Human Mutation

The SAAPdb web resource: a large scale structural analysis of mutant proteins

Journal:	*Human Mutation*
Manuscript ID:	humu-2008-0199.R1
Wiley - Manuscript type:	Research Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Hurst, Jacob; University College London, Institute of Structural and Molecular Biology McMillan, Lisa; University College London, Institute of Structural and Molecular Biology Porter, Craig; University College London, Institute of Structural and Molecular Biology Allen, James; University College London, Institute of Structural and Molecular Biology Fakorede, Adebola; University College London, Institute of Structural and Molecular Biology Martin, Andrew; University College London, Department of Biochemistry and Molecular Biology
Key Words:	Polymorphism, SNP, disease, database, amino acid mutation, structural bioinformatics

The SAAPdb web resource: a large scale structural analysis of mutant proteins

Jacob M. Hurst§, Lisa E. M. McMillan§, Craig T. Porter,

James Allen, Adebola Fakorede and Andrew C. R. Martin*

Institute of Structural and Molecular Biology,

Division of Biosciences, University College London, Gower Street,

London WC1E 6BT, United Kingdom

§ Joint first authors

* Corresponding author : andrew@bioinf.org.uk

Abstract

The Single Amino Acid Polymorphism database (SAAPdb) is a new resource for the analysis and visualisation of the structural effects of mutations. Our analytical approach is to map single nucleotide polymorphisms (SNPs) and pathogenic deviations (PDs) to protein structural data held within the Protein Data Bank. By mapping mutations onto protein structures, we can hypothesize whether the mutant residues will have any local structural effect which may 'explain' a deleterious phenotype. Our prior work used a similar approach to analyse mutations within a single protein. An analysis of the contents of SAAPdb indicates that there are clear differences in the sequence and structural characteristics of SNPs and PDs, and that PDs are more often explained by our structural analysis. This mapping and analysis is a useful resource for the mutation community and is publicly available at http://www.bioinf.org.uk/saap/db/.

Keywords

Polymorphism, SNP, disease, database, amino acid mutation, structural bioinformatics.

Introduction

Inherited genetic variation is critical in defining disease susceptibility. The term 'single nucleotide polymorphism' or SNP (International Hapmap Consortium, 2005) is often used to refer to any point mutation, but strictly SNPs are defined as allelic variants where the least common allele occurs in at least 1% of a normal population. They are estimated to occur once every 100-300 bases in the genome (Wang et al., 2006), giving rise to subtle phenotypic variation without causing severe and deleterious phenotypic changes. There have been two major efforts to catalogue and annotate SNPs: dbSNP (http://ncbi.nlm.nih.gov/projects/SNP ; Sherry et al., 2001) and the Human Genetic Variation database or HGVBase (http://hgvbase.cgb.ki.se; Fredman et al., 2004). HGVBase aims to provide a cleaned and validated set of data from dbSNP, augmented by disease causing mutations from locus-specific mutation databases or 'LSMDBs' (Claustres et al., 2002). HGVBase

provides validated annotations for all polymorphisms it describes, while dbSNP is a more general repository for SNP data. A third resource is the Human Gene Mutation Database (HGMD, Stenson et al., 2003) which collates disease associated genomic variation described in the literature. Much of the HGMD data however require purchase of a license and cannot be freely redistributed; consequently these data are not used here.

Disease-associated mutations occur at much lower frequencies and have a severe effect on phenotype. We use the term 'pathogenic deviation' (PD) to refer to any single base change reported to be correlated with disease. Online Mendelian Inheritance in Man (OMIM; http://www.ncbi.nlm.nih.gov/omim/) is a central repository for information about PDs. However a great deal more information is held and maintained by individual research groups in LSMDBs.

Non-synonymous SNPs (nsSNPs) and the vast majority of PDs result in a change at the protein level, to which we refer as 'single amino acid polymorphisms' (SAAPs). By mapping SAAPs onto protein structures, we hope to define how protein structure might be affected by mutant residues, and so 'explain' the functional effect (if any) of the mutation. It may also be possible to characterise SNPs and disease mutations differently in terms of their effect on protein structure. This information could then be used to predict whether a novel mutation would result in a disease phenotype or to design novel disease therapies. In this manuscript, we refer to all SAAP-PDs simply as PDs, and all SAAP-SNPs simply as SNPs.

Several approaches have been defined for predicting or classifying the effects of mutations on protein function. These include rule-based (Wang and Moult, 2001; Ramensky et al., 2002), probabilistic (Chasman and Adams, 2001), machine learning (Saunders and Baker, 2002; Krishnan and Westhead, 2003; Ferrer-Costa et al., 2004; Dobson et al., 2006) and purely evolutionarily based (Ng and Henikoff, 2003) methodologies.

A small number of these approaches are readily available on the web. MutDB (http://www.mutdb.org) annotates SAAPs from dbSNP and UniProtKB/SwissProt (Boeckmann et al., 2003) with structural information (Dantzer et al., 2005). PolyPhen (http://www.bork.embl-

heidelberg.de/PolyPhen) allows the user to analyse their own SAAPs, in addition to those in dbSNP (Ramensky et al., 2002). SIFT (http://blocks.fhcrc.org/sift/SIFT.html) is a method based entirely on evolutionary information derived from multiple sequence alignments (Ng and Henikoff, 2003), which calculates a sophisticated residue conservation estimate based on the observed and expected amino acid occurrence at each residue. It allows users to analyse their own SAAPs and SAAPs obtained from dbSNP. SNPeffect (http://snpeffect.vib.be/index.php) uses both sequence and structure based methods to predict the effect of dbSNP SAAPs on protein function (Reumers et al., 2006), the results of which the user can browse or search. SNPs3D (http://www.SNPs3D.org; Yue et al., 2006) employs several methods to assess the impact of SNPs on protein structure (Wang and Moult, 2001; Yue et al., 2005; Yue and Moult, 2006).

While these interfaces are useful, there is a need for more structural information, particularly molecular graphics, to be fully integrated into such resources in order to provide a complete and valuable service. In addition, a comprehensive structural analysis of both SNPs and PDs is required to contribute to the understanding of whether specific mutations will lead to disease. As such, we have developed the SAAPdb database and webserver, which provide two resources for the mutation community:

1. A website that can clearly present the location of mutated residues within solved protein structures
2. A fully automated and up-to-date structural analysis of these mutations that is accessible via the website

Materials and Methods

Statistical methods

χ^2 tests

Where possible, χ^2 statistics were calculated using truly representative expected values. To analyse rates of native and mutant residues, standard amino acid frequencies (Robinson and Robinson, 1991) were used to estimate the numbers expected in the dataset. To analyse mutation rates, the PAM30 (Dayhoff et al., 1978) matrix was normalised to include only positive values, and then normalised to sum to 100 to approximate the expected frequencies of a particular mutation. All χ^2 tests with one degree of freedom (i.e., a 2x2 contingency table) were carried out using the Yates correction.

Kolmogorov-Smirnov tests

Where distributions are being compared in the presence of ties, a bootstrapping method (n = 1000) was carried out using the ks.boot method in R (http://sekhon.berkeley.edu/matching/ks.boot.html). This more accurately estimates the p-value when comparing discontinuous distributions.

Log ratios

Log ratios demonstrate clearly where features in the dataset are seen more or less often than expected by some reference values. The log ratios for the mutant and native residues shown in Figure 1 were calculated using the standard amino-acid frequencies as described by Robinson and Robinson (1991). The log ratios for the mutation rates were calculated from the ratio of observed frequency of a mutation in the dataset to the expected frequency of a mutation in the dataset. The expected frequencies were also generated by transforming the PAM30 matrix to eliminate negative values, and normalising so that the values sum to 100. All log-ratios use $\log_2$.

Averaging across all mapped structures

Some SAAPs are mapped to a single structure (for example, mutations to P02766, human transthyretin), while others are mapped to over three hundred structures (e.g., mutations to the UniprotKB/Swissprot record P68871, human haemoglobin). Some structures may be of poorer quality and may give spurious solvent accessibility or torsion angle measurements. To account for this, the median of the relative accessibility scores and torsion angles is taken. In the case of the secondary structure DSSP codes (Kabsch and Sander, 1983), the mode of the data was used.

Mapping the PD dataset to protein sequence

The disease dataset in SAAPdb is derived mostly from Online Mendelian Inheritance in Man, or OMIM (http://www. ncbi.nlm.nih.gov/omim/). The dataset is currently augmented by five locus-specific mutation databases (LSMDBs): ADABase (http://bioinf.uta.fi/ADAbase/, adenosine deaminase deficiency); HAMSTeRs (http://europium.csc.mrc.ac.uk/WebPages/Main/main.htm, haemophilia A); G6PDdb (http://www.bioinf.org.uk/g6pd/, glucose 6-phosphate dehydrogenase); IARC TP53 Mutation Database (http://www- p53.iarc.fr/, TP53); a database of OTC mutations (Tuchman et al., 2002), and ZAP70base (http://bioinf.uta.fi/ZAP70base/index2.html, ZAP70 deficiency). See the PD section of Supplementary Table S1 for more details. These datasets are downloaded from their corresponding FTP sites. Currently, over 500 LSMDBs are recorded on the Human Genome Variation Society's website (http://www.hgvs.org/dblist/dblist.html). While SAAPdb only includes around 1% of these data, the system has been designed and implemented so that integrating more locus-specific data is straightforward (see below).

PDs can only be mapped to protein sequence where a UniProtKB accession number is provided or found for the mutated protein. The most significant challenge here is to verify the numbering provided by the datasets. In the case of OMIM, we have developed a mapping between OMIM mutations and UniProtKB which resolves discrepancies in numbering (http://www.bioinf.org.uk/omim/; Martin, in preparation). Here, a partial native sequence is reconstructred from the native residues in the list of mutations. This is then aligned with the named

UniProtKB sequence to calculate an offset and correct the numbering. Remaining mis-matched residues are rejected. The numbering provided by the LSMDBs is verified by comparison with the mapped sequence in the same way (see Supplementary Figure S1(a)).

The system has been designed to cope with the varying formats of the LSMDB datasets: each LSMDB is processed by a hand-crafted wrapper, which converts the data into a defined XML format. Writing this wrapper script is the only manual step required when importing a new set of PDs. The XML can then be processed identically by the pipeline.

Mapping the SNP dataset to protein sequence

For the results presented in this paper, SNP data were collected from dbSNP build 127 and HGVBase release 16.0. Both sets were downloaded in flatfile format via FTP.

Both dbSNP and HGVBase describe SNPs in the context of EMBL and/or Genbank records: the SNPs are mapped to an entry and flanking regions of 25 nucleotides are described. To map these data to protein sequences, we reconstruct the coding sequence from the EMBL/Genbank entry and identify where the SNP occurs (Cavallo and Martin, 2005). This method has recently been extended to include both dbSNP (Sherry et al., 2001) and HGVBase (Fredman et al., 2004) data, no longer uses EMBOSS (Rice et al., 2000) (see Supplementary Figure S1(b) for details) and is fully automated.

The second column in Table I indicates the number of mappings that are made from the SNP data to UniProtKB/SwissProt sequences (see the SNP section of Supplementary Table S1 for more details). The mapping method generates SQL that can be pushed straight into the database.

Mapping SAAPs to protein structure

If a SAAP is successfully mapped to a UniProtKB entry, it can then be mapped to one or more PDB structures using PDBSWS (Martin, 2005). The mapping process can yield multiple matches, as a protein sequence can map to multiple PDB structures, and to several chains within a single PDB structure. We retain all mappings for each SAAP.

Populating SAAPdb

The SNPs and disease-associated mutations are extracted from primary sources and mapped to protein structures as described above for import to the SAAPdb relational database. The PD and SNP importing and verification processes are depicted in Supplementary Figures S1(a) and S1(b) respectively.

In the second stage, all structurally mapped SAAPs are passed through a structural analysis pipeline, and the results are recorded in the relational database. The structural mapping and analysis phases are explained in more detail below.

SAAPdb is a PostgreSQL relational database for which we have developed methods to update the schema where data are more likely to change (for example, the inclusion of new structural analyses).

The Structural Analysis

The fully automated structural analysis pipeline examines the likely local effects of mutations on protein structure. The pipeline is implemented in python as a series of 'wrappers', which allow information to be passed to and from each individual analysis in a standard format (Supplementary Figure S2). The analyses themselves are implemented in various languages, including C and perl. Currently, the pipeline consists of fifteen structural analyses and one sequence analysis.

Seven of these analyses (hydrogen bonding, interface, proline, glycine, clash and void analyses) have been described previously (Kwok et al., 2002; Martin et al., 2002; Cuff and Martin, 2004; Cuff et al., 2006); the nine new analyses are described briefly below.

