loop linking beta-strands **, is spatially very close to position 43 and has probably given the arginine

residue enough room to accommodate at that position.

Amino acid level properties of CPD

Finally, we explored whether CPD correspond to intrinsically less deleterious than pathological mutations, that is, whether the change in amino acid properties associated to the mutation from human wild-type residue to the DAMU is less extreme, on the average. This is motivated by the fact that different studies in the characterization of DAMU (Chasman & Adams, 2001; Ferrer-Costa et al., 2002; Sunyaev et al., 2000) have shown that the intrinsic nature of the mutation can also contribute to define the pathological character of a mutation. For example, amino acid changes involving a large volume change are more likely to be damaging. Therefore, while compensatory mutations can clearly help to accomodate CPD, the nature of the latter may also play an important role. To measure this effect we have utilised three different properties that can be used as descriptors of the physico-chemical properties of the amino acids: volume, hydrophobic character, and mutation matrix elements (see *Methods*), and have shown their value in discriminating DAMUs from NEMUs (Ferrer-Costa et al., 2002).

In figures 5A-5C we show the distribution of the three amino acid descriptors chosen –hydrophobicity and volume indexes, and blosum62 matrix elements (see *Methods*), for the three mutation kinds: CPD, DAMUs and NEMUs. The latter two are displayed for comparison purposes. In general, we find that CPD distributions are intermediate between those from DAMUs and NEMUs. For example, in the case of the hydrophobic parameter, we see that there is a percentage of CPD that adopt more extreme values similar to those for DAMUs, e.g. when $\Delta\Delta G_{trf}$ = 1 Kcal/mol, the CPD and DAMUs frequencies are 15 % and 16 % respectively, while the percentage of

NEMUs with this value is 10 %. Conversely, for $\Delta\Delta G_{trf}$ values around 0, CPD are closer to NEMUs than to DAMUs (Figure 5A). A similar situation is found for size changes (Figure 5B), with some CPD having ΔV values closer to those of either NEMUs or DAMUs.

For the Blosum62 matrix elements we find again that CPD display an intermediate behaviour between DAMUs and NEMUs (Figure 5C). For example, when the matrix element is equal to 1, the corresponding frequencies for DAMUs, NEMUs and CPD are 16 %, 24 % and 31 %. On the contrary, for negative values of the Blosum62 elements, the corresponding frequencies are 20 %, 4 % and 9 %, respectively, with CPD closer to DAMUs than NEMUs. The fact that mutation matrix elements constitute summary measures of several properties reinforces the results obtained for the hydrophobic and size parameters. Overall, these results indicate that on the average CPD tend to be less deleterious than DAMUs, although more than NEMUs. This result would support the hypothesis formulated by Gao & Zhang (Gao & Zhang, 2003), 2003, according to which CPD, or at least a fraction of them, may correspond to slightly deleterious mutations.

CONCLUSIONS

In this article we review the properties of CPD, and of their protein locations the mechanisms, to understand how disease-associated amino acids in human can appear in other species as wild-type residues. Previous work, has mainly focused in the existence of compensatory mutations (Kondrashov et al., 2002), here we study whether either by their location or intrinsic properties, CPD could correspond to slightly deleterious mutations, as discussed by Gao & Zhang (Gao & Zhang, 2003). Our results show that

this is probably the case for a fraction of CPD, although they may also tend to happen at more resilient locations than the average DAMU.

ACKNOWLEDGEMENTS

The authors acknowledge the finantial support provided by the Fundación Areces. The work has been supported by the Spanish Ministry of Education and Science (BIO2002-06848, GEN2001-4758-C07-07 and BIO2003-09327, GENOM STRUCTURAL Project).

REFERENCES

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

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.

Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Howe, K. L. & Sonnhammer, E. L. (2000). The Pfam protein families database. *Nucleic Acids Res* 28, 263-6.

Chasman, D. & Adams, R. M. (2001). Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. *J Mol Biol* 307, 683-706.

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.

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

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

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

Gilis, D. & Rooman, M. (2000). PoPMuSiC, an algorithm for predicting protein mutant stability changes: application to prion proteins. *Protein Eng* 13, 849-56.

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

Huang, H., Winter, E. E., Wang, H., Weinstock, K. G., Xing, H., Goodstadt, L., Stenson, P. D., Cooper, D. N., Smith, D., Alba, M. M., Ponting, C. P. & Fechtel, K. (2004). Evolutionary conservation and selection of human disease gene orthologs in the rat and mouse genomes. *Genome Biol* 5, R47.

