Master in Proteomics and Bioinformatics

A new implementation for Swiss-Prot Variant Pages and a study on the conservation scores of all Swiss-Prot variants

Harris Procopiou

Supervisors: Yip Yum Lina and David Fabrice

University of Geneva – Faculty of Sciences
September 2007

Abstract

The rapid accumulation of Single Amino Acid Polymorphisms (SAPs) led the scientific community to better understand the mechanisms governing their disease-association. To facilitate the access of information related to variants, the UniProtKB/Swiss-Prot knowledgebase provides specific pages containing a summary of all the known relevant data for a given variant.

In this project, we developed an improved version of the Swiss-Prot Variant Web Pages by providing more sequence and structural information. New parameters, such as the conservation score, the structural local environment and the interface involvement of the variant, represent useful data to better understand the functional effects of these amino acid variations. Finally, we evaluated the predictive performance of the conservation score. Our results confirmed previous studies on the predictive power of this attribute for disease-associated variants.

1 Introduction

A Single Nucleotide Polymorphism, or SNP, is a small genetic variation that occurs within a genomic sequence. It is estimated that the world population contains a total of approximately 10 million SNP sites, resulting in an average density of one variant per 300 bases (Gibbs, *et al.*, 2003; Kruglyak and Nickerson, 2001; Reich, *et al.*, 2003). Because only three to five percent of DNA sequence codes for proteins, only a small number of SNPs is found in coding regions. Many of them do not produce phenotypic differences due to the degeneracy of the genetic code, while others can lead to a substitution of an amino acid by another in the sequence of the protein product. These are called non-synonymous SNPs (nsSNPs) or SAPs for *S*ingle *A*mino acid *P*olymorphism. SAPs are believed to play a major role in diseases or differing phenotypes (Gibbs, *et al.*, 2003; Pastinien, *et al.*, 2006). Indeed, whilst a large number of SAPs are functionally neutral, it is estimated that nearly half of these amino acid variations may cause deleterious effects on protein functions and are hence disease-associated (Krawczak, *et al.*, 2000). Despite the constant evolution of high-throughput techniques and the growth of SAP data in databases such as dbSNP (Sherry, *et al.*, 2001), HGVbase (Fredman, *et al.*, 2004), ModSNP (Yip, *et al.*, 2004) and many allele-specific databases, the mechanisms that explain why a SAP can be disease-associated is not yet fully understood.

Several methods have been developed in an attempt to discover potential predictors that could identify disease-associated nsSNPs from neutral nsSNPs. SIFT (Ng and Henikoff, 2001), for instance, is a sequence-based method that aims to predict if a substitution of an amino acid at an evolutionary specific position is tolerated or not. This method has shown great results but it deeply depends on the number of homologous sequences. Therefore, the necessity of combining structural and sequence data is crucial. To address this issue, rule-based approaches, support vector machines (SVMs), neural networks, decision trees and Bayesian methods are used (Bao and Cui, 2005; Cai, *et al.*, 2004; Dobson, *et al.*, 2006; Ferrer-Costa, *et al.*, 2004; Ramensky, *et al.*, 2002; Sunyaev, *et al.*, 2001; Yue and Moult, 2006). However, most of the methods offer merely the prediction result and users are given few hints on the basis of the prediction. More biologically informative structural and sequence attributes need to be investigated in order to better understand why SAPs can be associated with a disease.

A limited number of servers offer information on SNPs at the protein level. For instance, the *nsSNPAnalyzer* (http://snpanalyzer.utmem.edu) developed and maintained by Yan Cui's Lab at University of Tennessee Health Science Center allows predicting whether a SAP has a phenotypic effect. It provides additional information about the SNP to facilitate the interpretation of results, e.g., structural environment and multiple sequence alignment. However, it does not give a visual representation of the environment of the variant, rendering the interpretation of the results not trivial. Another automatic tool is *PolyPhen* (http://genetics.bwh.harvard.edu/pph) (Ramensky, et al., 2002). It aims to predict the possible impact of an amino acid substitution on the structure and function of a human protein. This prediction server provides information about the sequence features of the substitution site. Moreover, it computes several structural parameters like the accessible surface propensity, the residue side chain volume or the normed accessible surface area. It allows

also the computing of interchain contacts or contacts with functional sites. However, there is no visual indication of the chains and the positions of the residues in interchain contacts. Moreover, the degree of conservation of each position around the variant in the multiple sequence alignment is not defined.

1.1 Annotation of Single Amino acid Polymorphisms in UniProtKB/Swiss-Prot: Source of sequence data

The UniProtKB/Swiss-Prot database stores high-quality non-redundant protein entries that are manually annotated by curators. Annotation is a central process for the maintenance of the knowledgebase. It consists of analyzing, comparing and merging all available sequences for a given protein, as well as a critical review of associated data, either experimentally proven or predicted.

In UniProtKB/Swiss-Prot database, there are in total 28888 variants for 5234 human sequences (Release 52.3 of 17-Apr-2007). The variants of each protein are stored in the features table (FT) lines of the UniProtKB/Swiss-Prot.

```
FT   VARIANT      286      286      E->Q (in esophageal adenocarcinoma)
FT                                  /FTId=VAR_006030.
```

In this example, the P53 protein (*AC P04637*) is subjected to a substitution of a glutamic acid to a glutamine at the position 286 of its primary sequence. Each human variant is given a unique identifier, the FTid (*VAR_006030*).

The FT (Feature Table) lines also contain some UniProtKB/Swiss-Prot features that might aid in the understanding of the SAPs. Indeed, the table describes regions or sites of interest in the sequence. It lists principally post-translational modifications, binding sites and other particularities that are reported in the literature. Here is a non-exhaustive list of the features that can play a role in the mechanisms underlying why a variant can be disease-associated:

```
DOMAIN
CA_BIND
ZN_FING
DNA_BIND
NP_BIND
MOTIF
ACT_SITE
METAL
BINDING
SITE
MOD_RES
LIPID
CARBOHYD
DISULFID
CROSSLNK
VAR_SEQ
MUTAGEN
```

Fig 1 – Some Swiss-Prot features

Single amino acid polymorphisms are especially important for the research community. Consequently, they are treated by a specific medical annotation process in UniProtKB/Swiss-Prot. Information on the disease phenotypes associated with the variant can be found in the corresponding FT descriptions, as well as in the comment lines (CC), prefixed by the term DISEASE.

```
CC   -!-    DISEASE: Defects in TP53 are involved in esophageal
CC          cell carcinoma (ESCC) [MIM:133239]. ESCC is a tumor of
CC          The esophagus.
```

If the variant is a polymorphism not associated with a specific disease, information on alleles or polymorphisms can be found in the comment lines (CC), prefixed by the term POLYMORPHISM. For most variants, the corresponding literature references are stored in the references tables (in the lines coded: RN, RP, RC, RX, RA, RT and RL). The following example shows how an article on SAPs is cited in the P53 protein entry (AC P04637):

```
RN   [87]
RP   VARIANTS ESOPHAGEAL ADENOCARCINOMA PHE-176; SER-245; TRP-248;
RP   TRP-282 AND GLN-286.
RX   MEDLINE=96233927; PubMed=8829627 [NCBI, ExPASy, EBI, Israel,
RX   Japan];
RX   DOI=10.1002/(SICI)1098-1004(1996)7:2<109::AID-HUMU4>3.3.CO;2-0;
RA   Audrezet M.-P., Robaszkiewicz M., Mercier B., Nousbaum J.-B.,
RA   Hardy E., Bail J.-P., Volant A., Lozac'H P., Gouerou H., Ferec
RA   C.;
RT   "Molecular analysis of the TP53 gene in Barrett's
RT   adenocarcinoma.";
RL   Hum. Mutat. 7:109-113(1996).
```

Because all this information is scattered in different sections of a protein entry, the Swiss-Prot Variant Web Pages were created to provide a summary of all the known relevant data for a given variant. Moreover, while UniProtKB/Swiss-Prot is principally sequence-based, the variant pages aims to provide additional structural information for the understanding of the effects of disease-causing mutations. Each SAP entry contains five sections:

1. General Information
2. Information on the variant
3. Structural information on the variant
4. References for the variant
5. Cross-references for the variant

The data related to SAPs is recorded in ModSNP database (Yip, *et al.*, 2004). This database records information such as the classification of the variants, i.e. whether a variant is a polymorphism or disease-associated. It further manages the modeling of SAPs onto protein structures via an automatic modeling pipeline. In summary, ModSNP stores detailed information on mutations at the protein level that proves to be an essential support for the better understanding of variants. This is in contrary to other databases that are more focused on the genomic level, e.g. dbSNP.