Introducing unfavourable hydrophobicity on the surface

Hydrophobic residues are concentrated in the protein core. Introducing a hydrophobic residue on the surface could result in protein aggregation, and therefore a deleterious phenotype. Disease examples of this include sickle-cell disease (Embury, 1986) and amyloid disease such as Alzheimers (Masliah et al., 1996). We classify the amino acids F, I, L, M, V and W as hydrophobic

and D, E, H, K, N, Q, R, S, T and Y as hydrophilic. This analysis is implemented as an SQL query as all the required data exist in the database.

Introducing unfavourable hydrophilicity in the core

Introducing a hydrophilic residue in the hydrophobic core could destabilise the native protein fold as that the vast majority of hydrogen bonding capable sidechains are found to participate in hydrogen bonding (McDonald and Thornton, 1994). Again, we classify F, I, L, M, V and W as hydrophobic and D, E, H, K, N, Q, R, S, T and Y as hydrophilic. Since all these data exist in SAAPdb, the analysis can again be implemented as an SQL query.

Introducing a charge shift in the core

Electrostatics are important in protein folding and stability: interactions around 'charge centres' in protein structures improve the stability of protein architecture (Torshin and Harrison, 2001). Disrupting the net charge of such structurally critical regions could destabilise the protein and affect function. We identify any mutation with relative ASA $\leq$5% which involves a charge shift (H, R and K are positively charged, D and E are negatively charged). This analysis is carried out as a single SQL query.

Mutations from cis-proline

The vast majority of cis-peptide bonds precede proline residues. Mutations from a cis-proline to any other amino acid are likely to be energetically unfavourable and may destabilise the native protein fold. Further, it has been shown that cis-prolines play a specific functional role in protein structure (e.g., in the thioredoxins; Martin, 1995); in these cases, mutations from cis-proline will directly affect protein structure and therefore function (Charbonnier et al, 1999; Nathaniel et al, 2003). We identify any mutations from a proline with $-90 < \omega < 90$. This analysis is implemented as a single SQL query.

Mutations to UniprotKB annotated residues

Many protein sequences in UniprotKB are annotated with functional information. Mutations to such residues could disrupt protein function. We identify SAAPs at sites annotated as ACT_SITE, BINDING, CA_BIND, DNA_BIND, NP_BIND, METAL, LIPID, CARBOHYD, MOD_RES, MOTIF, DISULFID and/or CROSSLNK. These data are parsed from UniprotKB and stored in a database table. The analysis is then implemented as an SQL query.

Mutations to binding residues

The MMDBBIND database (Bader et al., 2001) annotates PDB residues as being involved in binding to proteins, DNA or small molecules. Mutations to such residues are likely to be deleterious (e.g., Casamassimi et al., 1998). Binding residues are parsed from the MMDBBIND flatfile and stored in SAAPdb. The analysis is implemented as an SQL query.

Mutations to disulphide bonding residues

Disulphide bonding can be critical to maintaining the protein fold. Mutations from a cysteine participating in a disulphide bond could disrupt native protein structure (e.g., Lavergne et al., 1992). We analyse PDB files to identify cysteines with Sγ1-Sγ2 bond distance ≤2.50° apart, with Cβ1-Sγ1-Sγ2 bond angles of 104°± 10% (Hazes and Dijkstra, 1988), and record mutations to potential disulphide binding cysteines in the database. The analysis is implemented as an SQL query.

Disruption of quaternary structure

The residues at a quaternary interface will be critical to the assembly of the functional quaternary structure (e.g., Scopes et al, 1998; Steward et al, 2008). Accessible surface area (ASA) statistics are calculated – using a local implementation of the Lee and Richards algorithm (Lee and Richards, 1971) – for the residues in the entire structure and in isolated chains. An increase in relative ASA of >10% defines an interface residue. We calculate the ASA for each residue in all relevant structures from the PQS database (Henrick and Thornton, 1998) and store these data in the database; the analysis is implemented as an SQL query.

Mutations to highly conserved residues

In addition to the fifteen structural analyses, we have implemented a sequence analysis that identifies highly conserved residues in families of functionally equivalent proteins (McMillan and Martin, two papers in preparation). ImPACT (IMproved Protein Alignment Conservation Threshold) considers the patterns of conservation within an alignment and calculates an alignment-specific threshold for high conservation by modelling the distribution of conservation scores with multiple Gaussian components. It is a generally applicable method for identifying high conservation, rather than a method for identifying deleterious mutations (e.g., SIFT (Ng and Henikoff, 2003)). ImPACT thresholds are calculated for the alignments of all relevant sequences and their fully sequenced functionally equivalent proteins (FEPs) (McMillan and Martin, submitted). These data are stored in SAAPdb and the analysis is implemented as an SQL query.

Results and Discussion

Table I summarizes the content of SAAPdb (see Supplementary Table S1 for data on the individual resources). After importing the raw data there are approximately ten thousand pathogenic deviations (PDs) and over 16 million neutral mutations described. We successfully mapped 9617 PDs (8972 of which are unique) and 24492 SNPs (of which 14015 are unique) to amino acid changes in a UniprotKB sequence. Using PDBSWS (Martin, 2005), the SAAPs are then mapped onto PDB structures. Of the 9617 mapped and coding PDs, 44.91% are mapped to at least one PDB structure, but only 8.26% of the neutral 24492 mutations are identified in a protein structure leaving a more balanced final dataset.

PD residues are more often 'unique'

There are significant differences between PDs and SNPs in terms of the native and mutant amino acids (see Supplementary Table S2). Cysteine, arginine, glycine, tryptophan and tyrosine are more often native residues in the disease dataset; alanine, glutamic acid, isoleucine, lysine, glutamine, threonine and valine are mutated more often in the neutral dataset. It is clear that the

native residues associated with the SNP dataset are more 'replaceable'; that is, there is at least one other residue that behaves similarly and can often replace it without affecting function (for example, aspartic acid/glutamic acid, isoleucine/valine and lysine/arginine). The PD-associated native residues are less likely to be replaced without affecting function.

In terms of the residues *introduced*, there are significantly more deleterious mutations to cysteine, aspartic acid, proline, arginine, tryptophan and tyrosine, and significantly more neutral mutations that introduce alanine, phenylalanine, isoleucine, leucine, asparagine and valine. Once again, those residues common in the SNP dataset are more often replaceable, while the PD associated residues are more difficult to introduce without affecting function.

Figure 1a expresses each native amino acid as the log-ratio of observed percentages over the expected percentages and Figure 1b shows the same data for the mutant residues. Significant results are denoted with stars, two where $p \leq 0.01$ and one where $p \leq 0.05$ (see Supplementary Table S2 for details). Positive values in Figure 1 indicate that the amino acid is over-represented compared with the standard amino acid frequencies and negative values indicate under-representation.

With a view to discriminating between the two types of SAAP, the most interesting results are those that are significantly different from what is expected ($p \leq 0.05$), and over-represented in one dataset and under-represented in the other. Using these criteria we identify glycine, cysteine and tryptophan as 'discriminating' native residues that are enriched in the PD dataset, and glutamic acid, lysine, isoleucine and valine as 'discriminating' native residues that are enriched in the neutral dataset. Asparagine, isoleucine, phenylalanine and valine are favoured as mutant residues in the SNP dataset, while proline is the only mutant residue favoured in the deleterious dataset. Discriminating residues associated with the deleterious dataset have unique roles in protein structure, while those associated with the neutral dataset have characteristics that are shared with other amino acids and so may more readily be replaced, without resulting in a disease phenotype. This supports previous findings that glycine, cysteine and tryptophan are targets of deleterious polymorphisms (Vitkup et al., 2003; Dobson et al., 2006).

PDs are more often between very different residues

In the previous section, we considered the native and mutant residues independent of their mutation partner; here, we consider the mutation/polymorphism as a residue pair. Table II lists the discriminating polymorphisms that (a) occur at significantly different rates compared with what is expected, and (b) are found to be over-represented in one dataset and under-represented in the other (see Materials and Methods). Ten of the eleven discriminating SAAPs that are associated with the deleterious dataset include at least one of glycine, cysteine, tryptophan or proline, the residues identified in the previous section as a PD-favoured native or mutant residue. Interestingly, no from-tryptophan mutations are identified; it is possible that the deleterious effect of from-tryptophan mutations is predominantly due to losing the characteristics of tryptophan, and that the characteristics of the introduced residue are not important. It is also interesting to note that both C→Y and Y→C are the mutations that generate the two most disparate $\log_2$-ratio results in the two datasets.

In contrast, there are nineteen SNP-associated discriminating mutations, which include five pairs of 'commutative' mutations (i.e., X→Y and Y→X). Again, those discriminating residues identified from Figure 1 commonly occur in this dataset (K, E, I, V, F and N). Once again, many of these mutations are between amino acids often regarded as interchangeable (for example, aspartic acid/glutamic acid, lysine/arginine, isoleucine/valine, leucine/valine and glutamine/glutamic acid). There are some SNP-associated discriminating mutations for which the reasons are less obvious (for example, A→S, S→I, K→N, H→Q and V→A); however the majority describe two residues that share some characteristic, for example, hydrophobicity (V→A/K→N) or size (A→S).

It is striking that all eleven of the discriminating mutations enriched in the PD dataset have a negative BLOSUM62 score (Henikoff and Henikoff, 1992), while eleven of the nineteen discriminating mutations enriched in the neutral dataset have a positive BLOSUM62 score, indicating that mutations characteristic of the SNP dataset tend to be between similar residues, while PD mutations are more likely to be between very different amino acids (see final column of

Table II). This difference is statistically significant ($\chi^2_{df=15} = 533.55$, $p \approx 0$); a similar statistical difference is found for the PAM30 matrix (Dayhoff et al., 1978) (see Supplementary Figures S3(a-c) for details of the BLOSUM62 results).

It is impossible at this stage to comment with any confidence as to whether the profile of residue substitutions will change if the site of the mutation is, for example, on the surface, in the core, or at a functional site. However, there is a clear and significant tendency for PDs to be mutations to and from amino acids known to have a unique role in protein structure, and for SNPs to be mutations between physico-chemically similar residues.

PDs more often affect conserved residues

Residues that are highly conserved across diverse species have been consistently selected for across different branches of evolution. It is therefore likely that they are critical to protein function. Comparing the conserved residues of SNPs and PDs shows that PDs more often occur at sites of high conservation ($D = 0.12$, $p \approx 0$) (see Supplementary Figures S3(d-f) for details).

PDs are more commonly found in the protein core

SAAPdb maps 4316 PDs and 2022 SNPs to at least one PDB structure (Table III). In this section, the SAAPs are characterised with respect to structural features that are derived from the PDB data files. To account for disparity in both the number and quality of mapping structures in the SAAP datasets, the median value across all structures was used (see Materials and methods).

There are proportionally more buried residues in the PD dataset than the SNP dataset. A bootstrapped Kolmogorov-Smirnov test ($D = 7.56\times10^{-2}$, $p < 4.27\times10^{-99}$) confirms that this difference is statistically significant (see Materials and methods, and Supplementary Figures S3(g-i) for details).

It appears that SNPs are less likely to be buried than disease mutations. Residues in the core of the protein are generally critical to the stability of the structure and it follows that mutations in the core of the protein could critically affect protein stability and be deleterious. It is also likely that

surface residues, unless at critical functional sites or at the quaternary interface, can change more readily without disrupting protein structure and/or function. This trend has been identified elsewhere (Chasman and Adams, 2001; Ferrer-Costa et al., 2002; Saunders and Baker, 2002; Krishnan and Westhead, 2003; Yue et al., 2005).

PDs are more easily 'explained'

We have analysed the SAAPs with respect to basic sequence and structural characteristics. We now present an analysis of SAAPs within the context of their protein structure to 'explain' what the effect of the mutation might be. Figure 2 shows the results of the analyses, for each unique sequence mutation which has at least one mapped structure that provides a positive result for the corresponding explanation.

We intend to use these data to train a prediction method to identify deleterious mutations and are therefore hoping to see significant differences between the analyses of disease and neutral mutations as this will result in good discriminatory power in the prediction method. In addition, we would like to see high overall explanation success for PDs, as this suggests that our current suite of analyses is comprehensive enough to account for effects of mutations.

Disease mutations are more often explained by at least one of our analyses than neutral mutations (see bars marked 'All explanations' in Figure 2): 87.17% of disease mutations are explained by at least one analysis compared with 58.68% of neutral mutations. This difference is found to be highly statistically significant ($\chi^2_{df=1} = 552.99$, $p \approx 0$).

