Compensated pathogenic deviations: analysis of structural effects

Anja Barešić, Lisa E.M. McMillan[1], Hubert H. Rogers[2], Jacob M. Hurst, Andrew C.R. Martin[*]

Institute of Structural and Molecular Biology, Darwin Building, University College London, Gower Street, London WC1E 6BT, United Kingdom

Abstract

Pathogenic deviations (PDs) in humans are disease-causing missense mutations. However, in some cases, these disease-associated residues occur as the wild-type residues in functionally equivalent proteins in other species and these cases are termed 'Compensated Pathogenic Deviations' (CPDs). The lack of pathogenicity in a non-human protein is presumed to be explained in most cases by the presence of compensatory mutations, most commonly within the same protein. Identifying structural features of CPDs, and detecting specific compensatory events, will help us to understand traversal along fitness landscape valleys in protein evolution.

We divided mutations listed in the OMIM database into PD and CPD datasets and performed two independent analyses: (i) we searched for potential compensatory mutations spatially close to the CPDs and (ii) using our SAAPdb database, we examined likely structural effects to try to explain why mutations are pathogenic, comparing PDs and CPDs. Our datasets were obtained from a set of 245 human proteins of known structure and contained a total of 2328 mutations of which 453 (from 85 structures) were seen to be compensated in at least one functionally equivalent protein in another (non-human) species.

Structural analysis results confirm previous findings that CPDs are, on average, 'milder' in their likely structural effects than uncompensated PDs and tend to be on the protein surface. We also showed that the residues surrounding the CPD residue in the folded protein are more often mutated than the residues surrounding an uncompensated mutation, supporting the hypothesis that compensation is largely a result of structurally local mutations.

Key words: disease mutations, structural analysis of mutations, sequence analysis of mutations, evolution, epistatic selection

1. Introduction

1.1. Compensated pathogenic deviations

Recent work on protein evolution and protein structure has focused on the phenomenon of Compensated Pathogenic Deviations (CPDs)[1,2,3,4,5] i.e. disease-associated mutations in a protein of one species (usually human), which occur as the wild-type in a 'functionally equivalent protein' (FEP) of another species – we define FEPs, and their potential differences from orthologues, in our recent paper[6]. The pathogenic effect of a CPD is assumed to be neutralized in the FEP by a compensatory mutation, usually within the same protein sequence[1].

From an evolutionary point of view, CPDs allow the crossing of unfit valleys between two known fit sequences by introducing protein sequence substitutions. Hence, the study of CPDs offers a unique and invaluable tool to access information on protein evolution and epistatic selection[1].

1.2. Evolution of CPDs

Almost all possible genetic sequences are unfit, so for a protein to evolve over time, only a discrete series of rare, fit sequences may be used as steps in the evolutionary journey[1]. Data on fitness landscapes are limited to the genetic sequences that are available: normally only wild-type sequences and disease-associated mutations, a representative subset of the latter being available from Online Mendelian Inheritance in Man (OMIM)[7,8] and other, locus-specific mutation databases. One of the ways to traverse between adjacent peaks in the fitness landscape is through CPDs: individually pathogenic mutations become fixed in the population through epistatic selection with compensatory mutations. (Epistasis is the dependency of the effect of a mutation on the genetic background in which it occurs[9].) Hence, sequence data on disease-causing mutations, and on CPDs, allow us to study valleys in the fitness landscape separating peaks of fit genotypes.

Previous studies by Kondrashov *et al.*[1] in the human genome and Kulathinal *et al.*[2] in the *Drosophila* genome have shown that 10% of deviations from a human/*Drosophila* wild-type sequence to a different residue in an orthologous sequence are of a residue type which causes disease in humans/*Drosophila*. In other words, 10% of substitutions are CPDs. This ratio of CPDs per total residue substitutions, is approximately stable over a wide range of human and some *Drosophila* FEPs, and is independent of phylogenetic distance and population size[1,2]. Hence, the stability of this ratio suggests frequent and regular evolution of compensatory mutations[1].

Focusing on co-occurrence of CPDs and compensatory mutations, DePristo *et al.*[3] proposed two hypotheses of CPD evolution based on models of biophysical properties. In the first

[*] Corresponding author

[1] Present address: Computing Science, Sir Alwyn Williams Building, Lilybank Gardens, University of Glasgow, Glasgow, G12 8QQ, Scotland

[2] Present address: Michael Smith Building, University of Manchester, Oxford Road, Manchester, M13 9PT

scenario, a compensatory mutation *C* is phenotypically neutral and stable, thus fixing itself quickly in the population. A pathogenic mutation *P* is unstable, and can become fixed only if it occurs *after* the compensatory mutation *C*, resulting in a CPD (the *P*–*C* pair) which has higher fitness owing to epistasis. In the second model, both *P* and *C* are individually deleterious, but together have a neutral effect, giving rise to a fitness valley. It is known that small frequencies of low-fitness mutations exist in large populations, so it is possible for the *P*–*C* genotype to fix itself within the population, while neither of the deleterious intermediates is fixed on its own. A less likely, but possible, scenario is that both *C* and *P* occur simultaneously. Additionally, Cowperthwaite *et al.*[4] propose a mechanism of compensation occurring after the appearance of the deleterious mutation. Their observations are based on RNA molecules' evolution *in silico*, and they show that, provided the mutation rate is sufficiently high, epistatic selection with compensatory mutations is the most prevalent mechanism of otherwise deleterious mutation fixation.

1.3. *Structural features of CPDs*

In a recent study, Ferrer-Costa *et al.*[5] demonstrated that both the structural environment and the nature of the substitution play an important role for the development of compensatory mutations facilitating a CPD. Their results show statistically significant differences in the solvent accessibility of mutated CPD residues as well as intrinsic properties of the mutation (change in amino acid volume, hydrophobicity and BLOSUM62 scores[10]) when compared with 'pathogenic deviations' (PDs). They suggested (i) that mutations to residues making a large number of contacts are more difficult to compensate than those making few contacts, and (ii) that CPDs are, on average, more conservative substitutions than PDs. We have built on this study to analyze a wide range of structural effects and their frequency of occurrence among compensated and uncompensated disease-associated mutations. We have also extended the analysis of the structural environment of disease-causing mutations by calculating the mutation rates among residues in close proximity to the pathogenic deviation.

Our data on the distribution of structural effects of CPDs in comparison with PDs provides an insight into what kinds of structural effects are easy, or more difficult, to neutralize through compensatory mutations. This may, in turn, help to shed light on the mechanisms of compensation, which are as yet poorly understood[1,2,3,5]. We analyzed local structural consequences of mutations on a large dataset of OMIM mutations, using methods of structural analysis previously developed in our group[11,12,13,14,15].

Thus, this paper sets out both to examine the location of compensatory mutations and the nature of pathogenic mutations which can be compensated.

2. Results and Discussion

2.1. *The CPD dataset*

2328 disease-causing mutations from OMIM[7,8] occurring in 245 human proteins were successfully mapped both to a residue in a UniProtKB/SwissProt[16] sequence and to a structure in the Protein Databank (PDB)[17]. Of these, 453 mutations were found as a native residue in at least one non-human aligned functionally equivalent protein sequence and annotated as CPDs.