1.2 The Protein Data Bank: Source of structural data

The Protein Data Bank (PDB) is the worldwide depository of information about experimentally determined three-dimensional structures of large biological molecules, including proteins and nucleic acids. These structures are typically obtained by X-ray crystallography or NMR spectroscopy. All this data is submitted by the international user community from around the world, and can be accessed for free.

The Protein Data Bank file format is a textual file that describes a given three-dimensional structure available in PDB. This file, divided in several sections, aims to provide the information about atomic coordinates, crystallization conditions, bibliographic citations, primary and secondary structure for a given PDB entry. Each column of the file has its own signification: Each parameter is positioned in an immutable way. Thus, the first six columns, i.e. the first six characters for a given line, determine the field of the file. One finds for example fields like:

TITLE_	The title of the studied macromolecule
KEYWDS	The keywords of the entry
EXPDTA	Gives information about the experimental method employed
SEQRES	The sequence of the studied molecule
ATOM__ Or **HETATM**	Gives information about a specific atom

Fig 2 – Fields one can find in a PDB file

The two last fields are of particular interest. Indeed, the ATOM and the HETATM sections present the atomic coordinates for standard residues and atoms within "non-standard" groups, respectively. Therefore, they indicate the localization of the atoms and residues in a three-dimensional space X, Y, Z. Here is the record format of the ATOM section:

COLUMNS	DATA TYPE	FIELD	DEFINITION
1 - 6	Record name	"ATOM "	
7 - 11	Integer	serial	Atom serial number.
13 - 16	Atom	name	Atom name.
17	Character	altLoc	Alternate location indicator.
18 - 20	Residue name	resName	Residue name.
22	Character	chainID	Chain identifier.
23 - 26	Integer	resSeq	Residue sequence number.
27	AChar	iCode	Code for insertion of residues.
31 - 38	Real(8.3)	x	Orthogonal coordinates for X
39 - 46	Real(8.3)	y	Orthogonal coordinates for Y
47 - 54	Real(8.3)	z	Orthogonal coordinates for Z
55 - 60	Real(6.2)	occupancy	Occupancy.
61 - 66	Real(6.2)	tempFactor	Temperature factor.
77 - 78	LString(2)	element	Element symbol, right-justified.
79 - 80	LString(2)	charge	Charge on the atom.

Fig 3 – The ATOM record format

Here is an example of the ATOM section of a PDB file:

```
1         2         3         4         5         6         7         8
12345678901234567890123456789012345678901234567890123456789012345678901234567890
ATOM    145  N   VAL A  25      32.433  16.336  57.540  1.00 11.92           N
ATOM    146  CA  VAL A  25      31.132  16.439  58.160  1.00 11.85           C
ATOM    147  C   VAL A  25      30.447  15.105  58.363  1.00 12.34           C
ATOM    148  O   VAL A  25      29.520  15.059  59.174  1.00 15.65           O
ATOM    149  CB AVAL A  25      30.385  17.437  57.230  0.28 13.88           C
ATOM    150  CB BVAL A  25      30.166  17.399  57.373  0.72 15.41           C
ATOM    151  CG1AVAL A  25      28.870  17.401  57.336  0.28 12.64           C
ATOM    152  CG1BVAL A  25      30.805  18.788  57.449  0.72 15.11           C
ATOM    153  CG2AVAL A  25      30.835  18.826  57.661  0.28 13.58           C
ATOM    154  CG2BVAL A  25      29.909  16.996  55.922  0.72 13.25           C
```

Fig 4 – The ATOM section of a PDB file

The atom serial number can be found in columns 7-11, whilst the atom name is placed in columns 13-16. The name of the corresponding residue is written in a 3-letters mode in columns 18-20. The chain in which the atom is localized is indicated in column 22 by one letter. In columns 23 to 26 is defined the residue sequence number. Finally, the orthogonal coordinates for X, Y and Z are in columns 31-54. These coordinates are in Angstroms.

Understand and making use of this information represents an essential step in the understanding of SAPs using structural information. Indeed, the Cartesian coordinates give the exact positions of the atoms in a three-dimensional space in the corresponding structure, it is thus possible to define the context of a variant in the 3D space.

1.3 Information aiding in the understanding of the relationship between SAPs and diseases

While the Swiss-Prot variant pages strive to provide essential information for the understanding of SAPs, some crucial information is still not available.

The conservation score has already been shown to represent a powerful sequence parameter for the prediction of the effect of mutation (Yip *et al.*, 2006). Indeed, mutations localized in very conserved positions were intolerant. This is hardly surprising as conserved regions often represent important structural and functional domains, and a mutation in these regions would mostly affect the protein's structure and function. The inclusion of this parameter in any web services that attempt to relay information on SAPs is thus important.

In terms of structural parameters, few have been systematically studied and none has yet been found to have similar predictive power as the conservation score. The chief characteristic of proteins that enables them to carry out their diverse cellular functions is their ability to bind other molecules specifically and tightly. Proteins can bind to other proteins, to DNA, as well as to small molecules. This binding ability is mediated by the three-dimensional structure of the protein that defines the binding site, and by the chemical properties of the surrounding amino acids. In this respect, these features represent promising structural parameters for the understanding of SAPs' relation to diseases.

The property of residues involved in *protein-protein interaction* has already been explored. One research group tried to correlate residues involved in protein-protein interactions and diseases (Keskin, *et al.*, 2005). Actually, they proposed that binding sites have preferred organizations. First, residues which have been found to be conserved in a multiple structure or sequence alignment (commonly called 'hot spots') cluster within densely packed 'hot regions'. Within these well compact organized regions, they form networks and interactions. Therefore, hot spots located within a hot region contribute cooperatively to the stability of the complex. Mutations could be able to broke this stability and prevent the interaction to occur, leading to putative disease-associated variants.

The local environment around a SAP is also essential for understanding the effect of surrounding residues (e.g. the proximity of a binding site or a chemically modified residue) on disease relationship. The local environment can be defined as the set of all residues having at least one atom within a sphere of radius *r* centered on each SAP atom.

In order to render this information available to users and to understand the significance of SAPs in these regions, it is essential to correctly map not only the variants themselves onto their known respective three-dimensional structures, but also map all UniProtKB/Swiss-Prot annotated features (ACT_SITE, BINDING, NP_BIND, MOTIF, DOMAIN etc.) that are important for the interpretation of the SAPs onto the structure. Conversely, residues involved in protein-protein interactions that are identified on the 3D structures should be mapped back to Swiss-Prot sequences to see if they coincide with variant positions. The mapping PDB-Swiss-Prot sequence is created within the Swiss-Prot group (SSMAP) (David, 2007) and will be used in this study.

1.4 Aim of the project

This study is divided into two parts. The first objective is to improve the ModSNP database by computing and storing more data related to SAPs. An improved version of the Swiss-Prot Variant Web Pages will be developed to provide more sequence and structural information on SAPs. The second objective is to evaluate the predictive performance of one of the newly computed parameters, namely the *conservation score*, over all Swiss-Prot variants.