PDs most often affect protein stability

Much research has suggested that the deleterious effects of mutations are predominantly due to their effect on protein stability (Wang and Moult, 2001; Ferrer-Costa et al., 2002; Ferrer-Costa et al., 2004; Yue et al., 2005). We use six analyses to assess whether a mutation will destabilise the protein structure. The data indicate that PDs more often break native hydrogen bonds (28.33% of PDs, 16.07% of SNPs, $\chi^2_{df=1} = 84.86$, $p \approx 0$); more often create voids in the core of the protein

(40.19% of PDs, 11.98% of SNPs, $\chi^2_{df=1} = 401.84$, p≈0); more often introduce hydrophilic residues in the core of the protein (5.54% of PDs, 1.84% of SNPs, $\chi^2_{df=1} = 32.91$, $p = 9.67\times10^{-9}$); more often create a buried, unsatisfied charge (12.13% of PDs, 3.61% of SNPs, $\chi^2_{df=1} = 86.69$, p≈0) and more often break disulphide bonds (1.25% of PDs, 0.54% of SNPs, $\chi^2_{df=1} = 4.07$, $p = 4.37\times10^{-2}$; $p = 1.75\times10^{-2}$ one-tailed Fisher exact test) than SNPs. Unexpectedly, it is found that SNPs more often introduce a hydrophobic residue on the surface of a protein (9.39% of PDs, 11.30% of SNPs, $\chi^2_{df=1} = 4.13$, $p = 4.22\times10^{-2}$).

Taking all the instability analyses together, most PDs (65.48%) are explained, while 35.06% of SNPs are explained; this is a significant result ($\chi^2_{df=1} = 956.01$, p≈0). With respect to instability analyses, it appears that PDs are more often associated with destabilising changes in the core of the protein. This is consistent with our finding that PDs are more often buried than SNPs.

The PQS interface analyses assess whether the polymorphism occurs at a chain interface a protein structure. Approximately 40% of disease polymorphisms are found at an interface. Less than 30% of SNPs are found at an interface and a χ^2 test indicates that this difference is significant ($\chi^2_{df=1} = 119.47$, p≈0).

The binding analysis identifies residues that form protein-ligand bonds. It explains 18.50% of PDs and 8.30% of SNPs. This is found to be highly significant ($\chi^2_{df=1} = 82.85$, p≈0). Furthermore, PDs are statistically more often annotated in MMDBBIND ($\chi^2_{df=1} = 54.31$, $p = 1.71\times10^{-13}$). Taken together, the functional analyses account for 30.10% of disease mutations and 17.30% of SNPs; this is found to be statistically significant ($\chi^2_{df=1} = 173.85$, p≈0).

A third analysis which identified mutations at sites annotated by UniprotKB suffered low counts, and on close inspection was affected by coarse-grained annotation; in particular, entire protein chains are annotated as DNA_BINDing, rather than just the functional site, leading to biases in the data. These data are not discussed here and will not be used to train the prediction method, but remain in the database in the hope that annotations will improve.

Unfavourable torsion angles and bulky residue replacements could inhibit native protein

folding. It is clear from Figure 2 that mutations to proline and from glycine where torsion angles cannot be accommodated, and the introduction of bulky, clash-causing residues are more common in the disease dataset than the SNP dataset; mutations from cis-proline however are very rare in both datasets. Two analyses show that the pro, gly and clash results are significant, explaining 6.29% ($\chi^2_{df=1} = 38.63$, $p = 5.13\times10^{-10}$), 5.77% ($\chi^2_{df=1} = 16.65$, $p = 4.49\times10^{-5}$) and 15.27% ($\chi^2_{df=1} = 113.52$, $p\approx0$) of disease mutations respectively. Taken together, the folding analyses account for significantly more PDs (18.80%) than they do SNPs (7.74%) ($\chi^2_{df=1} = 136.44$, $p\approx0$).

These structural results support our sequence-based findings that mutations to proline, and mutations from glycine are significantly more common in the disease dataset.

Sequence conservation discriminates most successfully between PDs and SNPs

22.82% of disease mutations occur at a site of high conservation – as defined by ImPACT (McMillan and Martin, in preparation) – whereas less than 5% of neutral polymorphisms affect highly conserved residues. This difference is highly statistically significant ($\chi^2_{df=1} = 239.39$, $p\approx0$) and is consistent with the hypothesis that mutating residues that are highly conserved is likely to be disease-causing.

Presenting the data: the SAAPdb website

SAAPdb is accessible from http://www.bioinf.org.uk/saap/db/, where the user can search for either a disease associated mutation or a neutral mutation, using gene, protein or mutation IDs. The results for each mutation can be viewed aggregated across all mapped protein structures, or for each individual structure (see Supplementary Figures S4(a-b) for screenshot examples). Furthermore, the results for all mutations (both disease associated and neutral polymorphisms) to each protein structure can be viewed together. It is possible to view any or all of the mutations in RasMol by clicking on any of the small RasMol icons on the page (see Supplementary Figures S5(a-b) for screenshot examples). The data are available for download at http://www.bioinf.org.uk/saap/db/download.

Conclusions

We have conducted a broad, but by no means exhaustive, profiling of PDs and SNPs in terms of their sequence and structure characteristics. Significant patterns have emerged that demonstrate there are differences between deleterious and neutral polymorphisms, both at the sequence and structure level.

Native and mutant residues found to be enriched in the PD dataset (cysteine, glycine, tryptophan and proline) have characteristics not shared by other residues, while those residues associated with the SNP dataset more often share characteristics with other residues, making them easier to replace while maintaining native protein function and structure.

Taking into account the native/mutant pairs, thirty amino acid changes have been identified as 'discriminating', i.e., as occurring at statistically different rates in the PD and SNP sets while being over-represented in one dataset and underrepresented in the other dataset. The PD-enriched discriminating mutations more often have negative BLOSUM62 scores, while the SNP-enriched discriminating mutations more often have positive BLOSUM62 scores. Taking all SAAPs into account, PDs have lower BLOSUM62 scores than SNPs.

It was also shown that deleterious mutations are more often buried than SNPs. There is no correlation between deleterious phenotype and secondary structure or torsion angles (data not shown).

In addition to profiling the SAAPs on the basis of general sequence characteristics, each SAAP was passed through a suite of fifteen analysis programs to ascertain whether it is likely to have a local effect on protein structure, which would 'explain' a deleterious phenotype. It was shown that PDs are statistically more likely to be explained, with more than 87.71% of the PDs in SAAPdb and 58.68% of SNPs in SAAPdb explained by at least one analysis. Statistical testing of all individual analyses showed that across twelve of the thirteen explanation categories, PDs are statistically more likely to be explained, with PDs most often being characterised as disrupting stability. An additional sequence analysis which uses a novel method to score and assess

conservation scores also found that 22.82% of PDs are found at sites of high conservation, compared with 4.63% of SNPs (McMillan and Martin, in preparation).

With a view to building a predictive model of the data, it is interesting to note that the most discriminating analysis (in terms of proportion of the dataset explained) is the sequence analysis, followed by clash causing; introduction of a buried unsatisfied charge; void creation; mutations from proline, and introducing a hydrophilic residue in the core. Three times more PDs than SNPs are explained by at least one of these analyes. At a more general level, mutations likely to prevent folding are most strongly associated with PDs; however it remains that mutations likely to have a less severe effect of simply destabilising the native fold (in particular by introducing voids) account for the highest proportion of PDs. It is clear that the current set of structural analyses is not able to 'explain' all PDs with respect to their structural effects, and as such we have included the sequence analysis. We hope that a future, comprehensive set of analyses will fully account for the structural effects of mutations, but retain the sequence analysis while the comprehensive analysis suite is being developed.

These results are encouraging for the next stage of analysis, where the datasets described here will be used to train a prediction method to identify deleterious SAAPs from neutral SAAPs. The present analysis concentrates on individual characteristics of PDs and SNPs. However, by combining these data in a multidimensional analysis, it may be possible to identify more subtle and complex characteristics of deleterious SAAPs, and therefore train effective prediction methods to discriminate between them. Preliminary analysis of these data using several un-optimised machine learning techniques show prediction performances which are already comparable with recently published methods.

Acknowledgements

This project is supported by the Welcome Trust, UK. LEMM is funded by a UK Medical Research Council Capacity Building Studentship in Bioinformatics. We thank Kenric Leung for work on Protein C, SangTae Kim for work on OTC and Colin Kwok for work on G6PD. We would like to acknowledge Amy Wai-ling Butler, Ammar Al-Chalabi and John Powell for helpful discussions.

References

Bader GD, Donaldson I, Wolting C, Ouellette BF, Pawson T, Hogue CW. 2001. BIND – The Biomolecular Interaction Network Database. Nucleic Acids Res 29:242-245.

Boeckmann B, Bairo ch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M. 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 31:365-370.

Casamassimi A, Guiseppina M, Porcellini A, De Vita C, de Nigris F, Zannini M, Di Lauro R, Russo R, Avvedimento VE, Fusco A. 1998. p53 genes mutated in the DNA binding site or at a specific COOH-terminal site exert divergent effects on thyroid cell growth and differentiation. Cancer Research 58: 2888-2894.

Cavallo A, Martin ACR. 2005. Mapping SNPs to protein sequence and structure data. Bioinformatics 21:1443-1450.

Charbonnier JB, Melin P, Moutiez M, Stura EA, Quemeneur E. 1999. On the role of the cis-proline residue in the active site of DsbA. Protein Sci 8:96-105.

Chasman D, Adams RM. 2001. Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: Structure-based assessment of amino acid variation. J Mol Biol 307:683-706.

Claustres M, Horaitis O, Vanevski M, Cotton RGH. 2002. Time for a unified system of

mutation description and reporting: a review of locus-specific mutation databases. Genome Res 12:680-688.

Cuff AL, Janes RW, Martin ACR. 2006. Analysing the ability to retain sidechain hydrogen-bonds in mutant proteins. Bioinformatics 22:1464-1470.

Cuff AL, Martin ACR. 2004. Analysis of void volumes in proteins and application to stability of the p53 tumour suppressor protein. J Mol Biol 344:1199-1209.

Dantzer J, Moad C, Heiland R, Mooney S. 2005. MutDB services: Interactive structural analysis of mutation data. Nucleic Acids Res 33:W311-W314.

Dayhoff MO, Schwartz RM, Orcutt BC. 1978. Atlas of Protein Sequence and Structure. In: Dayhoff MO, editor. Washington DC: National Biomedical Research Foundation. p345-352.

Dobson RJ, Munroe PB, Caulfield MJ, Saqi MA. 2006. Predicting deleterious nsSNPs: an analysis of sequence and structural attributes. BMC Bioinformatics 7:217-217.

Embury, SH. 1986. The clinical pathophysiology of sickle-cell disease. Annu Rev Med 37: 361-367.

Ferrer-Costa C, Orozco M, de la Cruz X. 2004. Sequence-based prediction of pathological mutations. Proteins 57:811-819.

Ferrer-Costa C, Orozco M, de la Cruz X. 2002. Characterization of diseaseassociated single amino acid polymorphisms in terms of sequence and structure properties. J Mol Biol 315:771-786.

Fredman D, Munns G, Rios D, Sjöholm F, Siegfried M, Lenhard B, Lehväslaiho H, Brookes AJ. 2004. HGVbase: a curated resource describing a human DNA variation and phenotype relationships. Nucleic Acids Res 32:D516-D519.

Hazes B, Dijkstra BW. 1988. Model building of disulfide bonds in proteins with known three-dimensional structure. Protein Eng 2:119-125.

Henikoff S, Henikoff JG. 1992. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A 89:10915-10919.

Henrick K, Thornton JM. 1998. PQS: a protein quaternary structure file server. Trends

Biochem Sci 23:358-361.

International Hapmap Consortium. 2005. A haplotype map of the human genome. Nature 437:1299-1320.

Kabsch W, Sander C. 1983. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577-2637.

Krishnan VG, Westhead DR. 2003. A comparative study of machine-learning methods to predict the effects of single nucleotide polymorphisms on protein function. Bioinformatics 19:2199-2209.

Kwok CJ, Martin ACR, Au SWN, Lam VMS. 2002. G6PDdb, an integrated database of glucose-6-phosphate dehydrogenase (G6PD) mutations. Hum Mutat 19:217-224.

Lavergne JM, DePaillette M, Bahnak BR, Ribba AS, Fressinaud E, Meyer D, Peitu G. 1992. Defects in type IIA von Willebrand disease: a cysteine 509 to arginine substitution in the mature von Willebrand factor disuprts a disrupts a disulphide loop involved in the interaction with platelet glycoprotein IbIX. Br J Haematol 82:66-72.

Lee B, Richards FM. 1971. The interpretation of protein structures: Estimation of static accessibility. J Mol Biol 55:379-400.

Martin ACR. 2005. Mapping PDB chains to UniProtKB entries. Bioinformatics 21:4297-4301.

Martin ACR, Facchiano AM, Cuff AL, Hernandez-Boussard T, Olivier M, Hainaut P, Thornton JM. 2002. Integrating mutation data and structural analysis of the TP53 tumor-suppressor protein. Hum Mutat 19:149-164.