Table 1 shows the numbers of CPDs analyzed in our study compared with those of Ferrer-Costa *et al.*[5] and Kondrashov *et al.*[1] These groups use the same definition of a CPD, but use different methods and datasets to identify CPDs. An important difference in our analysis is the use of functionally equivalent proteins (FEPs) rather than orthologues derived from Pfam[18] (as used by Ferrer-Costa), or from BLAST (as used by Kondrashov). Orthologues can diverge in function and, where they do, key functional residues will, by definition, be subject to mutation[6]. While the broader sets of sequences used in other work may lead to additional CPDs being identified, using our more restricted sets of FEPs obtained from our FOSTA database[6] ensures that this situation will not arise.

While Ferrer-Costa and colleagues identified a significantly larger set of 811 human proteins containing mutations (compared with our 245), many of these mapped only to sequence (Table 1), whereas our dataset includes only mutations mapped to structure. 35% of the larger (sequence-based) Ferrer-Costa disease-associated protein dataset contained at least one CPD location, while 29% of the smaller (structure-based) set in this study had compensated mutations (Table 1). In addition, they extracted mutation data from UniProtKB/SwissProt annotations, resulting in a different set of mutations from those we identified from OMIM. They only used protein structures for their relative accessibility analysis (24 proteins).

CPD detection by Kondrashov and colleagues was based on a small number of proteins reported to have large numbers of pathogenic deviations (at least 50 per protein). As a result, the percentage of human proteins containing a CPD is significantly higher than in the other two methods. Like Ferrer-Costa, most of their analysis was performed at the sequence level, with more detailed structural analysis, looking for potential compensatory mutations, being performed for just three proteins (β-hemoglobin, von Willebrand factor and transthery-tin) where structures are available for the human protein and for mammalian orthologues. Thus, to our knowledge, our results using 85 structures represent the largest structural analysis of CPDs.

Table 2 summarizes the general trends observed in the data. We also evaluated the diversity of the FEP families in which PDs and CPDs were obtained as shown in Figure 1. This shows first that CPDs are fairly evenly spread across families with different levels of diversity. Second, while compensatory events are more common in more diverse familes (i.e. those which, on average, contain more distantly related members), they occur even in families which show very low diversity.

Table 1: Analysis of CPDs detected in different studies.

	Present study	Ferrer-Costa[5]	Kondrashov[1]
Human proteins searched	245	811	32
Human mutations identified	2328	9334	4272
Total-observed CPDs[a]	3218	140 (30465)	608
Distinct CPDs[b]	453	52 (1658)	N/A[c]
Human proteins with one or more corresponding CPDs	85	24 (287)	3[d] (24)

In the present study, only sequences which could be mapped to structure were used. For other studies, where a distinction can be made, the main number refers to the number of structures, while the number in parentheses refers to sequence analysis. [a]A given disease-causing mutation in a human sequence may match the native residue in several different functionally equivalent proteins from other species. Thus the number of CPDs observed is greater than the number of human mutations. [b]The number of human disease-causing mutations having one or more CPD-containing functionally equivalent proteins. [c]Data not available in the Kondrashov paper. [d]While structures may have been available for more human proteins, the authors only analyzed those proteins where structures were also available for multiple mammalian orthologues.

Table 2: Summary data for PDs and CPDs

Characteristic		
Accessibility[a]	PDs: $\bar{x}$ = 26.9, σ = 27.2	CPDs: $\bar{x}$ = 43.4, σ = 28.0
Redundancy[b]	$\bar{x}$ = 6.79%, σ = 4.92, min= 0.00, max=99.32	

[a]Relative solvent accessibility was calculated using a local implementation of the Lee and Richards algorithm[19]. [b]Redundancy was calculated as the mean and standard deviation of the pairwise identity of the 85 human sequences used in this analysis; minimum and maximum identity are also provided – the most different pair being PTEN_HUMAN and RTN4R_HUMAN and the most similar pair being HBG1_HUMAN and HBG2_HUMAN.

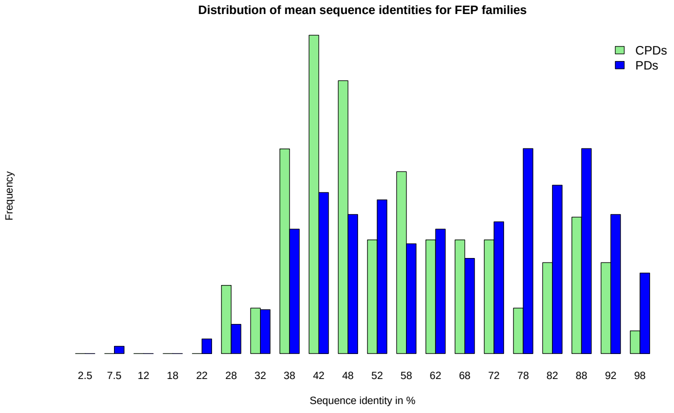

Figure 1: Diversity of FEP families containing PDs (264) and CPDs (85). Note that some families may occur in both datasets. The histogram is normalized such that the total height of the bars is the same between the two sets. Diversity was calculated as the mean pairwise sequence identity within the family.

2.2. *Potential local compensatory mutations*

This analysis calculated and compared the frequencies of mutations occurring in the residues surrounding a CPD or a PD in the structure. In common with Kondrashov *et al.*[1], we hypothesized that compensatory mutations, neutralizing a CPD's pathogenicity, are likely to be physically close to a CPD and involved in short-range interactions. Figure 2a shows the distribution of sequence variability in residues surrounding a CPD, compared with PDs. C/T ratios (where C was the number of local (potentially compensatory) mutations and T was the total number of 'in range' columns (in the alignment) checked for that sequence — i.e. the fraction of in-range residues that are mutated, see Materials and Methods) of PDs were taken in order to control for sequence variability. Owing to the great number of points on the graph, and in order to see if there is any major difference between the two datasets, we averaged the C/T ratio for every dataset and sequence identity, as shown in Figure 2b. Restrained linear regression was performed on the full datasets to obtain lines of best fit (the restraint being the biologically obvious condition that both lines have to pass through 0 mutations when the sequence identity is 100%). The line equations show a significant increase in slope for the CPD dataset (Z-statistic=7.860, with $p < 0.05$). This increase in the average number of diverged residues in the structural neighbourhood of CPDs strongly supports the hypothesis that compensation is commonly a local effect, as previously suggested by Kondrashov *et al.*[1]

For the CPDs in Figure 2b, the best-fit line has a slope of -1.007 indicating that CPDs reflect a set of random mutational events occurring during evolution of the environment in which the CPD occurs. In contrast, PDs occur at sites where conservation is higher (for structural or functional reasons) and thus compensation by random mutational drift in the surroundings is less likely to occur.