2 Methods

2.1 Data collection

All Single Amino acid Polymorphisms are classified into three distinct categories in Swiss-Prot; a variant is qualified as "Disease" if it has been proven to directly cause the disease; a variant is a "Polymorphism" if there is no proof of its association with a disease; and a variant is classified as "Unclassified" if its association to disease is not yet clearly defined. This information is available via the Swiss-Prot Variant web pages and is also recorded in the ModSNP database (Yip, *et al.*, 2004). The collection of SAPs was performed from this database (Release 52.2 3 Apr. 2007). We focused on the three categories previously quoted: "Disease", "Polymorphism" and "Unclassified". We obtained a dataset that contained 28888 variants in 5234 human entries; 14559 were disease-associated (50.4% of total variants), 12217 were polymorphisms (42.3% of total variants) and finally 2112 (7.3% of total variants) where considered as unclassified.

2.2 Calculation of structural and sequence parameters

We investigated one sequence and two structural attributes. The first one is the conservation score. The other two are the local environment around the variant and the involvement of the variant in interface residues.

2.2.1 Residue conservation

The calculation of the conservation score of all variants available in ModSNP was performed twice. Initially, the score was calculated using orthologous sequences from the Orthologs Matrix Project (OMA) project (Schneider et al., 2006). The OMA project is based on the comparison of genomes to identify orthologous genes. The conservation score was also computed using the orthologous sequences identified by a simple BLAST search.

2.2.1.1 Workflow with OMA data

First, sequences in ModSNP that have at least one variant are compared with sequences belonging to an OMA group, i.e. they belong to a group of orthologs. These groups are calculated by the OMA project and are stored in a fasta format file:

```
…
> OMA08137 | TETNG06552 | GSTENP00003612001
sequence 1
> OMA08137 | FUGRU16101 | SINFRUP00000152198
sequence 2
> OMA08137 | CIOSA11985 | ENSCSAVP00000006969
sequence 3
> OMA08138 | CHICK03023 | ENSGALP00000033975
```

```
sequence 4
> OMA08138 | BRARE08743 | ENSDARP00000073595
sequence 5
> OMA08138 | HUMAN18994 | ENSP00000264010
sequence 6
> OMA08138 | MOUSE24683 | ENSMUSP00000005841
sequence 7
…
```

Fig 5 – Part of the OMA file that contains orthologous sequences. A human sequence is highlighted in yellow.

The *OMA* term followed by a number represents the group of orthologs to which belongs the associated sequence. For instance, the sequences 1 to 3 in figure 5 belong to the group of orthologs OMA08137, whilst sequences 4 to 7 belong to the group OMA08138.

As the file is huge, only human sequences were first extracted from this file. This allows limiting the space search at the time of the comparison with ModSNP sequences. For each ModSNP entry having at least one variant, we first searched if the associated sequence matched with a human sequence previously extracted. If the sequence matched, all sequences belonging to the same orthologs group as the human sequence were retrieved from the entire OMA file and recorded in a file named `<accession_number_of_the_target_sequence.fas>`.

```
> OMA131782_HUMAN
MGGLEPCSRLLLLPLLLAVSGLRPVQAQAQSDCSCSTVSPGVLAGIVMGDLVLTVLIALAVYFLGRLVP
RGRGAAEAATRKQRITETESPYQELQGQRSDVYSDLNTQRPYYK
> OMA131782_MOUSE
MGALEPSWCLLFLPVLLTVGGLSPVQAQSDTFPRCDCSSVSPGVLAGIVLGDLVLTLLIALAVYSLGRL
VSRGQGTAEGTRKQHIAETESPYQELQGQRPEVYSDLNTQRQYYR
> OMA131782_CANFA
MGVLGPSNRLLFLPLLLSMGGFSPIQAQDGCNCPVVSPGVLAGIVLGDLVLTLLIALAVYSLGRMFPRG
RGSVEAVTRKQRITETESPYQELQGQRSDVYSDLNTQRPYYK
> OMA131782_RATNO
MGAPEPSWCFLFLPVLLTVGGLSPVQAQSDNYPGCECSSVSPGVLAGIVLGDLVLTLLIALAVYSLGRL
VSRGRGTADGTRKQHMAETESPYQELQGQRPEVYSDLNTQRQYYR
> OMA131782_BOVIN
MEGLRPSDRLLSLLLTVGGLSLVLAQSECNCSSVSPGVLAGIVLGDLMLTLLIALAVYYLGRLVPRGRG
ATEAVTRKQHIPETESPYQELQGQRTDVYSDLNTQRPYYK
> OMA131782_PANTR
MGGLEPCSRLLLLPLLLAVGGLRPVQAQAQSDCSCSTVSPGVLAGIVMGDLVLTVLIALAVYFLGRLVH
RGRGAAEAATRKQRITETESPYQELQGQRSDVYSDLNMQRPYYK
> OMA131782_MONDO
MRKGRTPTLGLALVLALLLTPGGESGWPRRSSAPCPAVGPGVLAGIVLGDLALTILIALAVYYLGRLTP
SGRNASEGKRRPEGMAEAESPYQELQDHRSDVYSDLTPQPGFYR
> OMA131782_MACMU
MGGLEPCSRLLLLPLLLAVGGLRPVQAQAQSDCSCSTVSPGVLAGIVLGDLVLTVLIALAVYFLGRLVP
RGRGAAEAATRKQRITETESPYQELQGQRSDVYSDLNTQRPYYK
> OMA131782_LOXAF
MVGIGPSGRLLLLPLLLAVGGLSPAQAQSECKCATVGPGVLAGIVLGDLLLTLLIALAVYSLGRLVPQK
RGAAETVTRKQRVNETESPYQELQGQRSDLYSDLNTQKPYYK
```

Fig 6 – The file O43914.fas. The human sequence is highlighted in yellow. All these sequences belong to the same OMA group. The group of orthologs is OMA131782.

These sequences must then be aligned. The alignment software *mafft* (Katoh, *et al.*, 2002) were used to address this issue, because it aims to produce good alignments and is faster than t-coffee.

Only alignments with at least five sequences were considered. Indeed, too few sequences may fail to provide an informative alignment. Each alignment was used as input of SCORECONS (method: valdar01) to compute the conservation score (Valdar, 2002). SCORECONS is a program that quantifies residue conservation in a multiple sequence alignment. Given a multiple sequence alignment file, it calculates the degree of amino acid variability in each column of the alignment and returns this information to the user. We used the option "focus" of the program that allows calculating the conservation score with respect to one specific sequence in an alignment. That means that the score is computed only in columns in which the sequence of interest has a residue. In this case, the score of each position of the human sequence was calculated.