Martin JL. 1995. Thioredoxin-a fold for all reasons. Structure 3:245-250.

Masliah E, Sisk A, Mallory M, Mucke L, Schenk D, Games D. 1996. Comparison of neurodegenerative pathology in transgenic mice overexpressing V717F beta-amyloid precursor protein and Alzheimer's disease. J Neurosci 16:5795-811.

McDonald IK, Thornton JM. 1994. Satisfying hydrogen bonding potential in proteins. J Mol

Biol 238:777-793.

Nathaniel C, Wallace LA, Burke J, Dirr HW. 2003. The role of an evolutionarily conserved cis-proline in the thioredoxin-like domain of human class Alpha glutathione transferase A1-1. Biochem J 372:241-246.

Ng PC, Henikoff S. 2003. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res 31:3812-3814.

Ramensky V, Bork P, Sunyaev S. 2002. Human non-synonymous SNPs: server and survey. Nucleic Acids Res 30:3894-3900.

Reumers J, Maurer-Stroh S, Schymkowitz J, Rousseau F. 2006. SNPeffect v2.0: a new step in investigating the molecular phenotypic effects of human non-synonymous SNPs. Bioinformatics 22:2183-2185.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the Europ ean Molecular Biology Open Software Suite. Trends Genet 16:276-277.

Robinson AB, Robinson LR. 1991. Distribution of glutamine and asparagine residues and their near neighb ors in peptides and proteins. Proc Natl Acad Sci U S A 88:8880-8884.

Saunders CT, Baker D. 2002. Evaluation of structural and evolutionary contributions to deleterious mutation prediction. J Mol Biol 322:891-901.

Scopes DA, Baustista JM, Naylor CA, Adams MJ, Mason PJ. 1998. Amino acid substitutions at the dimer interface of human glucose-6-phosphate dehydrogenase that increase thermostability and reduce the stabilising effect of NADP. Eur J Biochem 251: 382-388.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29:308-311.

Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NST, Abeysinghe S, Krawczak M, Cooper DN. 2003. Human Gene Mutation Database (HGMD): 2003 update. Hum Mutat 21:577-581.

Steward RE, Armen RS, Daggett V. 2008. Different disease-causing mutations in

transthyretin trigger the same conformational conversion. Protein Engineering, Design and Selection 21:187-195.

Torshin IY, Harrison RW. 2001. Charge centers and formation of the protein folding core. Proteins: Structure, Funciton and Genetics 43:353-364.

Tuchman M, Jaleel N, Morizono H, Sheehy L, Lynch MG. 2002. Mutations and polymorphisms in the human ornithine transcarbamylase gene. Hum Mutat 19:93-107.

Vitkup D, Sander C, Church GM. 2003. The amino-acid mutational spectrum of human genetic disease. Genome Biol 4:R72-R72.

Wang P, Dai M, Xuan W, McEachin RC, Jackson AU, Scott LJ, Athey B, Watson SJ, Meng F. 2006. SNP Function Portal: a web database for exploring the function implication of SNP alleles. Bioinformatics 22:e523-e529.

Wang Z, Moult J. 2001. SNPs, protein structure, and disease. Hum Mutat 17:263-270.

Yue P, Li Z, Moult J. 2005. Loss of protein structure stability as a major causative factor in monogenic disease. J Mol Biol 353:459-473.

Yue P, Melamud E, Moult J. 2006. SNPs3D: candidate gene and SNP selection for asso ciation studies. BMC Bioinformatics 7:166.

Yue P, Moult J. 2006. Identification and analysis of deleterious human SNPs. J Mol Biol 356:1263-1274.

Figure legends

Figure 1a: Profiling SAAPs with respect to native residues

Profiling SAAPs by native residues, normalising by standard amino acid frequencies and calculating the log ratio (see Materials and Methods for details). Data for PDs and SNPs are coloured light grey and dark grey respectively and asterisks indicate a significant result (two where $p \leq 0.01$ and one where $p \leq 0.05$).

Figure 1b: Profiling SAAPs with respect to mutant residues

Profiling SAAPs by mutant residues, normalising by standard amino acid frequencies and calculating the log ratio (see Materials and Methods for details). Data for PDs and SNPs are coloured light grey and dark grey respectively and asterisks indicate a significant result (two where $p \leq 0.01$ and one where $p \leq 0.05$).

Figure 2: Profiling SAAPs with respect to explanations

The results for PDs and SNPs are shown in light grey and dark grey respectively. **pqs**: the mutation disrupts an inter-chain interface as defined by a change in solvent accessibility in structures described by PQS; **bind**: the mutation disrupts a binding site identified as an H-bond or specific Van der Waals contact in the PDB; **mmdb**: the mutation disrupts a binding site as described by MMDBBIND; **ANY BINDING**: the mutation is positive for bind and/or mmdb; **pro**: the mutation introduces a proline where torsions are unfavourable; **gly**: the mutation replaces a glycine where torsions are unfavourable; **clash**: the mutation causes a steric clash in the hypothesised mutant structure; **cispro**: the mutation replaces a cis-proline; **ANY FOLDING**: the mutation is positive for pro, gly, clash and/or cispro; **hbond**: the mutation breaks an existing hydrogen bond; **void**: the mutation creates a void in the protein core; **corephilic**: the mutation introduces a hydrophilic residue in the core; **surfacephobic**: the mutation introduces a hydrophobic residue on the surface; **buriedcharge**: the mutation introduces a buried unsatisfied charge; **ssgeom**: the

mutation disrupts a disulphide bond as calculated from PDB coordinates; **ANY INSTABILITY**: the mutation is positive for hbond, void, corephilic, surfacephobic, buriedcharge and/or ssgeom; **highcons**: the mutation affects a highly conserved residue; **EXPLAINED**: the mutation is explained by at least one of the above analyses. For more information on these analyses, see Materials and Methods.

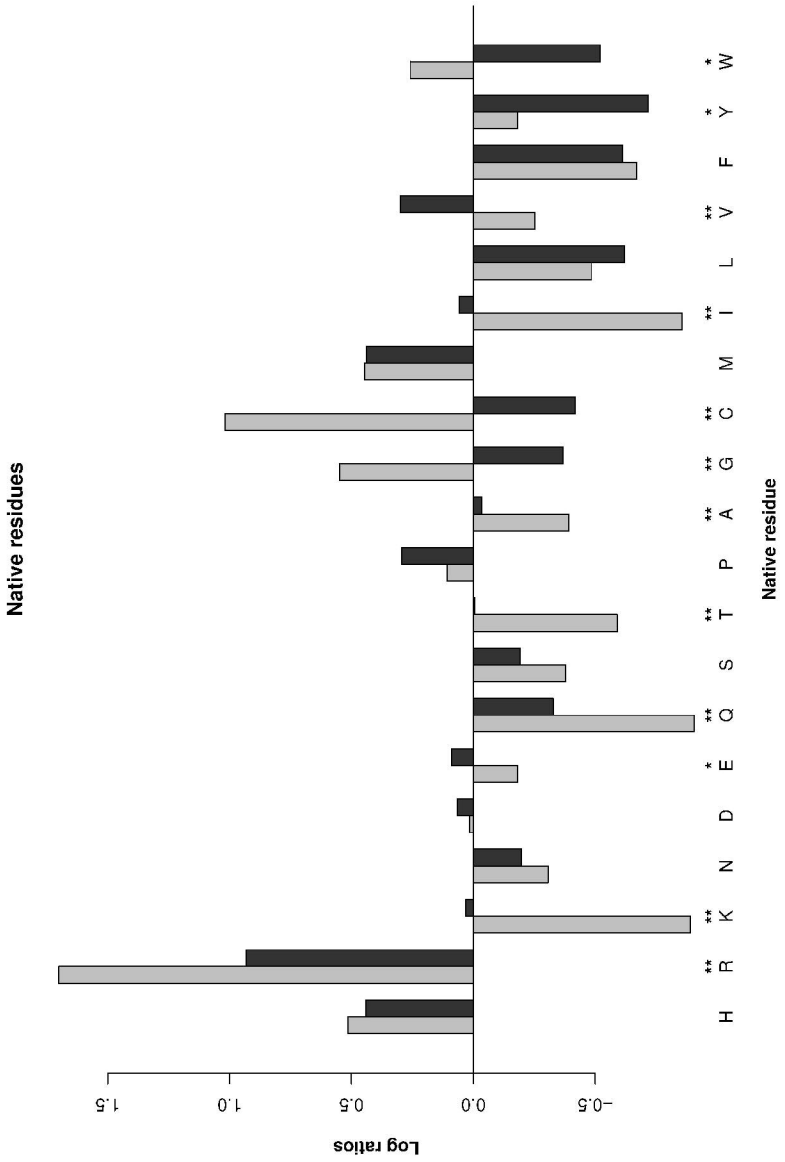
Table I : The contents of SAAPdb: mapping PDs and SNPs to protein sequences and structure

	PDs	SNPs
Raw number extracted from source database	9997	16227751
Raw number of SAAPs mapped to sequence	9617	24492
Unique sequence polymorphisms	8972	14015
Raw number mapped to structure	4319	2022

Table II : Mutations found to be significantly over-represented in one dataset, and underrepresented in the other

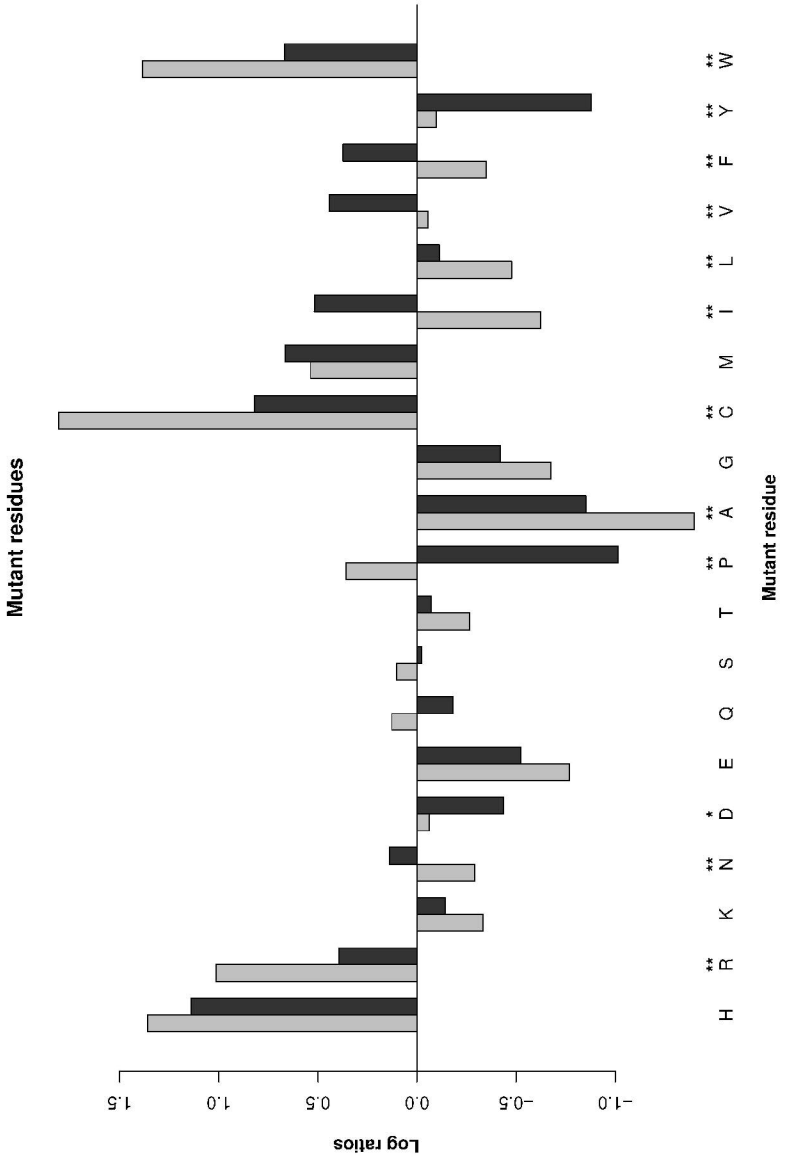
χ^2: the χ^2 statistic from a 2x2 Yates-corrected χ^2 test, testing for a difference in occurrence of the residue in the two datasets; **p**: the p-value (** denotes $p \leq 0.01$, * denotes $p \leq 0.05$); **set**: the SAAP set which has a higher occurrence of the corresponding amino acid; **L_2R**: the $\log_2$ ratio; **BLOSUM62**: the BLOSUM62 score for this mutation; results are ordered by the absolute difference between the two L_2R scores; all numbers are rounded to 2dp.