Hubbard, S. J. & Thornton, J. M. (1993). 'NACCESS', Computer Program,. Department

of Biochemistry and Molecular Biology, University College London, London.

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

Kulathinal, R. J., Bettencourt, B. R. & Hartl, D. L. (2004). Compensated deleterious mutations in insect genomes. *Science* 306, 1553-4.

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

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

Ortiz, A. R., Strauss, C. E. & Olmea, O. (2002). MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci* 11, 2606-21.

Pavlicek, A., Noskov, V. N., Kouprina, N., Barrett, J. C., Jurka, J. & Larionov, V. (2004). Evolution of the tumor suppressor BRCA1 locus in primates: implications for cancer predisposition. *Hum Mol Genet* 13, 2737-51.

Russell, R. B. & Barton, G. J. (1994). Structural features can be unconserved in proteins with similar folds. An analysis of side-chain to side-chain contacts secondary structure and accessibility. *J Mol Biol* 244, 332-50.

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

Shindyalov, I. N. & Bourne, P. E. (1998). Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng* 11, 739-47.

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.

Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., Antonarakis, S. E., Attwood, J., Baertsch, R., Bailey, J., Barlow, K., Beck, S., Berry, E., Birren, B., Bloom, T., Bork, P., Botcherby, M., Bray, N., Brent, M. R., Brown, D. G., Brown, S. D., Bult, C., Burton, J., Butler, J., Campbell, R. D., Carninci, P., Cawley, S., Chiaromonte, F., Chinwalla, A. T., Church, D. M., Clamp, M., Clee, C., Collins, F. S., Cook, L. L., Copley, R. R., Coulson, A., Couronne, O., Cuff, J., Curwen, V., Cutts, T., Daly, M., David, R., Davies, J., Delehaunty, K. D., Deri, J., Dermitzakis, E. T., Dewey, C., Dickens, N. J., Diekhans, M., Dodge, S., Dubchak, I., Dunn, D. M., Eddy, S. R., Elnitski, L., Emes, R. D.,

Eswara, P., Eyras, E., Felsenfeld, A., Fewell, G. A., Flicek, P., Foley, K., Frankel, W. N., Fulton, L. A., Fulton, R. S., Furey, T. S., Gage, D., Gibbs, R. A., Glusman, G., Gnerre, S., Goldman, N., Goodstadt, L., Grafham, D., Graves, T. A., Green, E. D., Gregory, S., Guigo, R., Guyer, M., Hardison, R. C., Haussler, D., Hayashizaki, Y., Hillier, L. W., Hinrichs, A., Hlavina, W., Holzer, T., Hsu, F., Hua, A., Hubbard, T., Hunt, A., Jackson, I., Jaffe, D. B., Johnson, L. S., Jones, M., Jones, T. A., Joy, A., Kamal, M., Karlsson, E. K., et al. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature* 420, 520-62.

Zemla, A. (2003). LGA: A method for finding 3D similarities in protein structures. *Nucleic Acids Res* 31, 3370-4.

Zuckerkandl, E. & Pauling, L. (1962). Molecular disease, evolution and genetic heterogeneity. In *Horizons in Biochemisty* (B., M. & Pullman, B., eds.), pp. 189-225. Academic Press, London.

LEGENDS TO FIGURES

Figure 1. Distribution of sequence identities between: (i) human proteins and different animal model SwissProt (Apweiler et al., 2004) homologs; (ii) human proteins and non-human homologs, where the latter carry in, at least one position, one amino that would be pathological in human (BLACK).

Figure 2. Distribution of sequence identity for neighbours of the CPD. Two kinds of neighbours were considered: (A) Local-in-sequence identity, where we only consider sequence neighbours of the CPD site (see *Methods*); (B) Local-in-space identity distribution, where we only consider spatial neighbors of the CPD site (see Methods). In both cases we only consider pairs human/non-human proteins with more than 90 % global sequence identity. For comparison purposes, we plot both observed distribution (dark-grey) and an expected/random distribution (light-grey), see *Methods*.

Figure 3. Residue accessibility distribution for CPD and DAMUs. The CPD sample is split in two, as explained in the text, human/non-human pairs with sequence identity higher (dark-grey) or lower (light-grey) than 60 %. Shown in white is the DAMUs distribution.