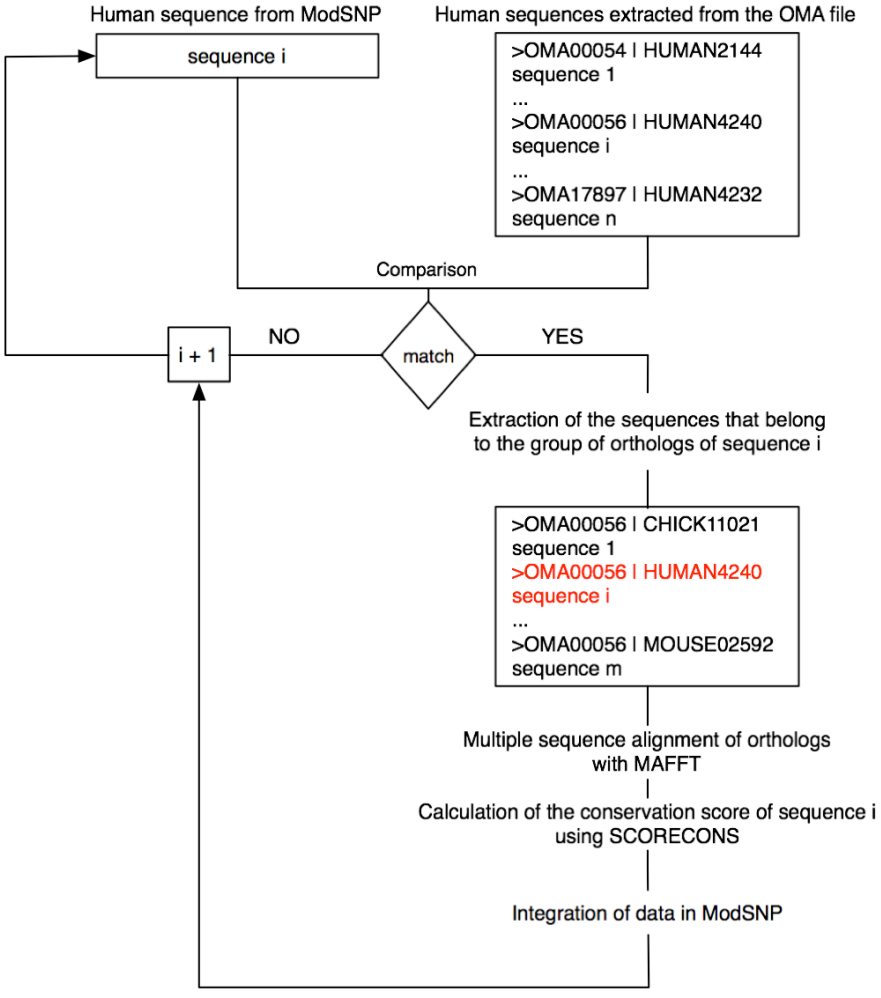


Fig 7 – Workflow of the conservation score calculation when OMA data are used

2.2.1.2 Workflow with BLAST data

As not all sequences stored in ModSNP have a corresponding OMA entry, it has been decided to complement the data by computing the conservation score using BLAST search. The way to reach the desired objective is similar to that used for OMA. However, some differences should be noted. First, each ModSNP sequence was now run against UniProtKB/Swiss-Prot database with *blastall*. The result was then parsed and analyzed. Alignments of less than five sequences were not considered, as for the OMA method. Moreover, to ensure that only orthologous sequences were used for alignment, only the sequences that have the same Swiss-Prot ID prefix (e.g. TYOBP-XXX) as the sequence of interest were kept. Once this task finished, these sequences were aligned with *mafft*. Finally, the conservation score was calculated with SCORECONS, as previously.

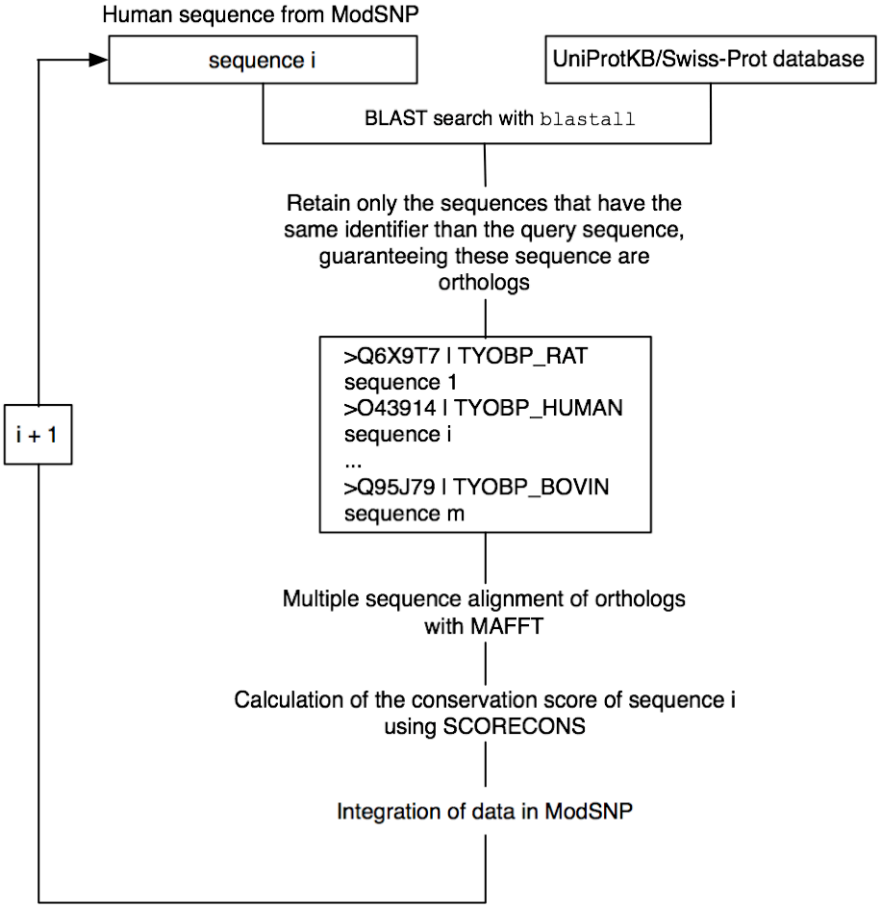


Fig 8 – Workflow of the conservation score calculation when BLAST data are used

2.2.1.3 Storage of related data in ModSNP database

Once the conservation score for a sequence calculated, we proceeded to the integration of the data into the ModSNP database. To address this issue, two

tables were initially created. The first one, called *sp_align* (Appendix – table 1), stored all the information related to an alignment, while the second one, called *conserv_score* (Appendix – table 2), recorded data concerning the conservation score of each variant present in the sequence of interest. This individual score was obtained by extracting the score at the variant's position in the corresponding sequence.

2.2.2 Determination of the best representative 3D structure for a Swiss-Prot variant

When we want to structurally analyze a mutation in a Swiss-Prot sequence, it is indispensable to associate a three-dimensional structure to that mutation. An algorithm was developed to find the best 3D structure for a given variant (Loichot, 2007). We considered that a "best structure" had to satisfy the following criteria:

1. To have a sufficiently good resolution (3 Angstroms)

2. To have the largest possible coverage with respect to the Swiss-Prot sequence

3. To cover the variant site

4. To have a sufficiently distinct structural environment around the variant. This criterion was used to eliminate similar structures to avoid redundancy.

Four filters were used in the algorithm, each of them filtering the data according to one of the points mentioned above. The order of these filters must be respected. The filters were classified from the most general to the most specific. The filter resolution made it possible to keep only the most precise structures, whereas the filter based on the variant's environment kept only the most complete structures in the zone of the variant.

The best representative structures were recorded in the *relftstructure* table (Appendix – Table 4) of ModSNP database. They will be used in all cases where a 3D display of the structure is necessary.

2.2.3 Structural local environment of the variant

We made the hypothesis that the three-dimensional environment around a single amino acid polymorphism site may have an influence on whether a SAP may associate to diseases. Therefore, we defined a new parameter, the structural local environment of the variant. A residue was considered as a neighbor of the variant when at least one of its atoms fell within a given radius around one of variant's atom (Fig 9) in its associated structure.

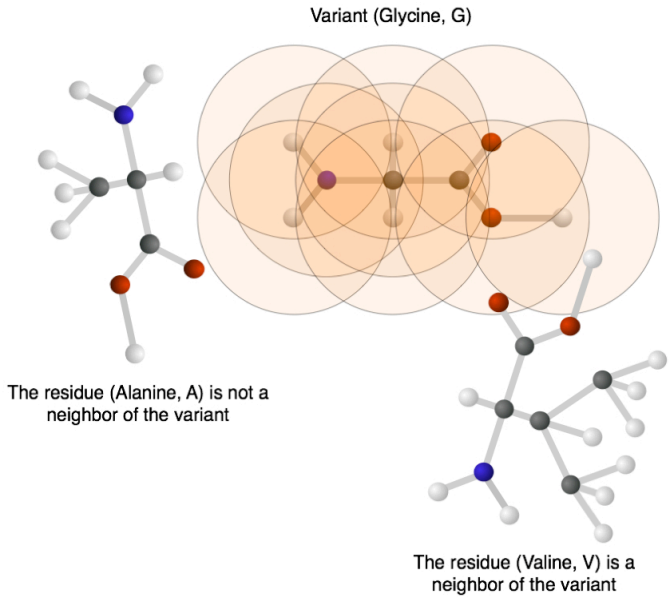


Fig 9 – The valine is a neighbor of the variant because at least one of its atoms falls within a given radius (here, a 6-angstroms radius defined by orange circles) around one atom of the glycine residue. The alanine is not in SAP environment.