PD Mutations		χ^2	p	PD L_2R	SNP L_2R	BLOSUM62
C→Y	**	14.72	1.25×10^{-4}	0.68	-3.53	-2
Y→C	**	16.78	4.21×10^{-5}	1.09	-1.94	-2
F→C	*	4.30	3.82×10^{-2}	0.86	-2.04	-2
L→P	**	29.45	5.74×10^{-8}	2.32	-0.37	-3
R→P	**	10.98	9.22×10^{-4}	0.80	-1.53	-2
G→D	**	16.21	5.66×10^{-5}	1.40	-0.63	-1
L→R	**	8.50	3.55×10^{-3}	1.21	-0.72	-2
S→P	*	5.14	2.34×10^{-2}	0.19	-1.40	-1
G→S	**	9.69	1.86×10^{-3}	1.18	-0.26	-1
C→R	*	5.79	1.61×10^{-2}	1.19	-0.23	-3
G→E	*	6.19	1.29×10^{-2}	0.95	-0.36	-2
SNP Mutations		χ^2	p	PD L_2R	SNP L_2R	BLOSUM62
D→E	**	36.05	1.92×10^{-9}	-1.21	0.77	2
R→K	**	17.13	3.49×10^{-5}	-1.88	0.02	2
I→F	**	16.13	5.91×10^{-5}	-1.65	0.19	0
K→R	**	29.26	6.34×10^{-8}	-1.02	0.81	2
V→I	**	36.47	1.55×10^{-9}	-0.79	1.00	3
I→V	**	21.51	3.51×10^{-6}	-1.28	0.42	3
L→V	**	31.98	1.55×10^{-8}	-0.42	1.28	1
Q→H	**	21.51	3.51×10^{-6}	-1.21	0.49	0
A→S	**	23.04	1.59×10^{-6}	-0.88	0.75	1
E→D	**	31.86	1.65×10^{-8}	-0.45	1.13	2
I→M	**	10.58	1.14×10^{-3}	-1.35	0.10	1
S→I	**	8.99	2.72×10^{-3}	-0.93	0.52	-2
K→N	**	17.80	2.45×10^{-5}	-0.03	1.19	0
E→Q	**	8.34	3.88×10^{-3}	-1.06	0.12	2
H→Q	**	8.34	3.88×10^{-3}	-1.06	0.12	0
M→I	**	16.05	6.16×10^{-5}	-0.03	1.15	1
V→F	*	5.97	1.46×10^{-2}	-0.07	0.96	-1
V→A	**	9.24	2.37×10^{-3}	-0.15	0.87	0
S→C	*	4.66	3.09×10^{-2}	-0.41	0.46	-1



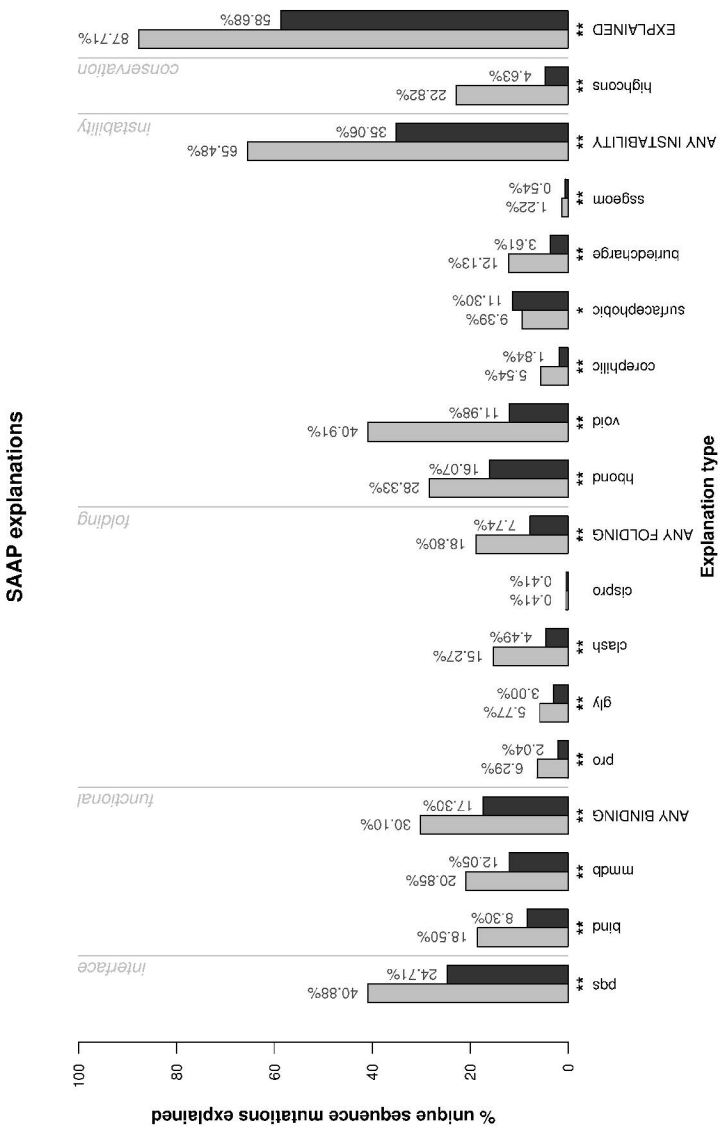
Profiling SAAPs by native residues, normalising by standard amino acid frequencies and calculating the log ratio (see Materials and Methods for details). Data for PDs and SNPs are coloured light grey and dark grey respectively and asterisks indicate a significant result (two where p≤0.01 and one where p≤0.05).

197x284mm (600 x 600 DPI)



Profiling SAAPs by mutant residues, normalising by standard amino acid frequencies and calculating the log ratio (see Materials and Methods for details). Data for PDs and SNPs are coloured light grey and dark grey respectively and asterisks indicate a significant result (two where p≤0.01 and one where p≤0.05).

197x284mm (600 x 600 DPI)



Profiling SAAPs with respect to explanations The results for PDs and SNPs are shown in light grey and dark grey respectively. pqs: the mutation disrupts an inter-chain interface as defined by a change in solvent accessibility in structures described by PQS; bind: the mutation disrupts a binding site identified as an H-bond or specific Van der Waals contact in the PDB; mmdb: the mutation disrupts a binding site as described by MMDBBIND; ANY BINDING: the mutation is positive for bind and/or mmdb; pro: the mutation introduces a proline where torsions are unfavourable; gly: the mutation replaces a glycine where torsions are unfavourable; clash: the mutation causes a steric clash in the hypothesised mutant structure; cispro: the mutation replaces a cis-proline; ANY FOLDING: the mutation is positive for pro, gly, clash and/or cispro; hbond: the mutation breaks an existing hydrogen bond; void: the mutation creates a void in the protein core; corephilic: the mutation introduces a hydrophilic residue in the core; surfacephobic: the mutation

introduces a hydrophobic residue on the surface; buriedcharge: the mutation introduces a buried unsatisfied charge; ssgeom: the mutation disrupts a disulphide bond as calculated from PDB coordinates; ANY INSTABILITY: the mutation is positive for hbond, void, corephilic, surfacephobic, buriedcharge and/or ssgeom; highcons: the mutation affects a highly conserved residue; EXPLAINED: the mutation is explained by at least one of the above analyses. For more information on these analyses, see Materials and Methods.
197x284mm (600 x 600 DPI)

Supplementary Materials

Supplementary Table S1 : Summary of SNP and PD data sources and the numbers mapping to sequence and structure

 # polymorphisms/mutations: the raw number of SNPs/PDs in the dataset; **# sequence mapped**: the number of SNPs/PDs mapped successfully to a UniprotKB sequence; **# structure mapped**: the number of SNPs/PDs mapped successfully to at least one PDB structure.

Dataset	Source	# polymorphisms	# sequence mapped	# structure mapped
PDs	**OMIM**	7298	7298	2557
	OTC	148	146	143
	G6PD	170	170	170
	HAMSTeRS	530	526	54
	IARC	1712	1258	1294
	ADABase	38	38	0
	ZAP70Base	5	5	5
	SOD1db	96	96	96
	Total	9997	9537	4319
SNPs	**HGVBase**	8274162	16302	1211
	DbSNP	7953589	8190	811
	Total	16227751	24492	2022

Supplementary Table S2 : Mutant and native residues in the SAAP datasets

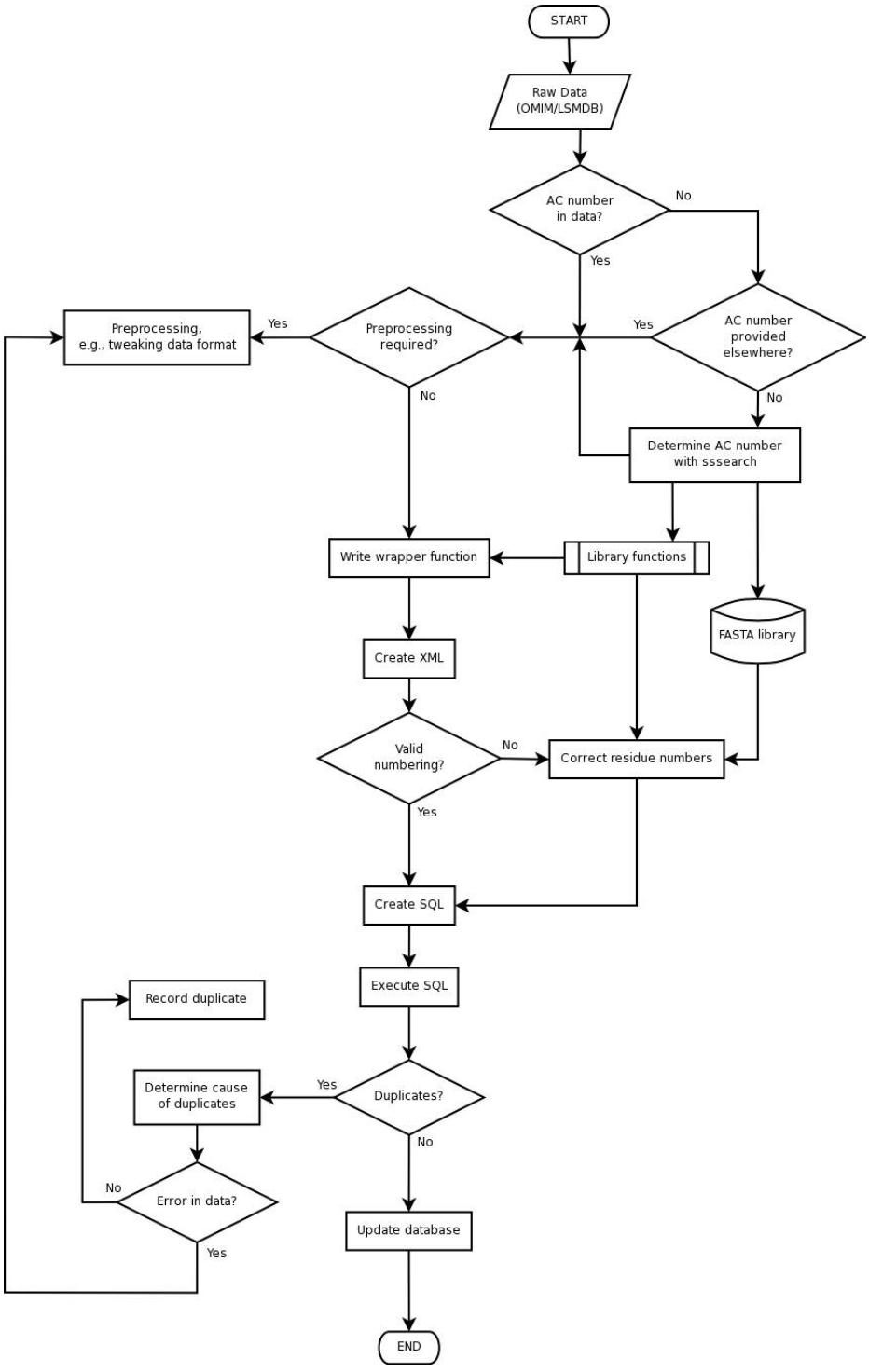
χ^2: the χ^2 statistic from a 2x2 Yates-corrected χ^2 test, testing for a difference in occurrence of the residue in the two datasets; **p**: the p-value (** denotes p≤0 .01, * denotes p≤0.05); **set**: the SAAP set which has a higher occurrence of the corresponding amino acid; all numbers are rounded to 2dp.