Figure 4. Example of CPD happening in superoxide dismutase. The disease-associated amino acid happens at location 43 in the human protein, and appears as wild-type in the yeast protein (also position 43). The structural alignment between both proteins (PDB codes: 1spd and 1f1g for the human and yeast proteins, respectively) is displayed in the figure. Shown in lilac and blue are the human and yeast protein, respectively.

Figure 5. Distribution of intrinsic properties for CPD: (A) Hydrophobicity distribution; (B) Volume distribution; and (C) Blosum62 matrix elements distribution. See Methods for the description of the different indexes. Three distributions are shown in each figure: CPD (dark-grey), DAMUs (ligh-grey) and NEMUs (white).

Table 1. Examples of CPD where there is an increase in the residue accessibility between the human and non-human protein, at the CPD location (see text).

FIGURE 1

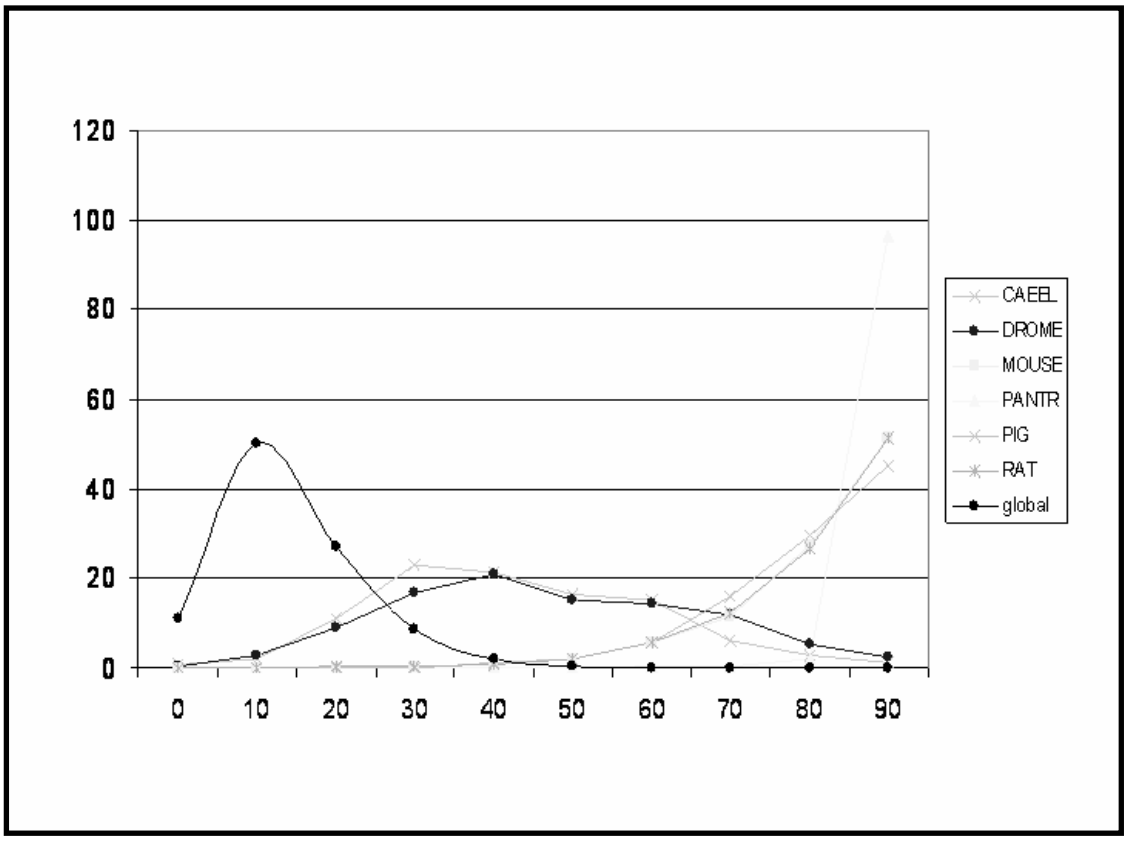

120
100
80
60
40
20
0
0
10
20
30
40
50
60
70
80
90
CAEEL
DROME