In addition we separated the data into buried ($<$ 10% relative accessibility) and exposed mutated residues (PD or CPD) and repeated the analysis shown in Figure 2 on the two sets separately. The lines of best fit were almost indistinguishable from

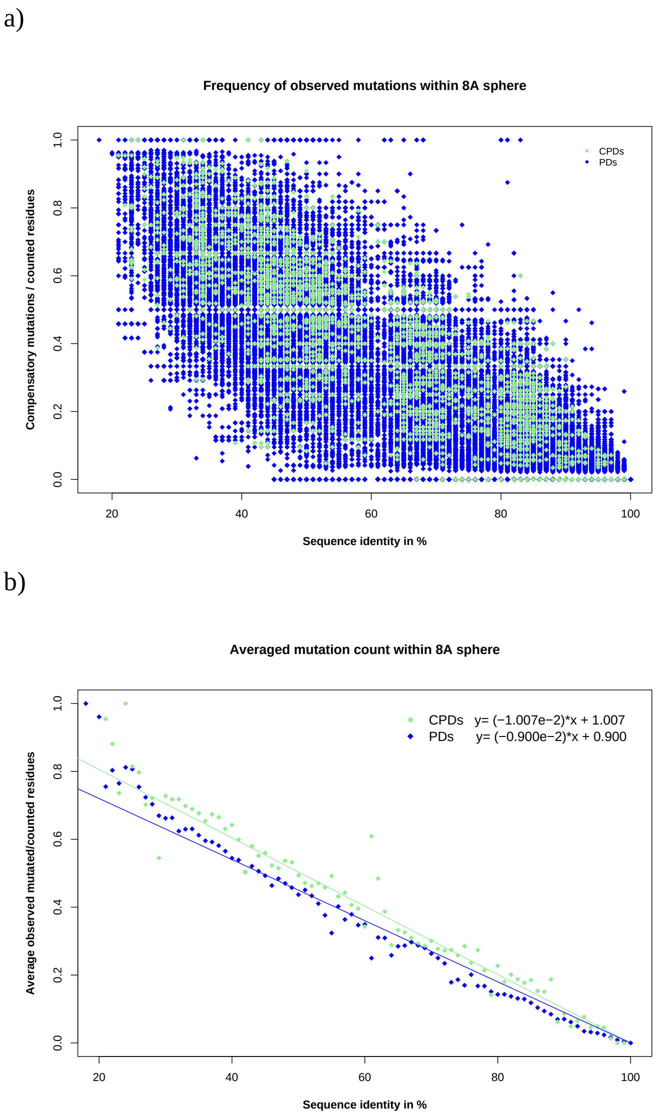

Figure 2: Dependency of the local mutation ratio on sequence identity. **a)** The C/T ratio for residues within an 8Å sphere of each mutation is plotted against sequence identity for both CPDs and PDs. See Materials & Methods. **b)** The line of best fit, obtained by linear regression with a (100, 0) constraint for both complete datasets (i.e. the data shown in a): 3138 datapoints for CPDs and 74429 datapoints for PDs) is shown together with the average C/T ratio for each 1% sequence identity bin to illustrate the trends in the data.

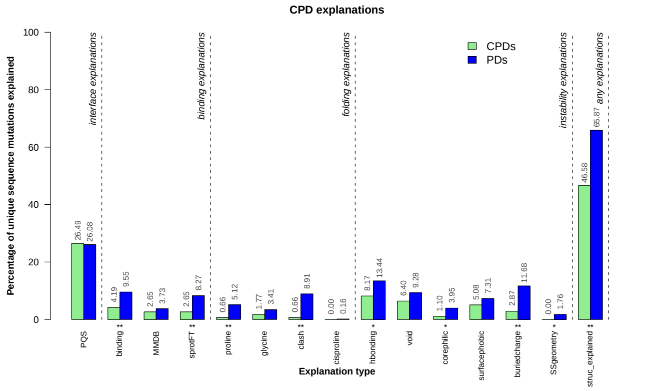

Figure 3: Relative frequencies (F_{cat}, see Materials and Methods) of predicted structural effects for CPDs and PDs. (See Table 3 for the meanings of the effect categories.) Values are indicated at the top of each bar. Significantly different bars (Fisher's exact test, see Table 3) after application of the Bonferroni correction for multiple testing are indicated (* $p < 0.05$, ** $p < 0.01$). The 'struc_explained' bar is a summary representing explanation by any of the other structural tests shown in the figure. In this case no correction was applied giving $p = 6.71 \times 10^{-14}$

the equivalent lines in the full datasets (data not shown), the only notable difference being that CPDs in the core showed a slightly greater slope of −1.025 suggesting that, when they occur, they are accompanied by a somewhat higher local mutation rate. CPDs on the surface showed a slope of −1.001 suggesting that compensation is indeed the result of random mutations.

2.3. *Mutation structural analysis*

Fractions of PDs and CPDs for which structural effects have been identified in SAAPdb are shown in Figure 3, divided into categories of likely structural effects. Analysis of relative frequencies in thirteen categories covered four classes of disrupting effects: protein interface, binding properties, protein folding and stability. These categories are summarized in Table 3 and have been explained in detail by Hurst *et al.*[15] Differences between the two datasets give an insight into which types of structural disruptions are more likely to be compensated, showing that the compensation of pathogenic mutations is highly dependent on the nature of the mutation's effect on the structure. We will now briefly discuss the results for each of the four general classes defined above. The examples shown were selected at random as examples where a simple 1-amino-acid compensatory event appears to be important. In other cases, a number of compensatory events may have an additive effect.

2.3.1. *Interface disrupting effects*

We define interface residues as surface residues in the monomer which undergo a change in relative accessibility of ≥10% on complex formation. Solvent accessibility is calculated using a local implementation of the Lee and Richards algorithm[19]. We find that 26.5% of CPDs and 26.1% of PDs occur in interface residues found in PQS files[20]. This is the only structural category for which the frequency of CPDs is the same, or greater than, the frequency of PDs. This confirms recent observations that CPDs are often found in residues having

Table 3: Structural effect categories.

Structural category	Effect of mutation	p-value	mc-value
PQS[a]	Affecting residues in the interface with a different protein chain or ligand identified from a PQS file (and therefore more likely to reflect biologically relevant interactions) by a change in solvent accessibility.	> 1	0.81
binding[b]	Affecting residues involved in specific binding interactions (a hydrogen bond, salt bridge, or packing interaction) with a different protein chain or ligand.	1.5×10^{-3}	0.00
MMDB[b]	Affecting residues in contact with a ligand, according to the MMDB database.	> 1	0.31
sprotFT[b]	Residues annotated in SwissProt Feature records as having a functional significance.	9.04×10^{-5}	0.00
proline[c]	Mutations to proline where the backbone angles are restrictive.	2.20×10^{-5}	0.00
glycine[c]	Mutations from glycine where the backbone angles are restrictive.	9.87×10^{-1}	0.07
clash[c]	Causing a clash between atomic radii of the neighbouring residues.	7.95×10^{-12}	0.00
cisproline[c]	Mutations from a cis-proline.	> 1	0.36
hbonding[d]	Causing the disruption of hydrogen bonds between residues.	2.79×10^{-2}	0.001
void[d]	Causing an internal void ≥ 275Å^3 to open in the protein owing to the substitution with a smaller residue.	7.22×10^{-1}	0.048
corephilic[d]	Introducing a hydrophilic residue in the protein core.	1.85×10^{-2}	0.083
surfacephobic[d]	Introducing a hydrophobic residue on the protein surface.	> 1	0.088
buriedcharge[d]	Introducing an unsatisfied charge in the protein core owing to the substitution with, or of, a charged residue.	6.47×10^{-9}	0.00
SSgeometry[d]	Causing the disruption of a disulphide bridge.	1.83×10^{-2}	0.0006