Native		χ^2	p	set
C	**	27.70	1.42×10^{-7}	PD
G	**	44.58	2.44×10^{-11}	PD
R	**	56.01	7.21×10^{-14}	PD
W	*	4.28	3.86×10^{-2}	PD
Y	*	4.24	3.95×10^{-2}	PD
A	**	8.05	4.56×10^{-3}	SNP
E	*	4.07	4.37×10^{-2}	SNP
I	**	32.62	1.12×10^{-8}	SNP
K	**	36.39	1.61×10^{-9}	SNP
Q	**	8.38	3.80×10^{-3}	SNP
T	**	15.32	9.06×10^{-5}	SNP
V	**	19.20	1.18×10^{-5}	SNP

Mutant		χ^2	p	set
C	**	29.66	5.16×10^{-7}	PD
D	*	4.15	4.16×10^{-7}	PD
P	**	46.38	9.74×10^{-7}	PD
R	**	22.62	1.98×10^{-7}	PD
W	**	8.76	3.07×10^{-7}	PD
Y	**	8.58	3.39×10^{-7}	PD
A	**	9.74	1.80×10^{-7}	SNP
F	**	19.43	1.04×10^{-7}	SNP
I	**	68.60	1.11×10^{-7}	SNP
L	**	9.23	2.38×10^{-7}	SNP
N	**	7.11	7.66×10^{-7}	SNP
V	**	17.59	2.74×10^{-7}	SNP

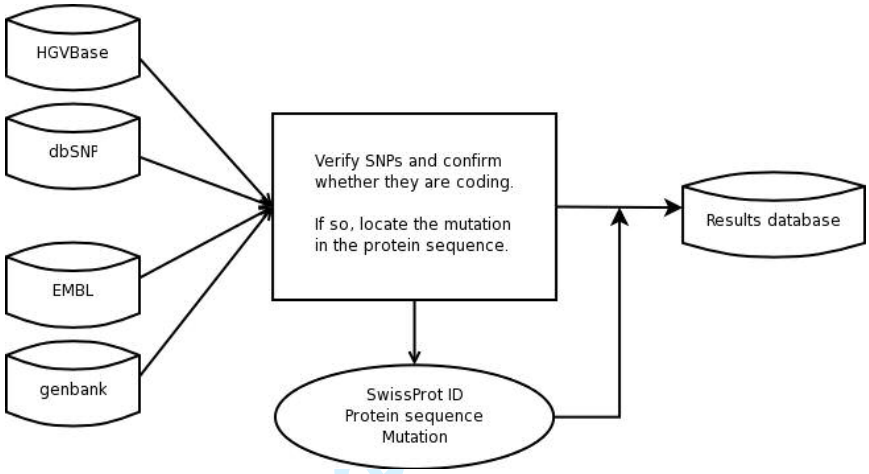
Supplementary Figure S1(a): Importing the PDs into SAAPdb

Disease associated SAAPs are entered a dataset at a time. The only manual process is the 'Write wrapper function' step where a dataset specific XML generation script must be written for each dataset.



Supplementary Figure S1(b): Importing the SNPs into SAAPdb

Importing SNPs is straightforward in that the data are already in the same format. However, the SNPs must be identified in the referenced databanks (EMBL or Genbank); this requires reconstruction of the coding sequence. This method is an extention of the method described in Cavallo and Martin, 2005, Bioinformatics, 21:1443-1450 (see text).

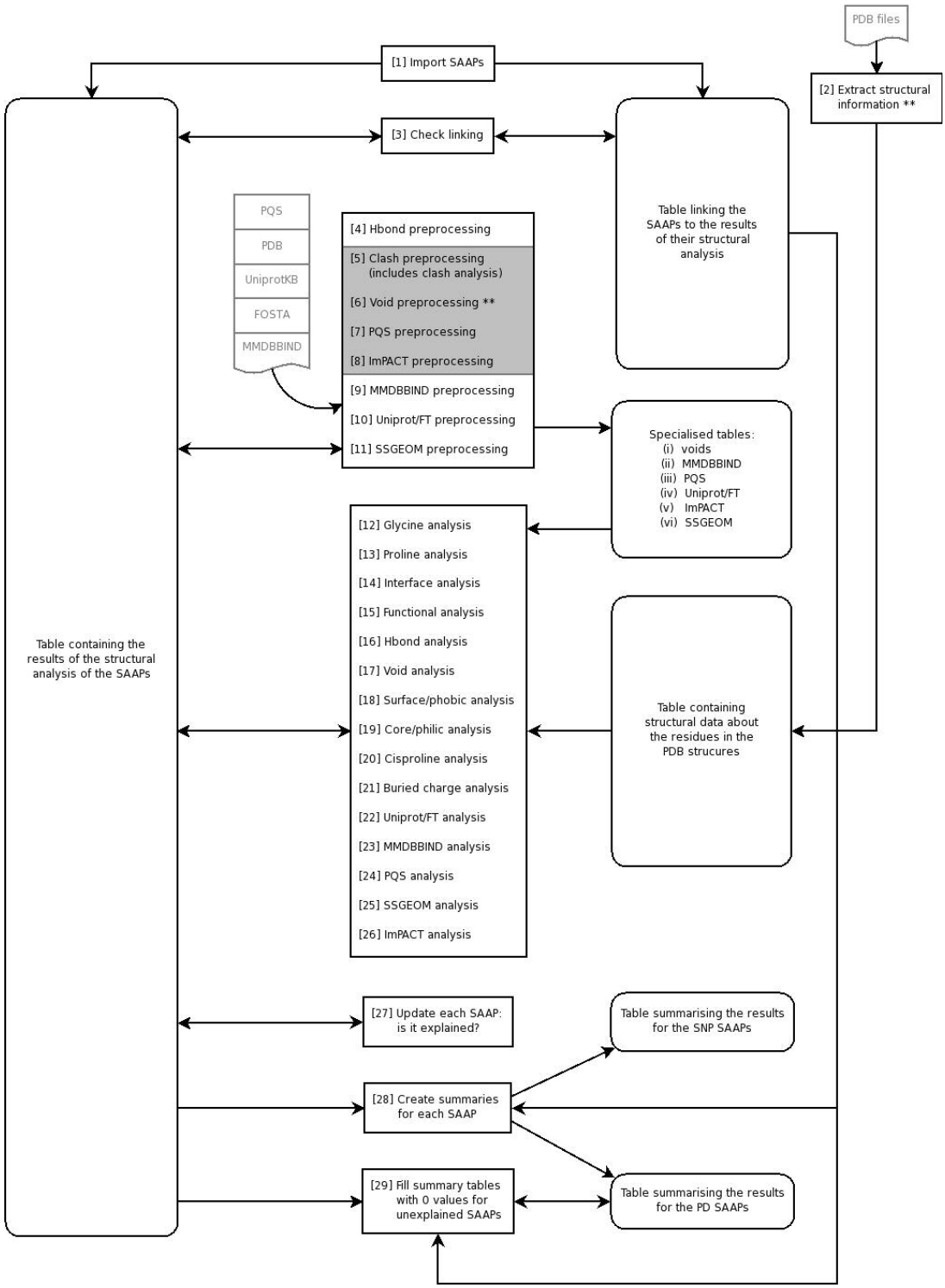


Supplementary Figure S2: Pushing the SAAPs through the pipeline

Square boxes indicate data processing; boxes with rounded corners represent database tables, and arrows indicate information flow.

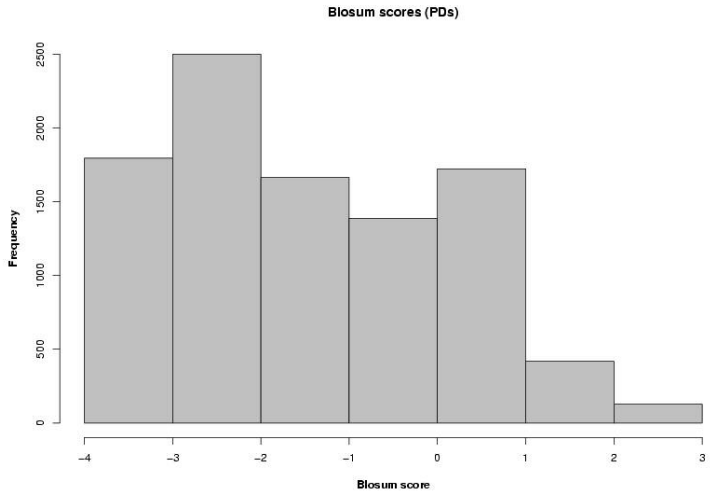
Processing is in four phases: (A) importing data; (B) pre-processing; (C) analyses; and (D) summarising results. These phases are named on the far left of the diagram and delineated by dashed grey lines. Steps [1-3] in phase (A) populate the database with all disease-associated SAAPs and structural information about all PDB structures; steps [4-11] in phase (B) generate mutant structures and carry out essential pre-processing for the hydrogen bonding, clash, void, MMDBBIND, UniprotKB/FT, PQS, ImPACT and SSGEOM analyses; steps [12-26] in phase (C) carry out the structural analyses; and steps [27-29] in phase (D) create summary tables for each SAAP.

Pre-processing requiring distributed processing is highlighted with a grey background, and data that are cached are highlighted with two asterisks (**).


PDB files
[1] Import SAAPs
[2] Extract structural information **
[3] Check linking
Table containing the results of the structural analysis of the SAAPs
Table linking the SAAPs to the results of their structural analysis
PQS
PDB
UniprotKB
FOSTA
MMDBBIND
[4] Hbond preprocessing
[5] Clash preprocessing (includes clash analysis)
[6] Void preprocessing **
[7] PQS preprocessing
[8] ImPACT preprocessing
[9] MMDBBIND preprocessing
[10] Uniprot/FT preprocessing
[11] SSGEOM preprocessing
Specialised tables:
(i) voids
(ii) MMDBBIND
(iii) PQS
(iv) Uniprot/FT
(v) ImPACT
(vi) SSGEOM
[12] Glycine analysis
[13] Proline analysis
[14] Interface analysis
[15] Functional analysis
[16] Hbond analysis
[17] Void analysis
[18] Surface/phobic analysis
[19] Core/philic analysis
[20] Cisproline analysis
[21] Buried charge analysis
[22] Uniprot/FT analysis
[23] MMDBBIND analysis
[24] PQS analysis
[25] SSGEOM analysis
[26] ImPACT analysis
Table containing structural data about the residues in the PDB strucures
[27] Update each SAAP: is it explained?
Table summarising the results for the SNP SAAPs
[28] Create summaries for each SAAP
[29] Fill summary tables with 0 values for unexplained SAAPs
Table summarising the results for the PD SAAPs

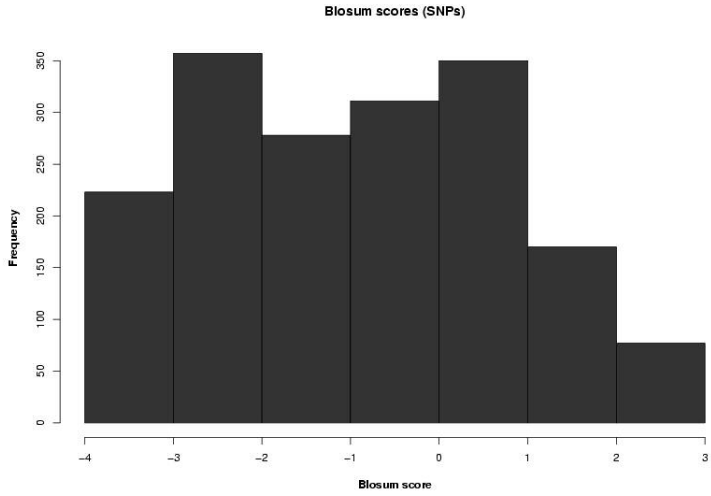
Supplementary Figure S3(a): The BLOSUM62 scores for PDs

The distribution of BLOSUM62 scores for PDs in SAAPdb.



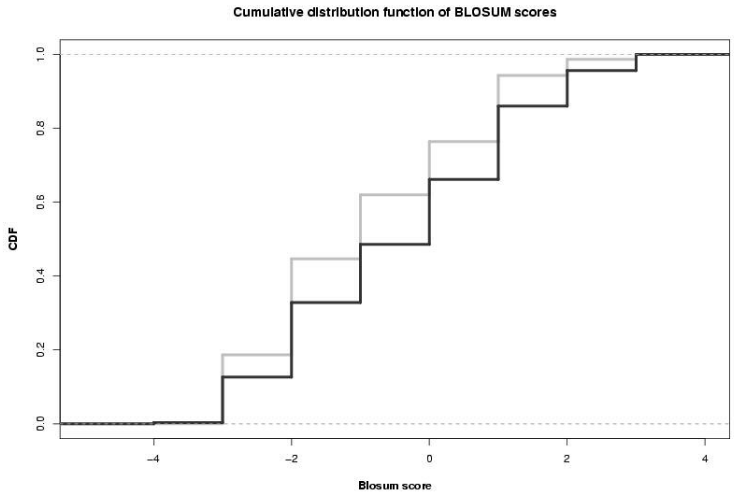
Supplementary Figure S3(b): The BLOSUM62 scores for SNPs

The distribution of BLOSUM62 scores for SNPs in SAAPdb.