MOUSE
PANTR
PIG
RAT
global

FIGURE 2A

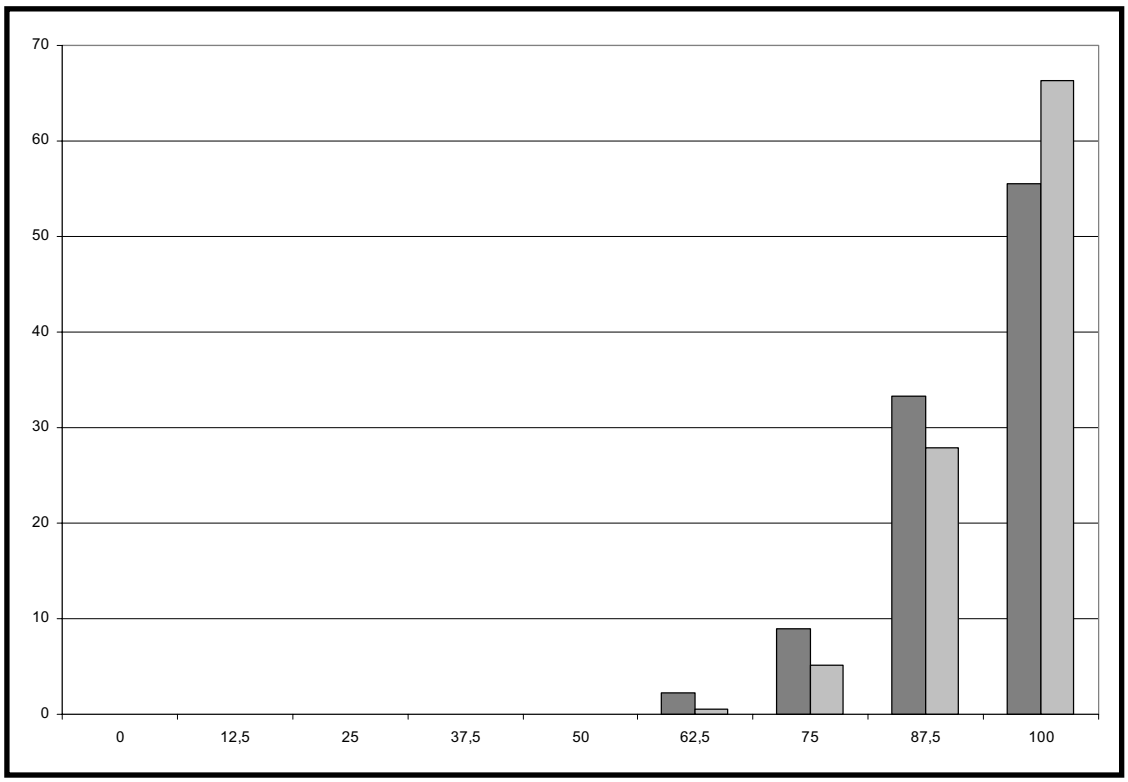


FIGURE 2B

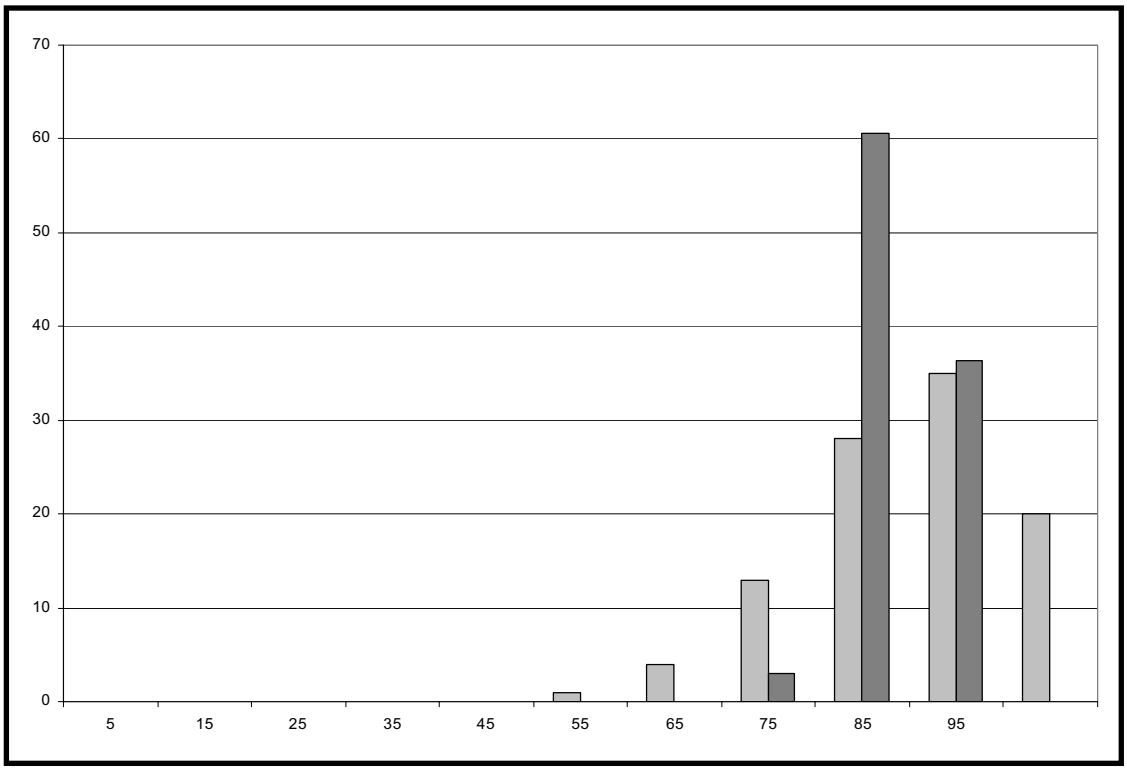


FIGURE 3

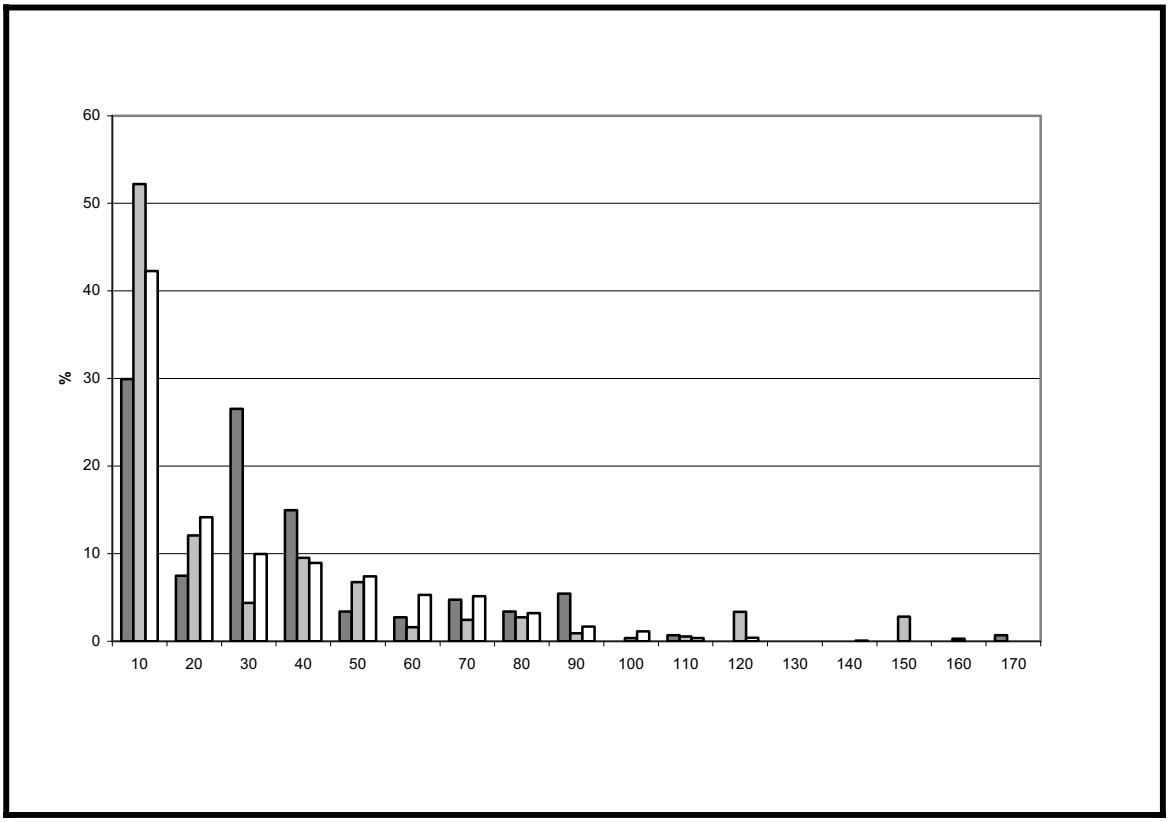

60
50
40
30
20
10
0
%
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170

FIGURE 4

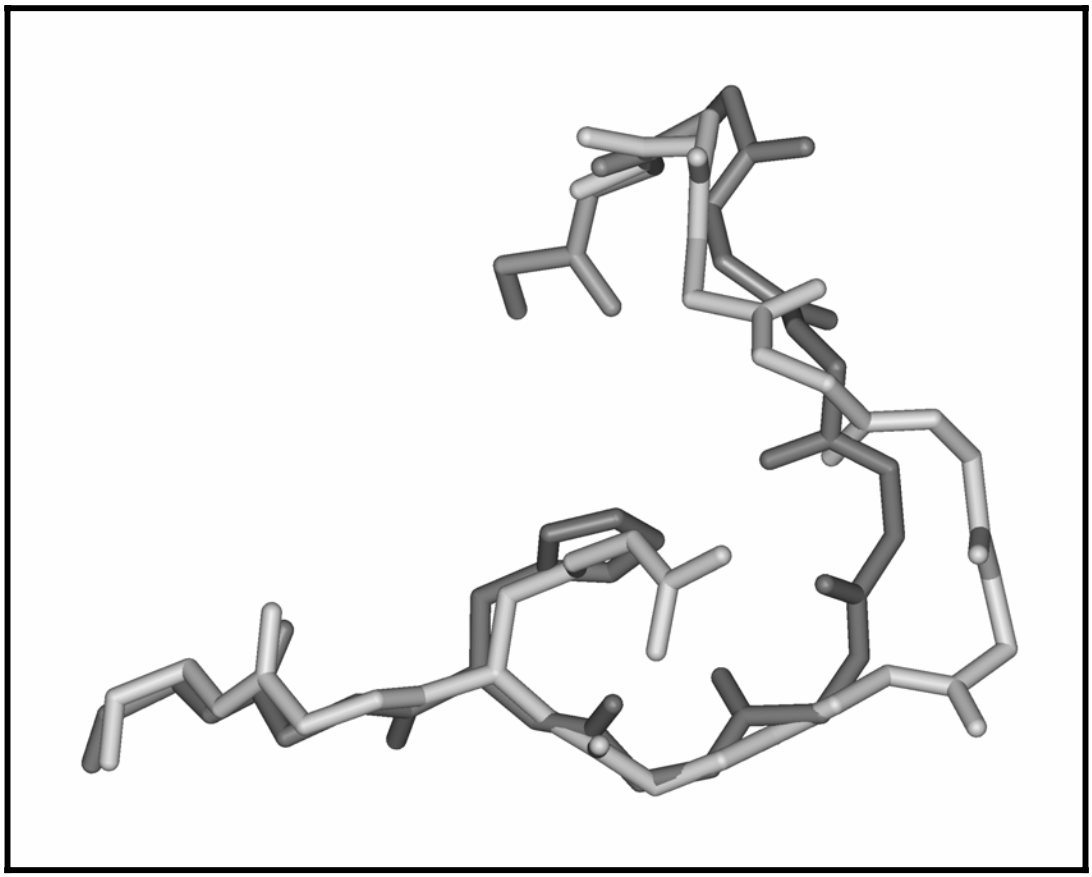


FIGURE 5A

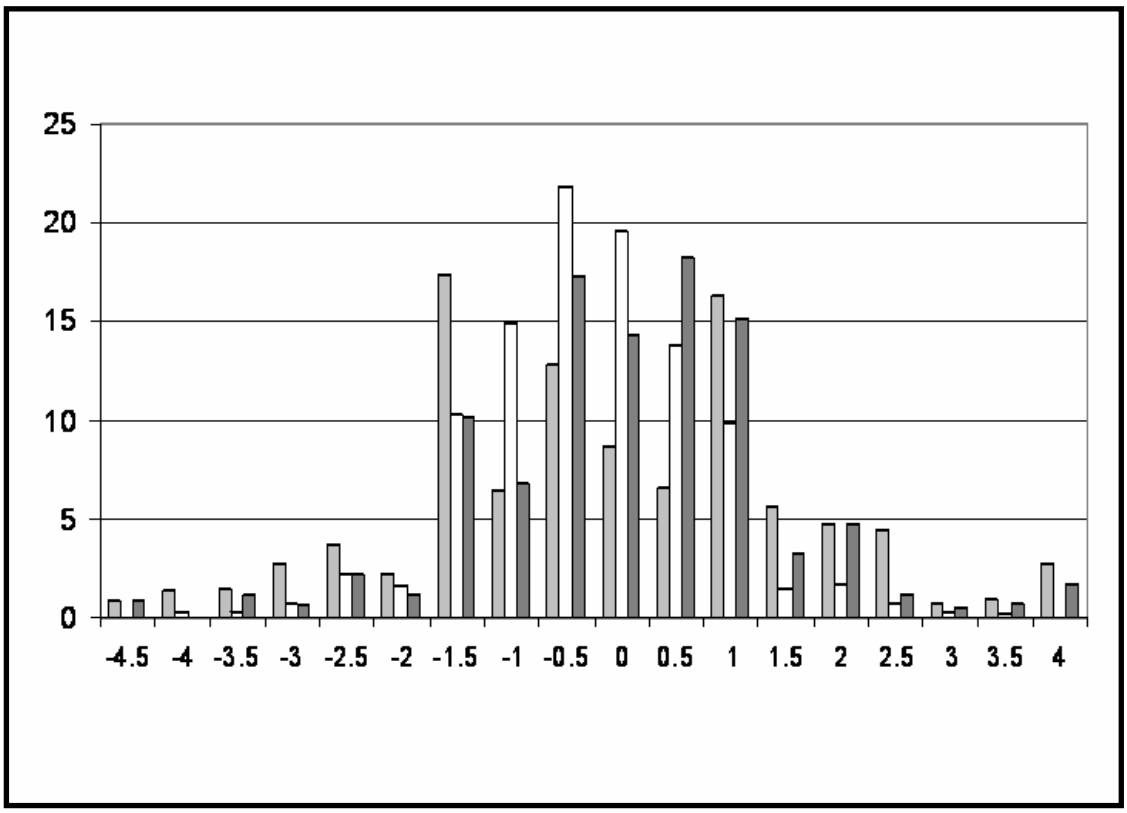


FIGURE 5B

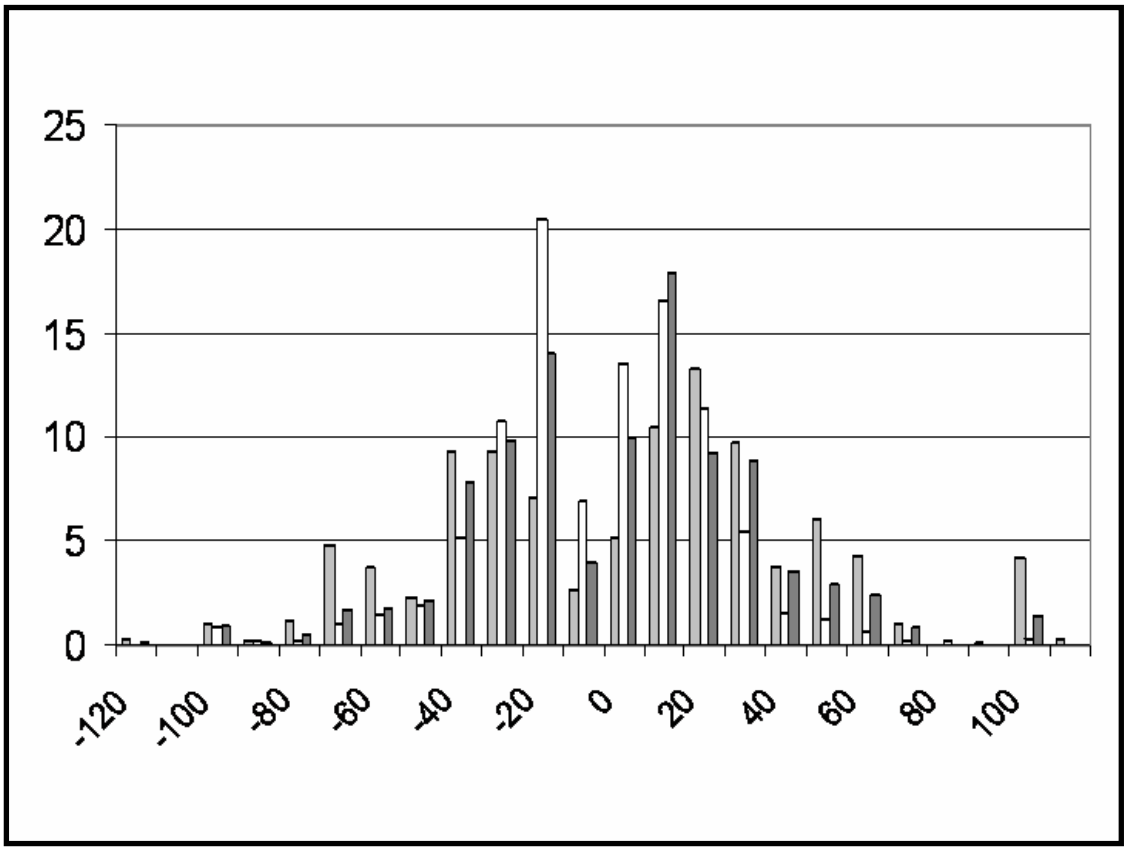
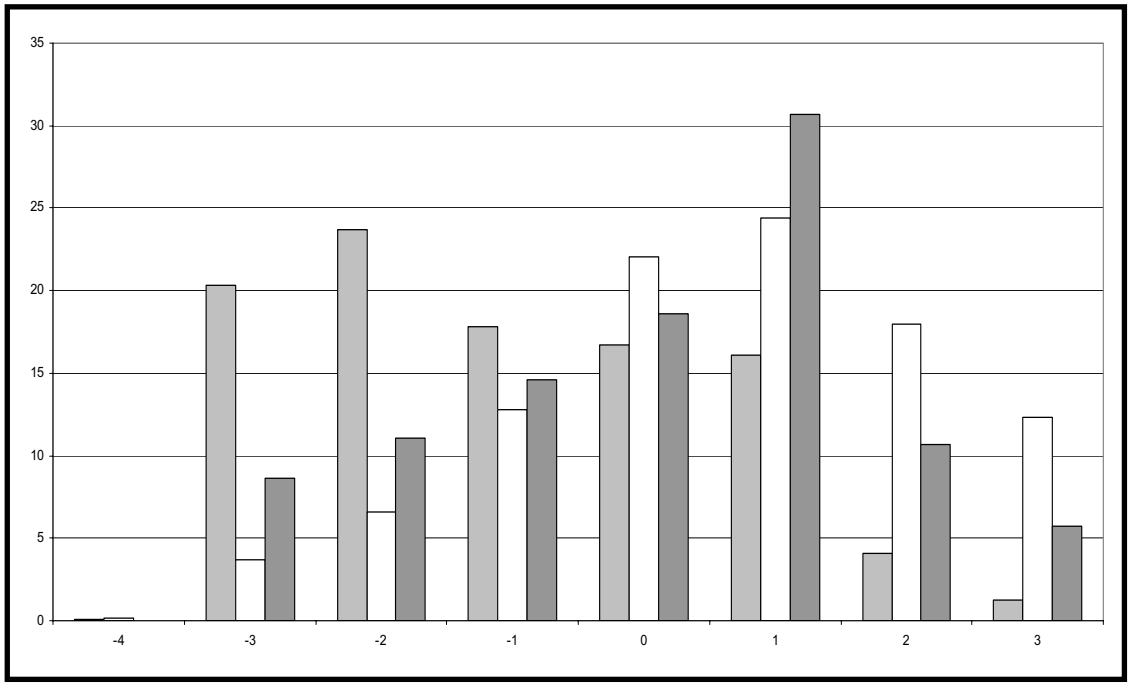


FIGURE 5C