The structural explanation categories are described in detail by Hurst *et al.*[15] [a]Interface explanations; [b]Functional explanations; [c]Folding (fold-preventing) explanations; [d]Instability (destabilizing) explanations. p-values are obtained from a Fisher's exact test ($d.f. = 1$) and then multiplied by 14 to apply a Bonferroni Correction to the p-values to allow them to be compared with conventional α values of 0.05 and 0.01. The mc-value shows the result of a Monte Carlo simulation and is the fraction of random divisions of the data which obtain the observed p-value or better (see text).

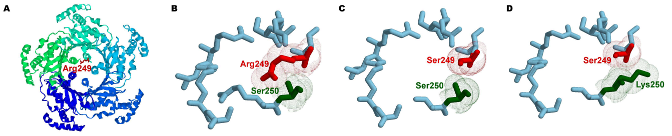

Figure 4: Potential compensation of a mutation affecting an interface residue. **a)** The position of Arg249Ser is shown on the human GTP cyclohydrolase pentamer structure, PDB:1fb1. This CPD occurs at an *interface* in the pentamer and causes dopa-responsive dystonia. **b)** Detail of Arg249 and its interaction with Ser250 from a neighbouring monomer. Multiple non-bond interactions between Arg249 and Ser250 contribute to pentamer stability. **c)** The Arg249Ser mutation causes the loss of function in GCH1_HUMAN by losing multiple non-bonded interactions (modelled structure shown) and hence destabilizing its structure. **d)** The *Rickettsia bellii* FEP has compensated for the Ser249 lost contacts by introducing Lys250 (modelled structure).

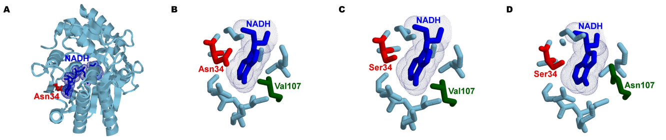

Figure 5: Potential compensation of a mutation affecting a binding residue. **a)** Asn34Ser position is shown on the human UDP-glucose 4-epimerase structure, PDB:1ek6. This CPD occurs in a *binding* site and in a *PQS* interface and causes epimerase-deficiency galactosemia. **b)** Detail of Asn34 and its interaction with NAD$^+$. **c)** The Asn34Ser mutation causes the loss of hydrogen bond with the exogenous NAD$^+$, needed for the normal function of the human protein (modelled structure). **d)** The *Streptococcus thermophilus* and *Streptococcus mutans* FEPs have compensated for the Ser34 by introducing Asn107, which in turn stabilizes protein-ligand interaction, shown on the modelled structure.

fewer intra-protein interactions[5] (and hence have fewer structural constraints) and may indicate that it is relatively easy to compensate for the deleterious effects of interface residues. An example of a compensated mutation in the protein interface is shown in Figure 4.

2.3.2. Mutations affecting binding

A significantly greater fraction of PDs than CPDs was assigned as making specific binding interactions (hydrogen bonds defined according to the rules of Baker and Hubbard[21], or non-bonded contacts) to a ligand or another protein chain (Figure 3, category 'binding'). Using data from the MMDBBIND database[22] to identify binding residues rather than the PDB data, also showed a greater fraction of PDs than CPDs, but the difference was not statistically significant.

It is not surprising that, owing to the specific properties required for H-bonds or interactions at interfaces, our results showed compensating for a mutation at a specific binding residue is usually difficult. An example of a compensated mutation at a binding residue is shown in Figure 5.

2.3.3. Folding disruption effects

This class of structural effects describes cases where the mutation is likely to prevent correct folding of the protein and is represented by (i) mutations from cis-proline, to proline and from glycine (where backbone torsion angles are unfavourable for the replacement residue), and (ii) introduction of a bulkier, clash-causing residue. In our analysis, mutations from cis-proline are very rare and are not considered further.

Mutations from another amino acid to proline are expected to be damaging to protein structure when the native residue has a backbone conformation disallowed by proline's cyclic sidechain. Our results show that such mutations occur significantly less frequently in the CPD set than the PD set indicating that compensation is difficult. An unusual example of neutralization of a mutation to proline is seen in antithrombin-III (ANT3, See Materials and Methods, Figure 10). In this example, compensation appears to be achieved by removing another nearby proline with both the compensated and compensatory mutations located in the same loop (PDB:2b5t chain I, structure not shown). In contrast, mutations from glycine (where

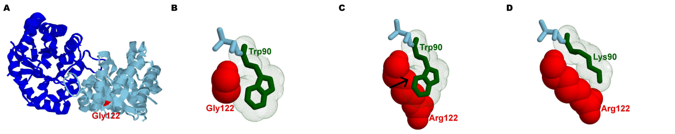

Figure 6: Potential compensation of a mutation affecting a folding residue. **a)** The position of Gly122Arg is shown on the human triosephosphate isomerase dimer structure, PDB:1wyi. This CPD causes a *clash* and a *buried charge*, and increases thermo-sensitivity of the human protein. **b)** Detailed position of Gly122 and Trp90. **c)** The Gly122Arg mutation causes atom clash, indicated by the arrow, between larger sidechain of Arg122 and native Trp90 (modelled structure). **d)** Substituting Trp90 with a smaller Lys compensates for the introduction of the Arg122 in several bacterial FEPs (*Aquifex aeolicus*, *Coxiella burnetii*, *Mycoplasma gallisepticum*, *Treponema pallidum*, *Xylella fastidiosa*, *Chromohalobacter salexigens*), shown on the modelled structure.

the glycine has a backbone conformation not accessible to other amino acids) showed no significant difference between PDs and CPDs.

Substitution to a clash-causing residue was extremely rare among CPDs compared with PDs. This is not surprising as compensating for a clashing residue would probably need several, chronologically earlier, cascading compensatory mutation events to create a void large enough to accommodate the clashing residue; such a void would itself be destabilizing. A rare example of a clash compensation is observed in human triosephosphate isomerase FEPs, as shown in Figure 6.

2.3.4. *Mutations affecting protein stability*

Mutations affecting protein stability introduce no physical barriers to *prevent* correct folding, but reduce the stability of the correctly folded form below that of unfolded or mis-folded states[15]. Disruption of hydrogen bonding, creation of voids, misplaced charges, hydrophilics, or hydrophobics, and disruption of disulphides all fall into this category. Such mutations may be temperature-sensitive (such as the Val143Ala mutation in p53[23]) and are the main category of interest in ‘rescuing’ protein function[24,25,26]. We observed very few cases of disruption of disulphides and this was not considered further.