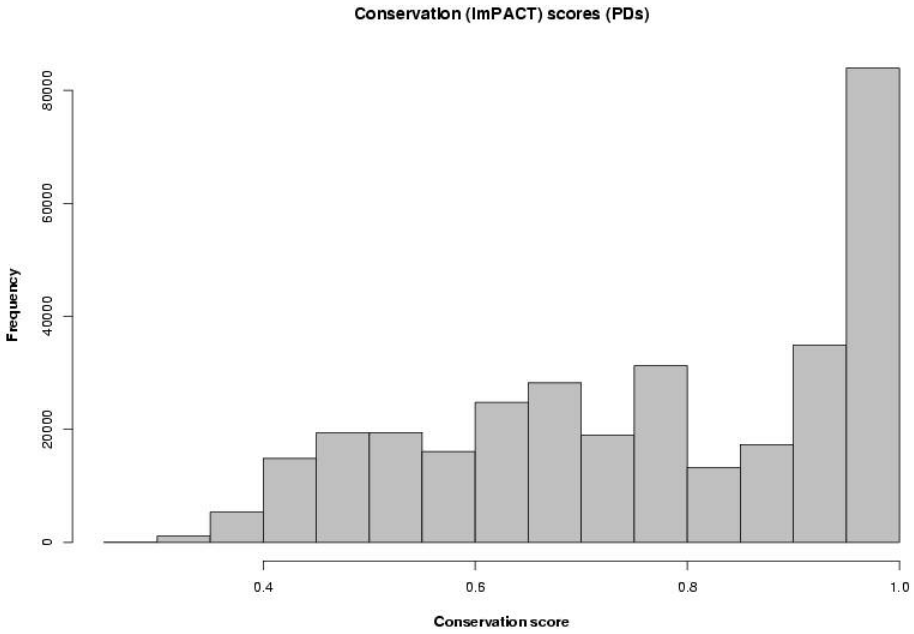
Supplementary Figure S3(c): Comparing PDs/SNPs using BLOSUM62 scores

The cumulative distribution plot, comparing PDs (light grey) and SNPs (dark grey) in terms of BLOSUM62 scores.



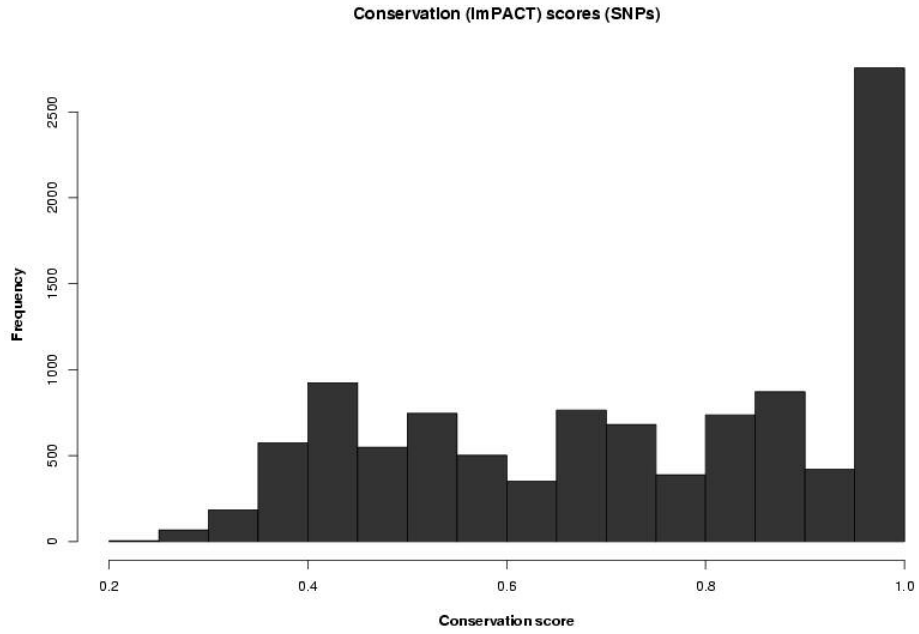
Supplementary Figure S3(d): The conservation scores for PDs

The distribution of conservation scores for PDs in SAAPdb.



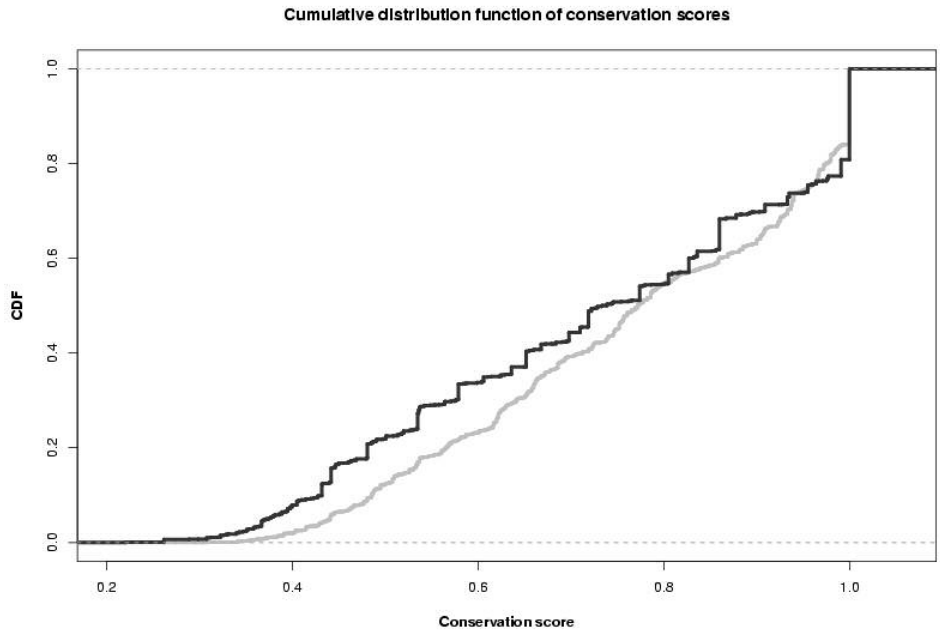
Supplementary Figure S3(e): The conservation scores for SNPs

The distribution of conservation scores for PDs in SAAPdb.



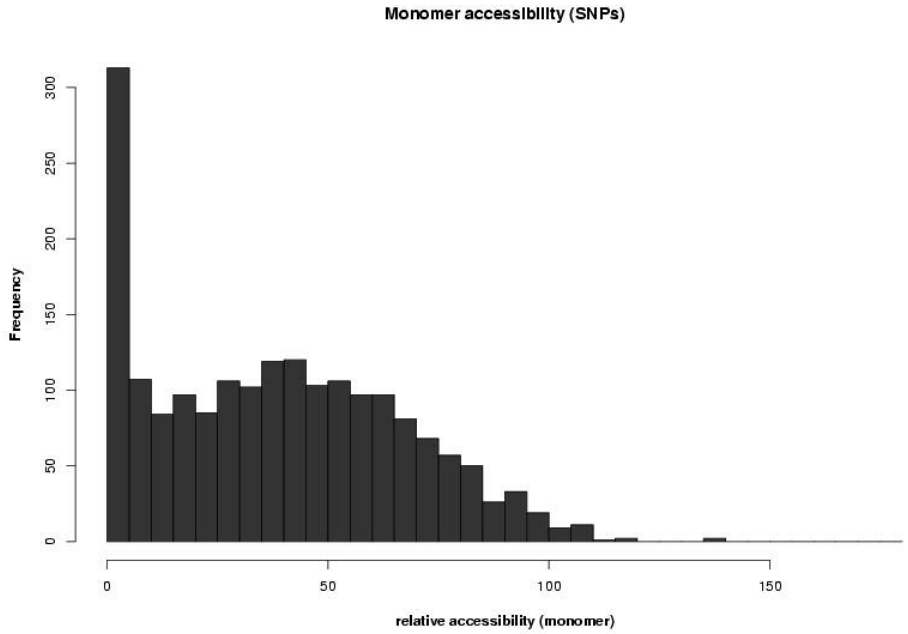
Supplementary Figure S3(f): Comparing PDs/SNPs using conservation scores

The cumulative distribution plot, comparing PDs (light grey) and SNPs (dark grey) in terms of conservation scores.



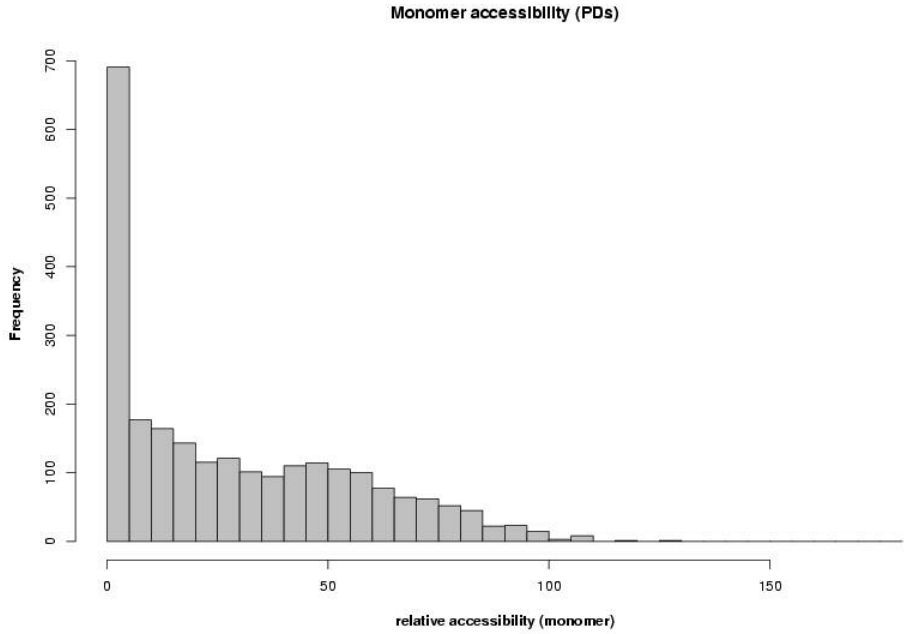
Supplementary Figure S3(g): The ASA for PDs

The distribution of ASA (accessible surface area) for PDs in SAAPdb.



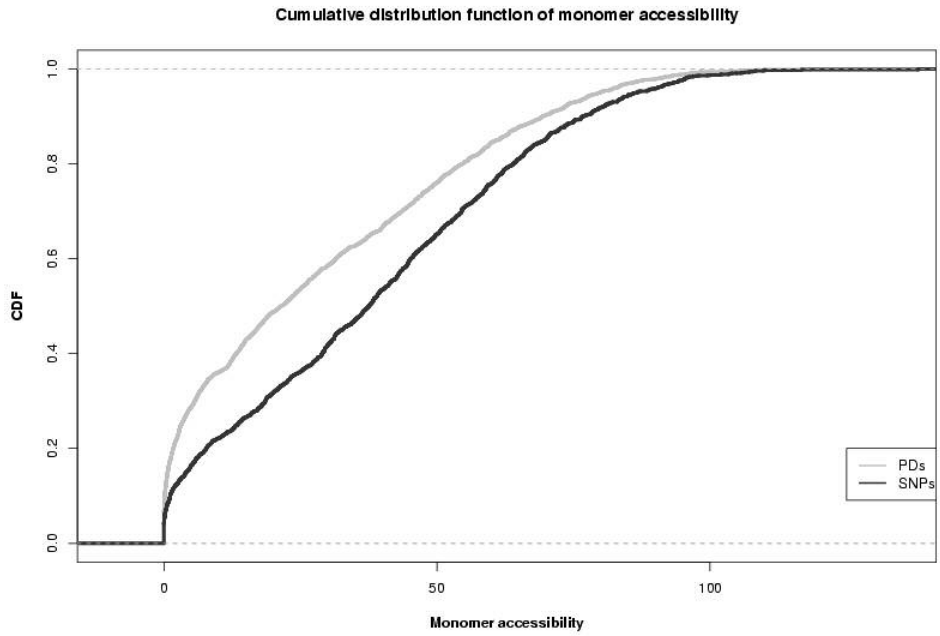
Supplementary Figure S3(h): The ASA for SNPs

The distribution of ASA (accessible surface area) for SNPs in SAAPdb.