In this study, the local structural environment of a given variant was calculated for a 6-angstroms radius. It allowed the identification of all the residues that were in proximity of the variant within a maximal distance of 6 angstroms.

2.2.3.1 Workflow

A perl script (David, 2006) was run on each best representative structure for each variant in the *relftstructure* table in ModSNP. This script required four parameters:

1. The *radius* for which to compute the structural local environment. As said before, a radius of 6 angstroms has been chosen.

2. The *position of the variant* in its representative three-dimensional structure. [1]

3. The *name of the chain* in which the variant is localized.

[1] The second parameter could not be derived directly from the variant itself. In other terms, it was not possible to find the structural position of a given variant directly from its sequence position. To solve this problem, a new table *mapping* (Appendix – Table 3) was constructed. This table stored the correspondence between a sequence position and a structural position for each variant having a related three-dimensional structure, i.e., a variant present in *relftstructure*. The mapping data for each position of the sequence were derived from the *ssmap* database developed in the Swiss-Prot group.

4. The *PDB file* of the three-dimensional structure.

This script parsed the given PDB file and provided as output the interactions between atoms. An interaction between the variant and a particular residue was defined by:

```
neighbor_residue -> variant_atom variant_atom_serial,;neighbor_atom
neighbor_atom_serial,;distance_of_the_interaction
```

Fig 10 – An excerpt of the output of the script describing an interaction between a variant and a neighbor residue

This structure of data can be read as "The atom *variant_atom*, identified by the atom unique serial number *variant_atom_serial* interacts with the *neighbor_atom* with the atom serial number *neighbor_atom_serial*. This *neighbor_atom* is part of the *neighbor_residue* residue. The distance between these two atoms is *distance_of_the_interaction*".

All interactions between the variant and another particular residue were separated by ":&". For instance:

```
neighbor_residue1 -> variant_atom variant_atom_serial,;neighbor_atom
neighbor_atom_serial,;distance_of_the_interaction:&
neighbor_residue2 -> variant_atom variant_atom_serial,;neighbor_atom
neighbor_atom_serial,;distance_of_the_interaction
```

Fig 11 – An excerpt of the output of the script describing two interactions between the variant and two different neighbor residues

Often, multiple SAP atoms interact with numerous other atoms of a particular residue. In this case, the script differentiates each interaction by the separator "<>", as shown below:

```
neighbor_residue -> variant_atom1 variant_atom1_serial,;neighbor_atom1
neighbor_atom1_serial,;distance_of_the_interaction<>variant_atom1
variant_atom1_serial,;neighbor_atom2
neighbor_atom2_serial,;distance_of_the_interaction<>variant_atom2
variant_atom2_serial,;neighbor_atom1
neighbor_atom1_serial,;distance_of_the_interaction<>variant_atom2
variant_atom2_serial,;neighbor_atom2
neighbor_atom2_serial,;distance_of_the_interaction
```

Fig 12 – Multiple interactions between a variant and a neighbor residue

All this information was first parsed and the relevant data extracted. Only the residues whose atoms interacted with the variant were retrieved as well as the distances. If multiple interactions exist between a residue and the variant, as shown in figure 12, only the minimal distance was extracted. Once the structural environment was completely defined within 6-Angstrom radii, we defined three additional values of radii: 3, 4, and 5 (Fig 13). These were served as cut-off option for the 3D display (see below, section 3.1, fig 15).

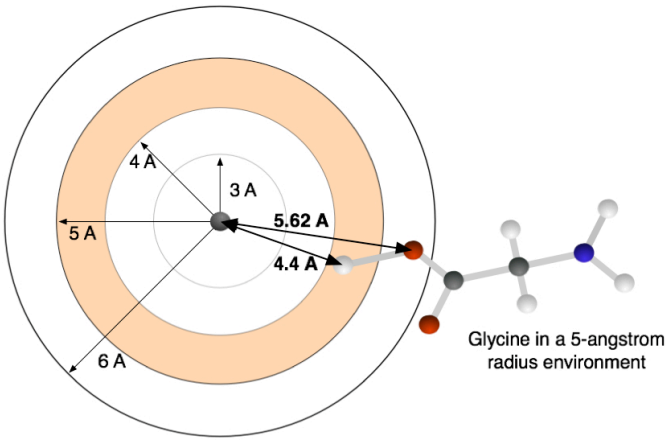


Fig 13 – A glycine around a SAP carbon atom (in black). Four radii are defined: 3, 4, 5 and 6 angstrom. The minimal distance (4.4 A) is used to define in which environment is localized the residue. Here, the residue is in a 5-angstrom radius environment (in orange).

All this information was then integrated to the ModSNP database, in a table called *environment* (Appendix – table 5).

2.2.4 Chain-chain interfaces

We considered a residue to be in a chain-chain interface when it is sufficiently close to another residue of another chain.

The computing of interfaces was not among the goals of this study. For further reading, please refer to the work of Gregory Loichot. All the information related to the interfaces was stored in the *interface* table in ModSNP (Appendix – Table 6).

2.3 Update of the parameters

All the parameters discussed till now are present in the tables mentioned in the Appendix. They must be updated regularly to assure the coherence of the data. Therefore, we developed a main script that controls the update process of these parameters. This script is called *modsnp_update_env.pl* and is launched during the ModSNP update process, i.e., every two weeks.

First, the data in the *mapping* table (Appendix – Table 3) is deleted and replaced by the new one. In the same way, the Swiss-Prot features listed in figure 1 are extracted and replaced in the table *sp_features* (Appendix – Table 7). Afterwards, the environments are calculated. It is important to notice that this step is carried out in an incremental way, meaning that only new environments are added into the table *environment* (Appendix – Table 4). This permits to gain time. Some additional tests are made to avoid redundancy and assure the integrity of data. For instance, we check if the position is still a variant position or if there is no modification at the structure level. Finally, the script that computes conservation scores is launched. The data is then replaced in the tables *sp_align* (Appendix – Table 1) and *conserv_score* (Appendix – Table 2). Concerning the update of the interfaces and the

computing of the best representative 3D structure, please refer to Gregory Loichot's report.

2.4 Web interface implementation

In this study, we improved the Swiss-Prot Variant Web Pages by providing more sequence and structural information to the user. Moreover, we created two additional pages. The first concerns the conservation score, whilst the other is about environments and interfaces. All these pages retrieve the information they need by accessing numerous tables in ModSNP.

2.4.1 The conservation score page

This page provides information on the conservation score of a given variant. It shows the alignment from which the conservation score was derived as well as the mapping of the scores on the variant's representative structure.

Concerning the alignment, we retrieved the multiple sequence alignment for the corresponding UniProtKB/Swiss-Prot sequence of the variant, as well as the scores of the sequence calculated from this alignment. From this data, we reconstructed the alignment with the sequence of interest at the top and with colored columns according to the degree of conservation of each position. The most conserved position (the position with the highest score in sequence of interest) is colored in red, whilst the most non-conserved position is colored in blue [2]. When a gap is present in the sequence of interest, the entire column is colored in grey.

The mapping of the scores onto the representative 3D structure was performed in two steps. First, we needed to establish the correspondence between sequence positions and the structural positions. The table *mapping* (Appendix – Table 3) was used to achieve this task. Then, we wrote to the scores directly in the PDB file. Indeed, the columns 61-66 of the PDB file contain values between 0 and 100, which represent the temperature factor of a given atom. This measure is generally used by softwares like Jmol to color the structure. Since the score is between 0 and 1, we multiplied it by 100 in order to comply with the scale required by the columns. This will allow us to color the residues on the structure, according to their conservation scores (see 3.1 fig 15).