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

IV. BIBLIOGRAFIA DEL CAPÍTOL

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

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.

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.

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

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

Huang, H., Winter, E. E., Wang, H., Weinstock, K. G., Xing, H., Goodstadt, L., Stenson, P. D., Cooper, D. N., Smith, D., Alba, M. M., Ponting, C. P. & Fechtel, K. (2004). Evolutionary conservation and selection of human disease gene orthologs in the rat and mouse genomes. *Genome Biol* 5, R47.

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

Kulathinal, R. J., Bettencourt, B. R. & Hartl, D. L. (2004). Compensated deleterious mutations in insect genomes. *Science* 306, 1553-4.

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

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

Ortiz, A. R., Strauss, C. E. & Olmea, O. (2002). MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci* 11, 2606-21.

Pavlicek, A., Noskov, V. N., Kouprina, N., Barrett, J. C., Jurka, J. & Larionov, V.

(2004). Evolution of the tumor suppressor BRCA1 locus in primates: implications for cancer predisposition. *Hum Mol Genet* 13, 2737-51.

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.

Zuckerkandl, E. & Pauling, L. (1962). Molecular disease, evolution and genetic heterogeneity. In *Horizons in Biochemisty* (B., M. & Pullman, B., eds.), pp. 189-225. Academic Press, London.