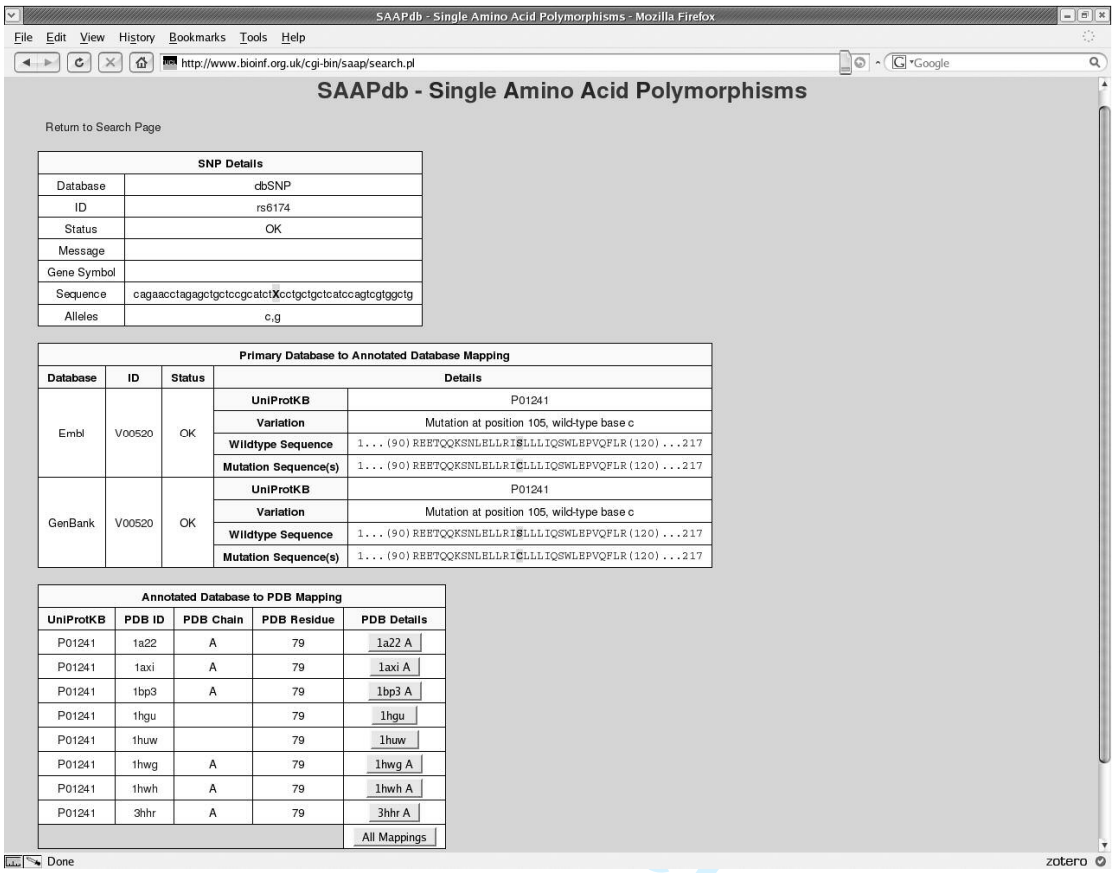
Supplementary Figure S3(i): Comparing PDs/SNPs using ASA

The cumulative distribution plot, comparing PDs (light grey) and SNPs (dark grey) in terms of conservation scores.



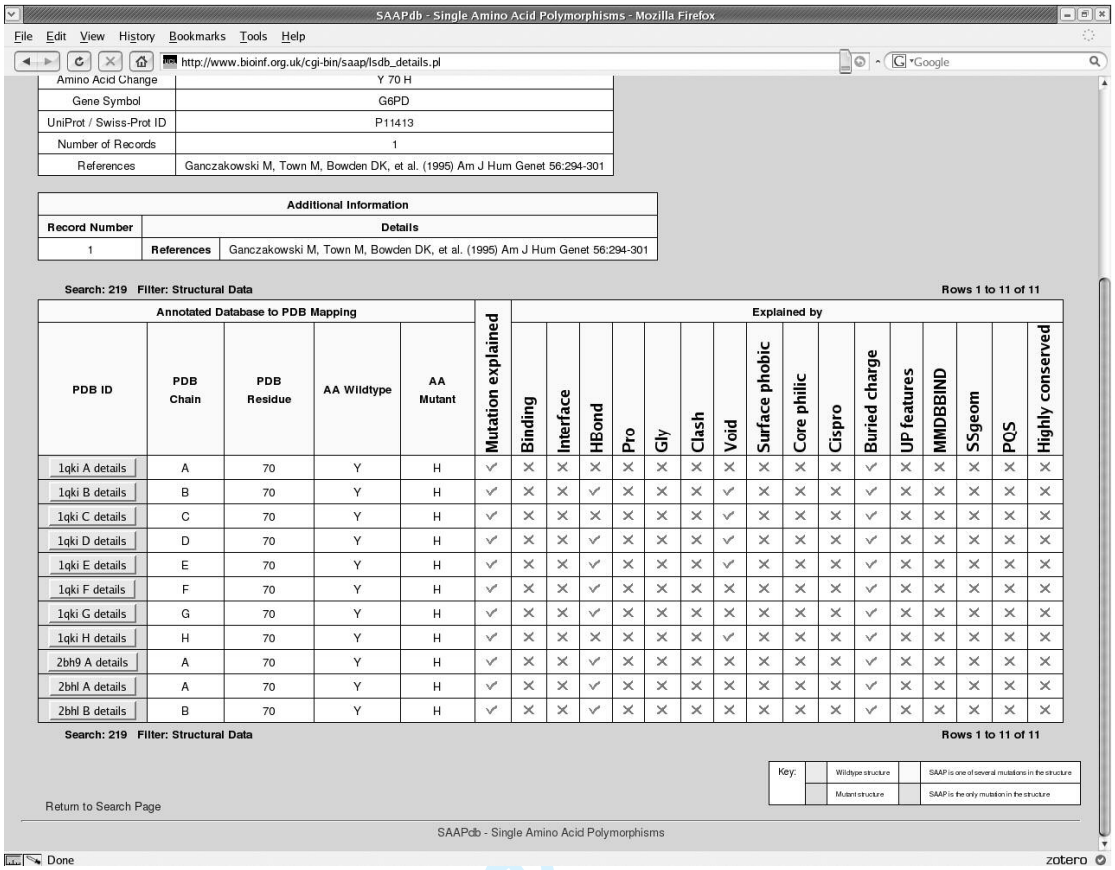
Supplementary Figure S4(a): A SNP record viewed on the SAAPdb webserver

An example of a SNP record, which describes the SNP itself, the SNP to UniProtKB mapping, and the UniProtKB to PDB mapping.



Supplementary Figure S4(a): Structural analyses results viewed on the SAAPdb webserver

The results of a structural analysis for a disease-associated SAAP for all mapped PDB chains; all SAAP analyses for a particular chain are obtained by pressing the corresponding button in the far left column.



SAAPdb - Single Amino Acid Polymorphisms - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://www.bioinf.org.uk/cgi-bin/saap/lsdb_details.pl

Amino Acid Change	Y 70 H
Gene Symbol	G6PD
UniProt / Swiss-Prot ID	P11413
Number of Records	1
References	Ganczakowski M, Town M, Bowden DK, et al. (1995) Am J Hum Genet 56:294-301

Additional Information		
Record Number	Details	
1	References	Ganczakowski M, Town M, Bowden DK, et al. (1995) Am J Hum Genet 56:294-301

Search: 219 Filter: Structural Data Rows 1 to 11 of 11

Annotated Database to PDB Mapping					Mutation explained	Explained by														
PDB ID	PDB Chain	PDB Residue	AA Wildtype	AA Mutant		Binding	Interface	HBond	Pro	Gly	Clash	Void	Surface phobic	Core philic	Cispro	Buried charge	UP features	MMDBBIND	SSgeom	PQS	Highly conserved
1qki A details	A	70	Y	H	✓	×	×	×	×	×	×	×	×	×	×	✓	×	×	×	×	×
1qki B details	B	70	Y	H	✓	×	×	✓	×	×	×	✓	×	×	×	✓	×	×	×	×	×
1qki C details	C	70	Y	H	✓	×	×	×	×	×	×	✓	×	×	×	✓	×	×	×	×	×
1qki D details	D	70	Y	H	✓	×	×	✓	×	×	×	✓	×	×	×	✓	×	×	×	×	×
1qki E details	E	70	Y	H	✓	×	×	✓	×	×	×	✓	×	×	×	✓	×	×	×	×	×
1qki F details	F	70	Y	H	✓	×	×	✓	×	×	×	×	×	×	×	✓	×	×	×	×	×
1qki G details	G	70	Y	H	✓	×	×	✓	×	×	×	×	×	×	×	✓	×	×	×	×	×
1qki H details	H	70	Y	H	✓	×	×	×	×	×	×	✓	×	×	×	✓	×	×	×	×	×
2bh9 A details	A	70	Y	H	✓	×	×	✓	×	×	×	×	×	×	×	✓	×	×	×	×	×
2bhl A details	A	70	Y	H	✓	×	×	✓	×	×	×	×	×	×	×	✓	×	×	×	×	×
2bhl B details	B	70	Y	H	✓	×	×	✓	×	×	×	×	×	×	×	✓	×	×	×	×	×

Search: 219 Filter: Structural Data Rows 1 to 11 of 11

Key:		Wildtype structure		SAAP is one of several mutations in the structure
		Mutant structure		SAAP is the only mutation in the structure

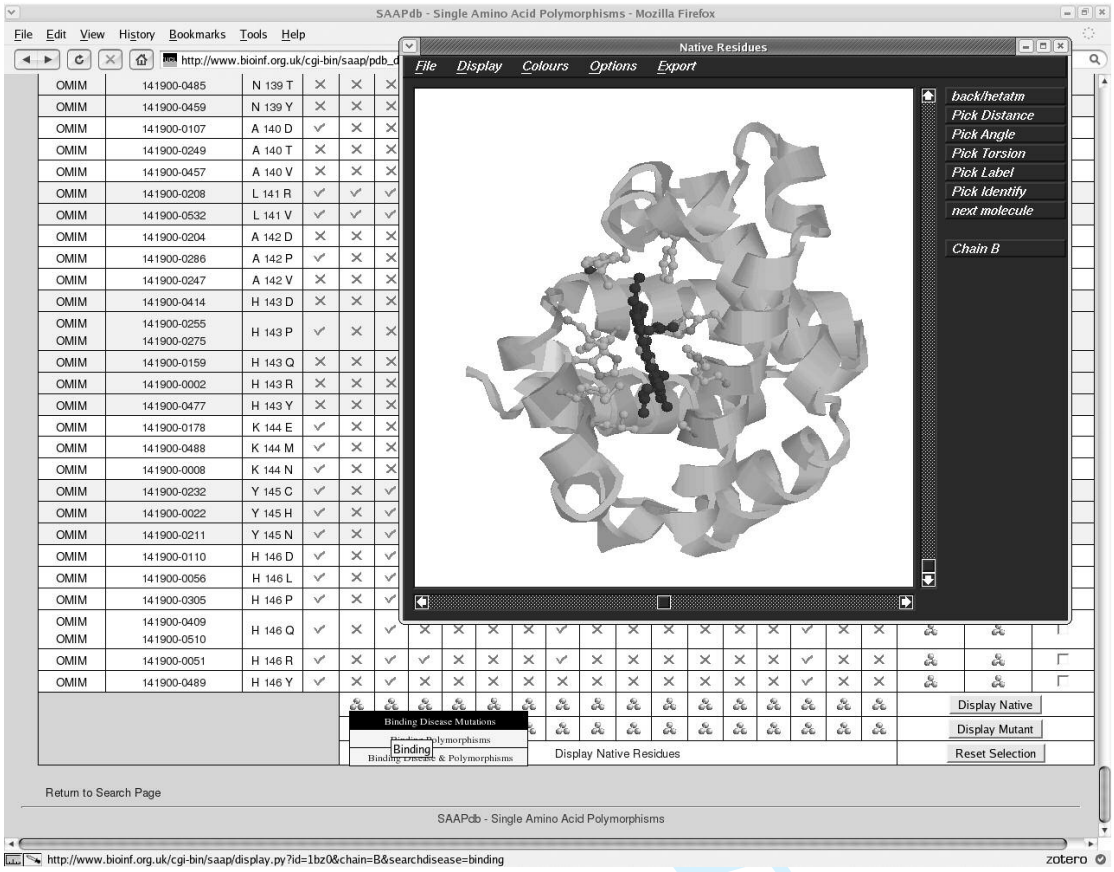
Return to Search Page

SAAPdb - Single Amino Acid Polymorphisms

Done zotero

Supplementary Figure S5(a): Using Rasmol to view a mutation at a binding site

Mutations explained by binding in a haemoglobin molecule are highlighted using the 'ball-and-stick' display option; it is clear that these mutations are clustered around the haem group, in the centre of the structure.



Supplementary Figure S5(a): Using Rasmol to view a mutation that introduces a hydrophobic residue on the surface

Mutations explained by introducing a hydrophobic residue on the surface of a haemoglobin molecule are highlighted using the 'ball-and-stick' display option.