Obviously, the mapping of the scores is made only when a representative structure for a given variant exists. Indeed, it may be that a best structure is not available for a given variant, in which case only the alignment is shown.

2.4.2 The 3D Display page: Interface residues and local environment

This page provides a view of the interface residues as well as residues located in the local environment of a given variant in a 3D manner. All the information on the interfaces is obtained by accessing the table '*interface*' (Appendix – Table 6). Concerning the environments, the table '*environment*' (Appendix – Table 5) is used to select all the residues present around a SAP.

[2] The Jmol colors were used (see Jmol documentation on http://jmol.sourceforge.net/jscolors)

We also developed a function that allowed checking for each residue in the environment of a given variant whether it corresponds to a Swiss-Prot feature or not. To address this issue, the corresponding sequence position of a given structural residue is first given by the table *mapping*. Then, the function checks in the table *sp_features* (Appendix – Table 7) whether the residue's position is part of a UniProtKB/Swiss-Prot feature.

2.4.3 The Swiss-Prot Variant Page

The new Swiss-Prot Variant Page was implemented on a server called *perceval* at the Swiss Institute of Bioinformatics. The page is reachable via the following address: http://perceval.isb-isb.ch/cgi-bin/variant_pages/get-sprot-variant.pl?VAR_XXXXXX. It can be accessed through the features (FT) lines from a UniProtKB/Swiss-Prot protein entry, too.

The new pages provide more structural and sequence information than the current ones. However, the access to these data varies depending on the features of the variant. Indeed, not all the variants have a conservation score. Also, 3D structures or models are not always solved. Therefore, we developed a flexible page able of showing the appropriate information according to available source data.

First, the residue conservation parameter is shown when a score could be calculated, either by the OMA method or the BLAST method, or both methods, and when the diversity of the alignment is more that 50%. In this case, a link to the conservation page is provided. Otherwise, it is mentioned that there is no conservation score available for the variant. The conservation score attribute is obtained from both the *sp_align* (Appendix – Table 1) and the *conserv_score* (Appendix – Table 2) tables.

Then, for information concerning structural data, the display of the Swiss-Prot variant page depends on the availability of 3D structures or models for the given variant. If there is no structure available (neither experimental or model), no additional information will be shown. If the variant has one or more representative 3D structures (retrieved from the table *relftstructure* (Appendix – Table 4) in ModSNP), additional information concerning the 3D structure itself would be shown and links would be provided to the 3D display page mentioned above. Finally, if there is no available experimental structure but homology models exists (information from the table 'models' in ModSNP), then the current 'Structural information on the variant' section will be displayed, i.e. only homology models will be accessible and no additional structural parameters calculated in this study will be available.

2.4.4 Technologies used

The technologies used to implement the web site are listed below:

1. All scripts were written in PERL language.

2. PostgreSQL was used to query ModSNP database.

3. CGI (Common Gateway Interface) perl module was used to generate the HTML code for the new Swiss-Prot Variant Web Pages.

4. Jmol was used to show structures in a three-dimensional space.

5. Javascript was needed to respect the interaction between the user and the web browser.

3 Results

3.1 Web interface

An updated Swiss-Prot Variant Web Page was created in this study in order to provide to the scientific community the new sequential and structural attributes computed in this work. Additional references retrieved by text-mining for each variant are also made available via the new web interface (please refer to the report of Nathalie Lachenal). The page is now subdivided into seven main sections, instead of five previously:

1. General information
2. Information on the variant
3. Sequence features
4. Structural features
5. References for the variant
6. Additional references for the variant
7. Cross references for the variant

Sections 1, 2, 5 and 7 are not changed compared to the old version. Additional information is provided in sections 3, 4 and 6. An example of the new Swiss-Prot Variant Web Pages is available on http://perceval.isb-sib.ch/cgi-bin/variants/get-sprot-variant.pl?VAR_000399 (Fig 14).

Sequence features

Location on the sequence

334 KKLADMLTEITLGLHACLQL G RLKDQDKAAPEMVSLLKRNN 374
↓
S

Protein features in neighborhood

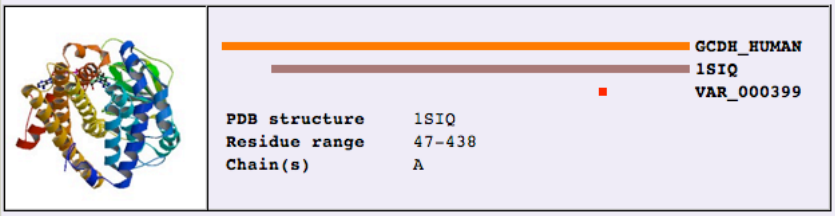
Key	From	To	Length	Description
CHAIN	45	438	394	Glutaryl-CoA dehydrogenase
HELIX	330	358	29	

Residue conservation

From	Score	Diversity	Nb. of sequences in alignment	
oma	0.929	87.34 %	56	View oma alignment

Structural features

3D structure(s)



Protein features in structural neighborhood	View local structural neigborhood of variant
Interface(s) involvement	View interface(s)
Surface accessibility	To come
Physico-chemical property	Change from glycine (G) to small size and polar (S)

References for the variant

[1] REVIEW ON VARIANTS.
DOI=10.1002/(SICI)1098-1004(1998)12:3<141::AID-HUMU1>3.0.CO;2-K; MEDLINE=98375689; PubMed=9711871; [NCBI, ExPASy, EBI, Israel, Japan]
Goodman S.I., Stein D.E., Schlesinger S., Christensen E., Schwartz M., Greenberg C.R., Elpeleg O.N.
"Glutaryl-CoA dehydrogenase mutations in glutaric acidemia (type I): review and report of thirty novel mutations."
Hum. Mutat. 12:141-144(1998)

Additional references for the variant

View additional references for the variant

Fig 14 – The new four sections of the Swiss-Prot Variant Web Page.

In section 3 “Sequence features”, the conservation score of the variant is added. Here, the user has the possibility to obtain the conservation score of the variant, the diversity of the alignment from which the score is derived, the number of the sequences aligned, as well as the method used (OMA or BLAST). Moreover, a link to an external page is provided. This page shows the entire alignment where all the columns are colored according to the degree of conservation (Fig 15). If a representative 3D structure exists for the variant, the mapping of the conservation score of each residue onto the structure is also shown. The user can then decide to view only the variant by selecting the checkbox “Show only variant” or all the residues of the 3D structure by checking “Color by conservation score”.