Mutations that affect hydrogen-bonding were identified in SAAPdb according to the method of Cuff *et al.*[14] Considering the fact that hydrogen bonds have a strong effect on protein stability[14] and that precise geometries are involved, it is not surprising that mutations affecting hydrogen-bonding were found very commonly in both datasets. The high frequencies in both datasets, 8.17% of CPDs and 13.44% of PDs, indicated a common occurrence of both mutation types in hydrogen bonding residues, although there are significantly fewer hydrogen-bond disrupting CPDs than PDs. This suggests that it is difficult to make compensatory mutations which counteract the disruption of the intricate hydrogen-bonding network in the protein core.

The creation of voids of volume $> 275\text{Å}^3$ did not show a significant difference between the CPD and PD datasets. Our void calculation method[13] calculates the volume of voids assuming that no movements occur in the protein structure. In reality it is likely that several small movements of sidechains and backbone will occur to fill the void (at least partially). Only if these movements are too great will the stability and function of the protein be disrupted. It appears that in the CPDs, voids can be compensated for by replacing one or more local sidechains with a larger residue. A number of small changes can compensate as effectively as a single larger change and these may be accommodated more easily if, in evolution, they occur before the CPD. Figure 7 shows an example of a compensated void mutation in glucose-6-phosphate dehydrogenase.

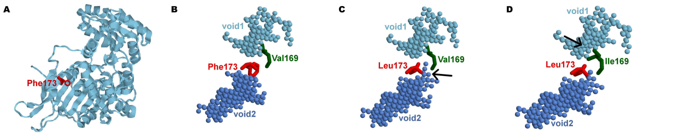

Figure 7: Potential compensation of a stability-reducing mutation. **a)** The Phe173Leu is shown on the human glucose 6 phosphate dehydrogenase structure, PDB:2bh9. This CPD creates a *void* and causes neonatal jaundice. **b)** Detail of Phe173 and its relative position to Val169. **c)** Substitution of aromatic Phe173 with a smaller leucine creates an enlarged ‘void2’ in the protein core, indicated by the arrow (modelled structure). **d)** Several bacterial FEPs have compensated for the void creation by substituting Val169 with a larger residue: Leu, Ile or Met. The compensatory effect of Val169Ile in *Buchnera aphidicola subsp. Schizaphis graminum* and *subsp. Baizongia pistacia* FEPs, shown here on a modelled structure. Introducing a larger isoleucine reduces the ‘void1’ size, increases the distance between the two voids, and in turn stabilizes the structure (indicated by an arrow). In **b)**–**d)**, only the two residues of interest are shown. The small spheres fill buried voids surrounding the residues and bounded by the rest of the protein structure.

Introducing a hydrophilic residue or an unsatisfied charge in the protein core[15] were significantly less likely to be compensated for, again, showing the great complexity of interactions among tightly packed buried residues. Compensating for a buried hydrophilic or charge would require introduction of a compensatory hydrophilic or charged residue (which, by itself, would be destabilizing) in a precise orientation in the core. The observation that such events are rare argues for the first DePristo hypothesis described above, in which phenotypically neutral compensatory mutations are introduced before the compensated mutation. Introducing a hydrophobic residue on the surface seems to be easier to compensate for, although a detailed analysis of multi-chain proteins and complexes with ligands would be required in order to explain these mechanisms fully.

In summary, frequencies of structural effects in both datasets presented here were quite similar to PD frequencies presented by Hurst *et al.*[15] The differences in frequencies between our overall counts per category and PD counts in that earlier work are a result of that PD dataset including other mutation sources in addition to OMIM. However, some categories typical for protein core residues (such as introducing a hydrophilic residue, buried charge, clash and SS-geometry) show a striking difference between PDs and CPDs, indicating these effects are less likely to be compensated for.

2.4. *Validity of the results*

Assignment of mutations as PDs or CPDs is based on a ‘negative’ observation (i.e. that this mutation, known to cause disease in humans, has *not* been observed as the native residue in a FEP from another species). Consequently, the number of CPDs may be an under-estimate simply because FEPs have not yet been observed demonstrating that compensation can take place.

In order to test that the significance of the results observered above was not a result of random partitioning of the data, a 10,000-iteration Monte Carlo simulation was run as described in the Materials and Methods. The results, shown in Table 3, indicate that where the observed (Bonferroni-corrected) *p*-value was < 0.01, the probability of seeing this *p*-value by chance was zero (i.e. *mc*-value = 0.00 when $p < 0.01$). Where $p < 0.05$, there was a >91.7% chance that the results were not obtained by chance (i.e. *mc*-value ≤ 0.083 when $p < 0.05$).

We can thus be confident that the results were not obtained by random chance divisions of the dataset.

3. Conclusions

The results presented here have three main novel aspects: (i) the orthologous proteins have been chosen on the basis of functional equivalence rather than sequence identity thresholds, (ii) CPDs have been surveyed in a structural context on a much larger scale than previous work and (iii) the range of surveyed effects of CPDs on protein structure is greater than in previous work. We used our SAAPdb database[15] to analyze the specific structural effects of CPDs in a range of structural categories, comparing them with PDs. The reliability of our analyses was increased by using data on functionally equivalent proteins for the multiple sequence alignments, because even relatively similar sequences can diverge in function[6]. We believe that the large size of the dataset and its wide spread across different protein families was sufficient for a broad structural analysis of human disease-associated single amino acid mutations and cases where these have been compensated in other species.

Our analysis of sequence divergence showed that residues local to a CPD tend to have random variability reflecting what would be expected from the overall sequence similarity (Figure 2b). In contrast, the local sequence divergence around PDs is significantly lower reflecting a requirement for conservation in these regions. Thus it appears that the surroundings of PDs are less able to undergo compensatory mutational events than one would expect by chance. For example, CPDs tend to be closer to the surface than PDs (as shown previously and confirmed in this work, see Table 2). The higher conservation around PDs may therefore be reflecting the conservation required within the core in order to maintain the structure of the protein. CPDs are more common near the surface simply because these regions can be less conserved. These observations also confirm that compensation tends to be a local effect in the majority of compensated mutations and suggests that compensation results from random mutational drift.

Structural analysis by the SAAPdb pipeline, which indicates the likely local structural effects of a mutation, showed important features of the CPD dataset. First, CPDs in humans were less often assigned any likely local structural effect, suggesting that they cause less significant disruption of local structure. This confirmed results by Ferrer-Costa *et al.*[5], suggesting that CPDs cause 'milder' changes than PDs in physico-chemical properties.