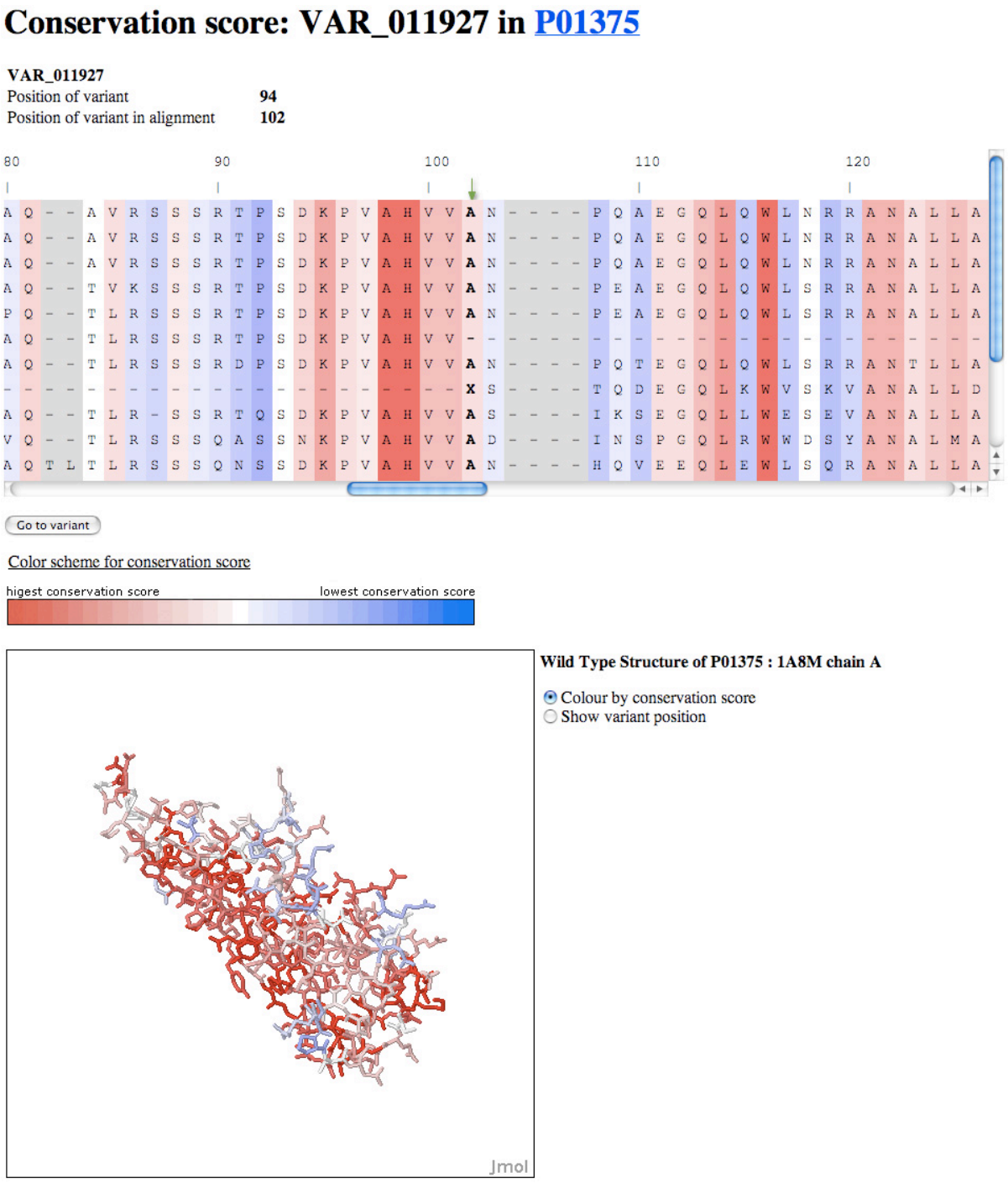


Fig 15 – The conservation score page indicates the position of the variant in the alignment. It allows the user to see the entire alignment with columns colored according to the degree of conservation. If a representative 3D structure exists, the conservation score of the sequence of interest (the first in the alignment) is mapped onto it.

In section 4 'Structural Information on Variant', several new features are present. First, the best representative 3D structure for a given variant is shown. This information is available only when an experimentally resolved 3D structure exists. Each structure is represented by a little image, with a link to the corresponding PDB entry. The PDB code, the residue range and the chain(s) involved are indicated, as well as a linear view of the PDB sequence, the Swiss-Prot sequence and the variant's position.

In the same section, the row 'Protein features in Structural neighborhood' provides a link to access the page containing information on the structural local environment of the variant. This page shows all the residues that are localized in a given radius around the variant in a given PDB chain. The user can choose this radius. This varies between 3 to 6 angstrom, where 3 is the default value. The possibility to visualize the whole PDB chain is also provided. The main functionality of the page resides in the viewing of all the Swiss-Prot features (Fig 1) involved in the environment for a given radius. Indeed, a residue in the environment is colored in blue when it represents a Swiss-Prot feature. Otherwise, the residues are in red, except for the variant, which is shown in green (Fig 16).

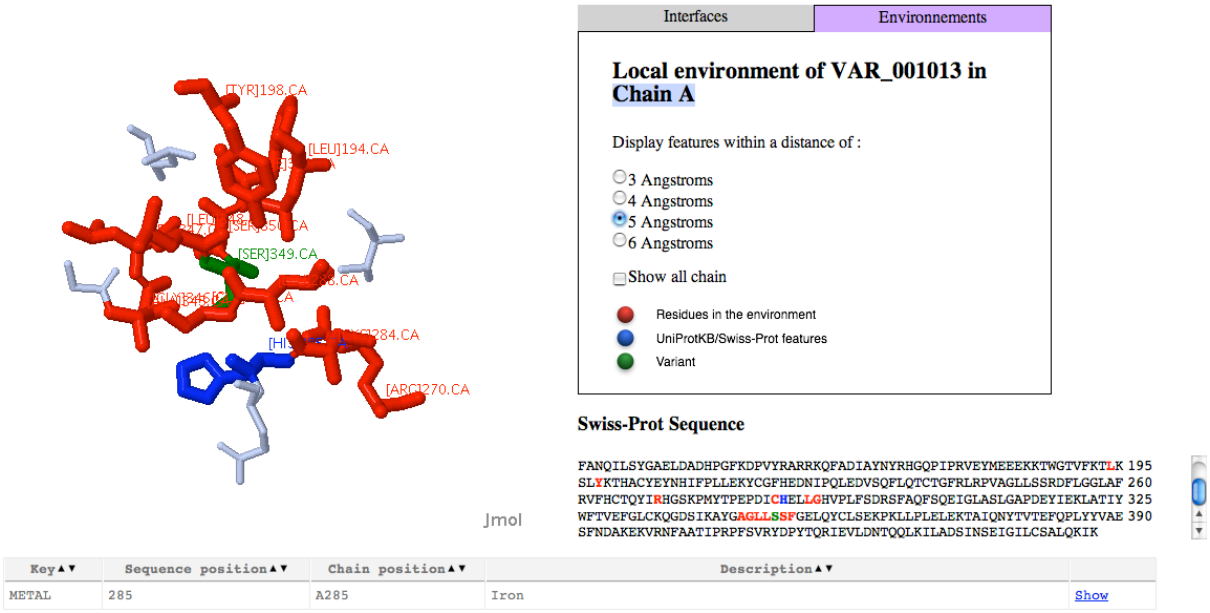


Fig 16 – The section "Environment" of the 3D display page indicates the residues surrounding the variant in a 5-angstrom radius. The residues that represent a feature are colored in blue, whilst the others are in red. The variant is in green. Each feature with its description is shown in the bottom of the structure and a link "Show" allows the user to visualize the specific feature on the 3D structure. The location of the displayed structural residues on the UniProt sequence is also shown.

The interfaces between chains of a given protein are also provided by a link in the "Structural features" section. The corresponding page has the same web address of the previous one, but in this case, the user arrives directly to the "Interface" section of the page, unlike previously where the "Environment"

section is shown. Here, two PDB chains involved in the interaction are present. These chains can belong to the same protein or belong to different proteins, depending on the PDB entry chosen for the display. All the residues of a chain that interact with another chain are shown. That means that all residues that are localized in an interface are colored and can be viewed. The user can choose which interface to be shown, as well as the method employed (Carbon alpha or Wan der Waals) (Loichot, 2007). Similar to the display of structural local environment, all the chains involved can be viewed in order to have a global three-dimensional view of the context. Finally, this page indicates whether a variant is implicated or not in the interface (Fig 17).