Second, CPDs often occur in interfaces. According to the first evolutionary model proposed by DePristo *et al.*[3], introduction of phenotypically neutral mutations (which are then able to compensate for a CPD) is a necessary first step before a CPD mutation can occur. Previously we have shown a high occurrence of neutral mutations in interface residues[15] and this may thus create an amenable environment for CPD occurrence. Thus it was not surprising to find the PQS-interface category being the only structural category having a slightly higher frequency of assigned CPDs than PDs (Figure 3). In contrast, disease-associated mutations were less likely to be compensated for when the residue had more complex intra-protein interactions (i.e. in the protein core), which would often require multiple compensatory events. Our results show that, based on structural categories as defined by SAAPdb, CPDs are more likely to be found among surface residues, with the exception of specific binding residues which make key hydrogen-bonding or van der Waals interactions across an interface. It is also possible that other factors may result in compensation such as changes in expression levels or accumulated biochemical differences.

In conclusion, we have performed a detailed structural comparison of the occurrence of compensated pathogenic deviations. Our structure-based results have confirmed an earlier proposal by Ferrer-Costa *et al.*[5] (based on sequence analysis) that the effects of CPDs are less drastic than uncompensated pathogenic deviations. Our larger dataset has also confirmed their result that CPDs are more likely to occur on the protein surface. Through a large-scale structural analysis, we have also confirmed the hypothesis that compensation tends to be a local effect, since local sequence variation around a CPD is greater than around sites of PDs in functionally equivalent proteins of the same sequence identity. Thus we have begun to differentiate compensated and uncompensated mutations on the basis of their effects on protein structure. This gives us insights into evolutionary mechanisms and may shed light on pathogenicity in humans.

4. Materials & Methods

An extensive set of 2328 missense human disease-causing mutations, extracted from OMIM[7,8], was mapped to the sequence data (Martin, manuscript in preparation, `http://www.bioinf.org.uk/omim/`). In brief, the method (which is described in more detail on the web site) uses cross references from UniProtKB/SwissProt to OMIM; a partial sequence is then constructed from the 'native' residues in OMIM and matched to the complete sequence in order to identify any offset that needs to be applied to the OMIM numbering to map mutations to UniProtKB/SwissProt. Subsequently mutations are mapped to structural data using PDBSWS[27]. The mutations were divided into two datasets, each mutation being either a 'PD' or a 'CPD', as shown in Figure 8. Two distinct analyses were performed on the datasets: (i) an analysis of the frequency of mutated residues within 8Å of the disease-associated mutation, by mapping aligned sequence data to structural data in the Protein DataBank (PDB)[17] as shown in Figure 9, and (ii) an

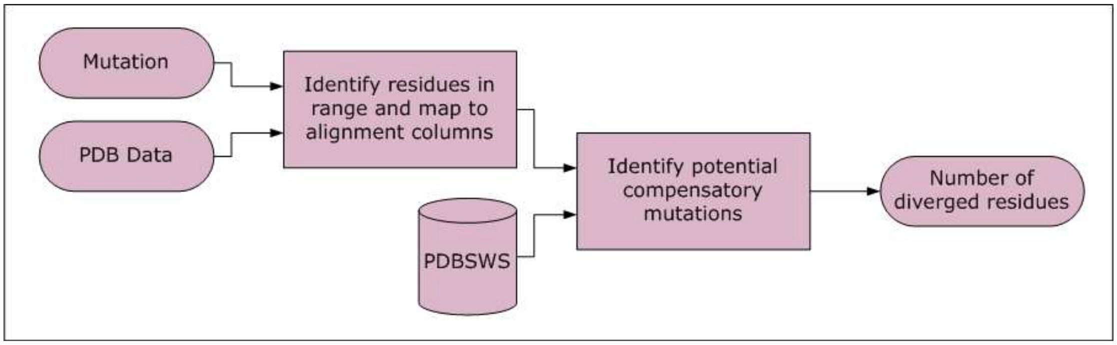
Mutation
PDB Data
Identify residues in range and map to alignment columns
PDBSWS
Identify potential compensatory mutations
Number of diverged residues

Figure 9: The number of mutated residues within 8Å of a CPD/PD mutation was counted.

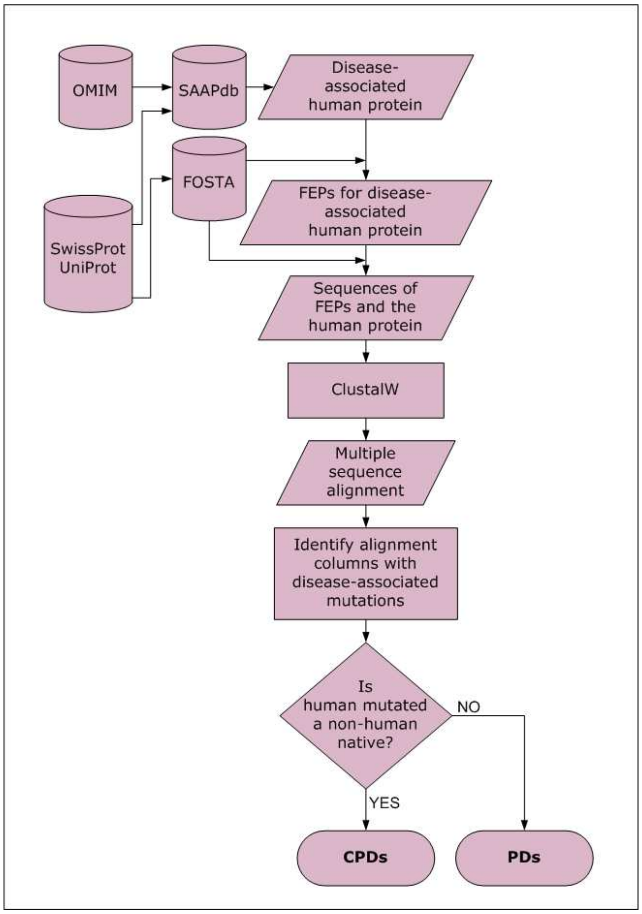
OMIM
SAAPdb
Disease-associated human protein
FOSTA
SwissProt UniProt
FEPs for disease-associated human protein
Sequences of FEPs and the human protein
ClustalW
Multiple sequence alignment
Identify alignment columns with disease-associated mutations
Is human mutated a non-human native?
NO
YES
CPDs
PDs

Figure 8: Creating CPD and PD datasets from the mutation data, the structural and sequence data and data about functionally equivalent proteins (FEPs).

analysis of local structural effects, using 14 structural explanations implemented in SAAPdb[15].

4.1. CPD dataset creation

We obtained a set of distinct CPDs from the OMIM missense mutations mapped to a residue in a UniProtKB/SwissProt[16] sequence and at least one PDB structure in SAAPdb, as shown in Figure 8. Although 2907 OMIM missense mutations were identified (April 2008 version of OMIM), 579 mutations could not be mapped to a residue in the PDB leaving 2328 mutations to be sorted in the two datasets. For every human sequence containing an OMIM mutation, a list of functionally equivalent proteins (FEPs) and their UniProtKB/SwissProt sequences were extracted from the FOSTA database[6]. FOSTA initially identifies a set of homologues from UniProtKB/SwissProt using BLAST. It then uses a series of text analyses of the UniProtKB/SwissProt annotations, initially looking for a match in the protein name element of the UniProtKB/SwissProt identifier, followed by the EC number and finally by matching synonyms at multiple levels of specificity from the UniProtKB/SwissProt description field. The sequences of the FEPs were then aligned with the human sequence using ClustalW [28]. Columns from the multiple sequence alignment containing disease-associated mutations in the human protein were then identified. If any of the non-human residues aligned to the human pathogenic mutation matched the amino acid causing the disease in humans, that mutation was sorted into the CPD dataset. An example of a CPD defined in this way is shown at residue 323 in Figure 10. Ser323Pro in humans causes disease, yet proline is the wild-type residue in sheep. Where the disease-causing mutation was not observed as the native residue in any other species, the mutation was placed in the PD dataset.

4.2. Detection of potential local compensatory mutations

Potential compensatory mutations were identified as follows, using the sequence alignment shown in Figure 11 as an example. After identifying the Ala419Val mutation in human sequence P01008 as a CPD because P41361 and P32262 (from cow and sheep respectively) contain a native valine at position 419, the best quality PDB structure (PDB:2b5t) mapped to the human P01008 was checked for all residues having at least one atom within 8Å of the Ala419 (the native CPD residue). These

```
                                    320  323
                                      |  |
Q5R5A3|ANT3_PONPY QVLELPFKGDDITMVLILPKPEKSLAKVEKELTPEVLQEWLDELEEMMLVVHMPRFRIED
P01008|ANT3_HUMAN QVLELPFKGDDITMVLILPKPEKSLAKVEKELTPEVLQEWLDELEEMMLVVHMPRFRIED
P32261|ANT3_MOUSE QVLELPFKGDDITMVLILPKPEKSLAKVEQELTPELLQEWLDELSETMLVVHMPRFRTED
P32262|ANT3_SHEEP QVLELPFKGDDITMVLILPKLEKPLAKVERELTPDMLQEWLDELTETLLVVHMPHFRIED
P41361|ANT3_BOVIN QVLELPFKGDDITMVLILPKLEKTLAKVEQELTPDMLQEWLDELTETLLVVHMPRFRIED
                  ******************** **.*****:****::******** * :******:** **
```

Figure 10: A CPD example in the human antithrombin-III (ANT3) protein aligned to its non-human FEPs. A Ser323Pro mutation in the human protein causes antithrombin-III deficiency, while Pro occurs in the wild-type sheep protein at the same position. These residues are highlighted in column 323, while a potential compensatory mutation is highlighted at column 320.

'in range' residues were then mapped back onto the alignment and these positions were checked for sequence divergence from the human sequence in the P41361 and P32262 sequences (the two sequences which showed a native Val419). We made the approximation that, having identified residues within 8Å of the mutation in the human structure, the equivalent residues in non-human sequences would also be within 8Å of the CPD residue in their respective structures. Thus all 'in range' differences in both of the FEP sequences compared with the human sequence were considered to be potential local compensatory mutations. A C/T ratio was calculated for each of the CPD-containing FEP sequences (P41361 and P32262), where C was the number of local (potential compensatory) mutations and T was the total number of 'in range' columns checked for that sequence. In other words, this ratio is the fraction of spatially neighbouring residues which are mutated. C/T was recorded together with the overall pairwise sequence identity.

Figure 11 also contains a pathogenic deviation (PD). At column 416, a mutation to proline causes disease and no proline is identified at this location in the FEPs from other species. For PDs, the C/T ratios were calculated for every non-human sequence aligned to the PD-containing human sequence and recorded with the pairwise sequence identity.

This was repeated for every alignment of sequences, examining both CPDs and PDs. Every C/T ratio was recorded together with the pairwise sequence identity of the FEP compared with the human sequence.

4.3. Structural analysis

After being divided into the CPD and PD datasets, every mutation was mapped to a residue in a PDB structure. The PDBSWS database[27], provides a mapping between PDB chains and UniProtKB/SwissProt or trEMBL entries derived from cross-links provided in the source data and enriched by 'brute-force' scanning of unmatched PDB chain sequences against UniProtKB/SwissProt and trEMBL using FASTA[29]. PDBSWS also provides alignments and residue-level equivalences. A given sequence may map to multiple PDB crystal structures, so a single entry was chosen on the basis first of sequence identity with the UniProtKB/SwissProt sequence, second of resolution and third of R-factor. The disease-associated mutation was labeled by the SAAPdb pipeline as 'explained' or 'unexplained' for every likely structural effect (Table 3)[15]. Note that one mutation can be assigned multiple likely structural effects.

The fraction of CPDs (or PDs) whose likely structural effect was explained by a given category, F_{cat}, was calculated as: $F_{cat} = N_{cat}/T_{cat}$, where N_{cat} is the number of CPDs (or PDs) predicted to cause that structural effect, and T_{cat} is the total number of mutations in the CPD (or PD) dataset. The difference between calculated fractions of the two datasets was tested by a two-tailed Fisher's exact test for statistical significance in each structural category (Table 3).

4.4. Potential compensatory mutation examples

The four compensation examples presented in the Results and Discussion section were created using RasMol[30]. Simple modelled structures were obtained using mutmodel[11] which replaces sidechains using a minimum perturbation protocol[31] where the sidechain's torsion angles are rotated to find the optimum orientation.

4.5. Monte Carlo simulations

Because of the division of data into PDs and CPDs via a negative observation (i.e. a mutation is defined as a PD because no compensatory event is observed), we tested whether the same significance values could be obtained by chance by a Monte Carlo simulation. Our dataset contained 447 CPDs and 1753 PDs; these data were merged and 447 mutations were chosen at random to create set A, the remaining 1753 being set B. For each of the structural explanation categories, a p-value was calculated (as before using a Fisher's exact test) based on this random division of the data. The random division and calculation of p-values was repeated 10000 times and for each structural explanation, the fraction of 'random p-values' that were lower than the observed p-value was recorded (Table 3).

5. Acknowledgments

AB was supported by an Overseas Research Scholarship and by the UCL Graduate School. LEMM was supported by an MRC Capacity Building studentship in Bioinformatics. The authors thank Miljenko Huzak for help with the Z-statistic for the comparison of the slopes in linear regression and the referees for useful suggestions.

```
                                                                         416 419
                                                                          |  |
P01008|ANT3_HUMAN GFSLKEQLQDMGLVDLFSPEKSKLPGIVAEGRDDLYVSDAFHKAFLEVNEEGSEAAASTAVVI
Q5R5A3|ANT3_PONPY GFSLKEQLQDMGLVDLFSPEKSKLPGIVAEGRDDLYVSDAFHKAFLEVNEEGSEAAASTAVVI
P32261|ANT3_MOUSE GFSLKEQLQDMGLIDLFSPEKSQLPGIVAGGRDDLYVSDAFHKAFLEVNEEGSEAAASTSVVI
P41361|ANT3_BOVIN SFSVKEQLQDMGLEDLFSPEKSRLPGIVAEGRSDLYVSDAFHKAFLEVNEEGSEAAASTVISI
P32262|ANT3_SHEEP SFSVKEQLQDMGLEDLFSPEKSRLPGIVAEGRNDLYVSDAFHKAFLEVNEEGSEAAASTVISI
                  .**:********* ********:****** **.************************** : *
```

Figure 11: A CPD example in the human antithrombin-III (ANT3) protein aligned to its non-human FEPs. The Ala419Val CPD mutation is assigned two C/T ratios, one for each CPD-containing sequence (ANT3_BOVIN and ANT3_SHEEP), while for the Ala416Pro PD mutation, four C/T ratios were calculated, one for every FEP sequence.

References

1. A. S. Kondrashov, S. Sunyaev, F. A. Kondrashov, Dobzhansky-muller incompatibilities in protein evolution., Proc. Natl. Acad. Sci. USA 99 (23) (2002) 14878–14883. doi:10.1073/pnas.232565499.
2. R. J. Kulathinal, B. R. Bettencourt, D. L. Hartl, Compensated deleterious mutations in insect genomes., Science 306 (5701) (2004) 1553–1554. doi:10.1126/science.1100522.
3. M. A. DePristo, D. M. Weinreich, D. L. Hartl, Missense meanderings in sequence space: a biophysical view of protein evolution., Nat. Rev. Genet. 6 (9) (2005) 678–687. doi:10.1038/nrg1672.
4. M. C. Cowperthwaite, J. J. Bull, L. A. Meyers, From bad to good: Fitness reversals and the ascent of deleterious mutations, PLoS Comput. Biol. 2 (2006) e141.
5. C. Ferrer-Costa, M. Orozco, X. de la Cruz, Characterization of compensated mutations in terms of structural and physico-chemical properties., J. Mol. Biol. 365 (1) (2007) 249–256. doi:10.1016/j.jmb.2006.09.053.
6. L. E. M. McMillan, A. C. R. Martin, Automatically extracting functionally equivalent proteins from SwissProt, BMC Bioinf. 9 (2008) 418.
7. V. A. McKusick, Online Mendelian Inheritance in Man (OMIM)(TM). McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD) (2000).
8. J. Amberger, C. A. Bocchini, A. F. Scott, A. Hamosh, McKusick's Online Mendelian Inheritance in Man (OMIM), Nuc. Ac. Res. 37 (2009) D793–D796.
9. G. Schlosser, G. P. Wagner, A simple model of co-evolutionary dynamics caused by epistatic selection, J. Theor. Biol. 250 (2008) 48–65.
10. S. Henikoff, J. G. Henikoff, Amino acid substitution matrices from protein blocks, Proc. Natl. Acad. Sci. USA 89 (1992) 10915–10919.
11. A. C. R. Martin, A. M. Facchiano, A. L. Cuff, T. Hernandez-Boussard, M. Olivier, P. Hainaut, J. M. Thornton, Integrating mutation data and structural analysis of the TP53 tumor-suppressor protein., Human Mut. 19 (2) (2002) 149–164. doi:10.1002/humu.10032.
12. C. J. Kwok, A. C. R. Martin, S. W. N. Au, V. M. S. Lam, G6PDdb, an integrated database of glucose-6-phosphate dehydrogenase (G6PD) mutations, Human Mut. 19 (2002) 217–224.
13. A. L. Cuff, A. C. R. Martin, Analysis of void volumes in proteins and application to stability of the p53 tumour suppressor protein, J. Mol. Biol. 344 (2004) 1199–1209.
14. A. L. Cuff, R. W. Janes, A. C. R. Martin, Analysing the ability to retain sidechain hydrogen-bonds in mutant proteins, Bioinformatics 22 (2006) 1464–1470.
15. J. M. Hurst, L. E. M. McMillan, C. T. Porter, J. Allen, A. Fakorede, A. C. R. Martin, The SAAPdb web resource: a large-scale structural analysis of mutant proteins, Human Mut. 30 (2009) 616–624.
16. B. Boeckmann, A. Bairoch, R. Apweiler, M.-C. Blatter, A. Estreicher, E. Gasteiger, M. J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, M. Schneider, The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003, Nuc. Ac. Res. 31 (2003) 365–370.
17. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne, The Protein Data Bank, Nuc. Ac. Res. 28 (2000) 235–242.
18. R. D. Finn, J. Mistry, B. Schuster-Bockler, S. Griffiths-Jones, V. Hollich, T. Lassmann, S. Moxon, M. Marshall, A. Khanna, R. Durbin, S. R. Eddy, E. L. L. Sonnhammer, A. Bateman, Pfam: clans, web tools and services., Nuc. Ac. Res. 34 (Database issue) (2006) D247–51. doi:10.1093/nar/gkj149.
19. B. K. Lee, F. M. Richards, The interpretation of protein structures: Estimation of static accessibility, J. Mol. Biol. 55 (1971) 379–400.
20. K. Henrick, J. M. Thornton, PQS: a protein quaternary structure file server, Trends Biochem. Sci. 23 (1998) 358–361.
21. E. N. Baker, R. E. Hubbard, Hydrogen bonding in globular proteins, Progr. Biophy. Molec. Biol. 44 (1984) 97–179.
22. G. D. Bader, I. Donaldson, C. Wolting, B. F. Ouellette, T. Pawson, C. W. Hogue, BIND–The Biomolecular Interaction Network Database, Nuc. Ac. Res. 29 (2001) 242–245.
23. A. C. R. Martin, A. M. Facchiano, A. L. Cuff, T. Hernandez-Boussard, M. Olivier, P. Hainaut, J. M. Thornton, Integrating mutation data and structural analysis of the TP53 tumor-suppressor protein, Human Mut. 19 (2002) 149–164.
24. A. N. Bullock, A. R. Fersht, Rescuing the function of mutant p53, Nature Reviews: Cancer 1 (2001) 68–76.
25. A. Friedler, D. B. Veprintsev, L. O. Hansson, A. R. Fersht, Kinetic instability of p53 core domain mutants: Implications for rescue by small molecules, J Biol Chem 278 (2003) 24108–24112.
26. A. Friedler, L. O. Hansson, D. B. Veprintsev, S. M. V. Freund, T. M. Rippin, P. V. Nikolova, M. R. Proctor, S. Rüdiger, A. R. Fersht, A peptide that binds and stabilizes p53 core domain: Chaperone strategy for rescue of oncogenic mutants, Proc Natl Acad Sci U S A 99 (2002) 937–942.
27. A. C. R. Martin, Mapping PDB chains to UniProtKB entries, Bioinformatics 21 (2005) 4297–4301.
28. J. D. Thompson, D. G. Higgins, T. J. Gibson, CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, Nuc. Ac. Res. 22 (1994) 4673–4680.
29. W. R. Pearson, D. J. Lipman, Improved tools for biological sequence comparison, Proc. Natl. Acad. Sci. USA 85 (1988) 2444–2448.
30. R. A. Sayle, E. J. Milner-White, RASMOL: biomolecular graphics for all, Trends Biochem. Sci. 20 (1995) 374–374.
31. H. L. Shih, J. Brady, M. Karplus, Structure of proteins with single-site mutations: A minimum perturbation approach, Proc. Natl. Acad. Sci. USA 82 (1985) 1697–1700.