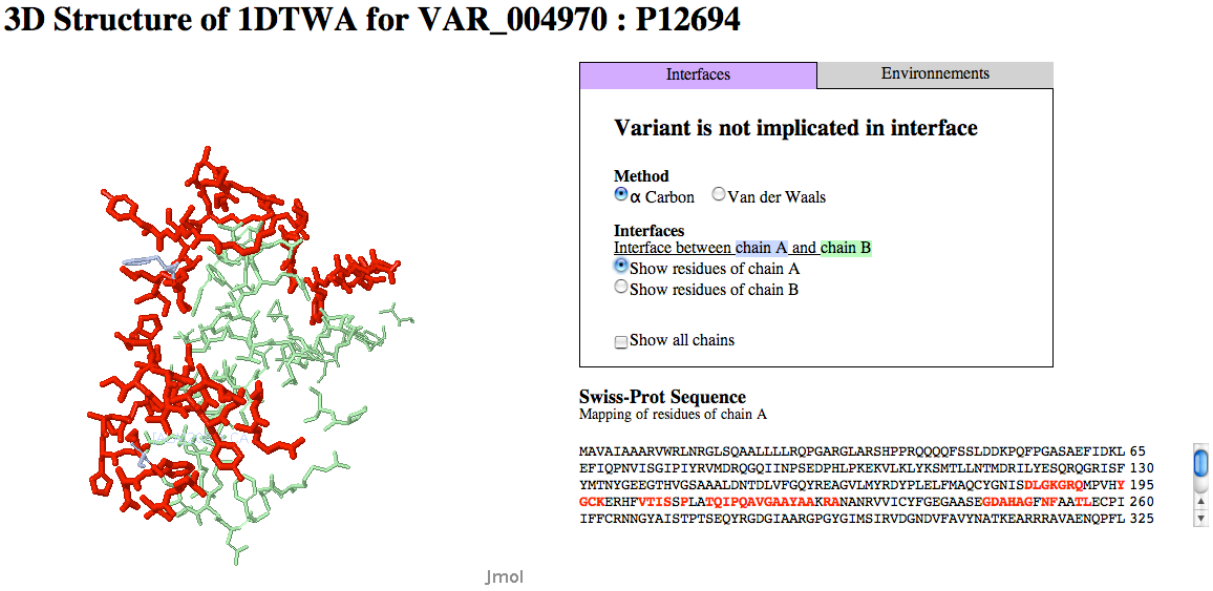


Fig 17 – The section "Interfaces" of the 3D display page indicates the residues implicated in a chain-chain interaction. The user can choose to visualize the residues issued either from the Alpha-carbon method or the Van der Waals method (Loichot, 2007). It can also be decided which residues of which chain to visualize (here the residues of chain A are shown in red. These residues interact with the chain B). The mapping of these residues onto the corresponding Swiss-Prot sequence is also shown. Finally, this page tells whether the variant is implicated in the interface or not.

In section 6 'Additional references for the variant', new variant-related articles are provided through the hyperlink "View additional references for the variant". These references were retrieved by text-mining (Yip, *et al.*, 2007).

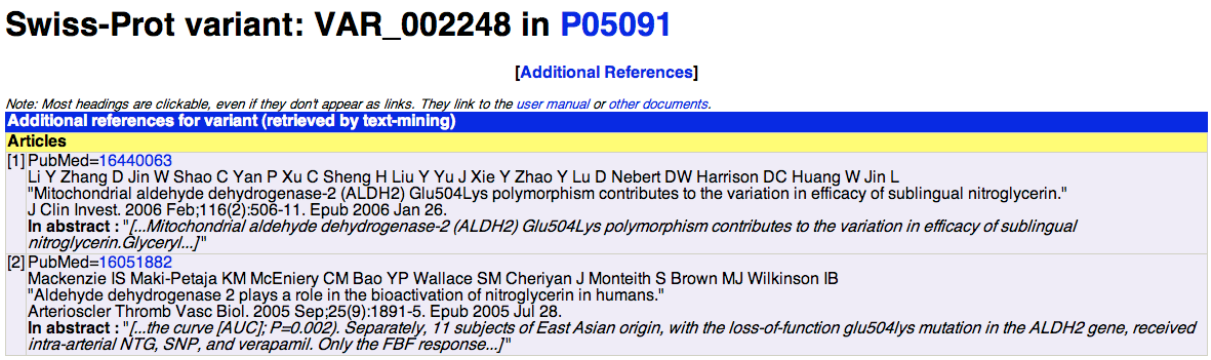


Fig 18 – The 'Additional references for variant' page indicated all the new references related to the variant.

3.2 Conservation score predictive power

In this study, the predictive power of the conservation score was evaluated. The conservation score for each variant present in ModSNP was calculated using SCORECONS from an alignment of multiple species (including human). We used two methods to calculate the conservation score: OMA and BLAST.

We analyzed the distribution of all the scores in order to find a correlation between the score and the classification of the variant (disease, polymorphism, unclassified). This was done for both methods.

3.2.1 Calculation of the score of conservation using OMA and BLAST

Among the 28888 human variants available in UniProtKB/Swiss-Prot (Release 52.2 3 Apr. 2007), we found 8116 disease-associated variants for which a conservation score could be calculated with the OMA method. We have also computed the conservation score for 5604 polymorphisms and 1173 unclassified variants using this method. For BLAST method, scores were computed for 8421 disease-associated variants, 4099 polymorphisms and finally 1598 unclassified variants (Fig 19).

	SP	***OMA***	***BLAST***	***None of the methods***
Total	28888	14893 (51.55%)	14118 (48.87%)	7690 (26.62%)
Disease	14559	8116 (28.1%)	8421 (29.15%)	2765 (9.57%)
Polymorphism	12217	5604 (19.4%)	4099 (14.19%)	4637 (16.05%)
Unclassified	2112	1173 (4.06%)	1598 (5.53%)	288 (1%)

Fig 19 – Number of variants having a conservation score. The percentages represent the proportion of variants found among the 28888 available

We proceeded to see how the BLAST procedure complemented that of OMA. For this, we computed the number of scores computed only by BLAST and not by OMA.

	SP	*BLAST*	*Only BLAST*
Total	28888	14118 (48.87%)	6305 (21.82%)
Disease	14559	8421 (29.15%)	3678 (12.73%)
Polymorphism	12217	4099 (14.19%)	1976 (6.84%)
Unclassified	2112	1598 (5.53%)	651 (2.25%)

Fig 20 – Number of variants found only by the BLAST method (column 4). The percentages represent the proportion of variants found among the 28888 available

The table (Fig 20) showed that overall almost 22% of the variants present in UniProtKB/Swiss-Prot were not identified by OMA method. Therefore, BLAST could represent a good complementary method to increase the coverage of the dataset.

3.2.2 Distribution of diversities using OMA and BLAST

Before analyzing the conservation score predictive power, it is interesting to assess the quality of the alignments produced either by the OMA method or the BLAST method. Therefore, we analyzed the distribution of diversities of all the alignments.

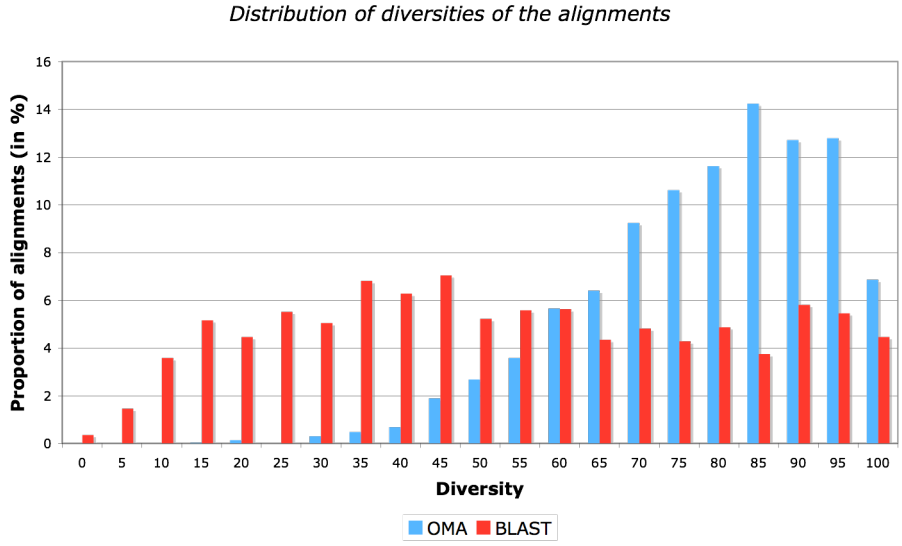


Fig 21 – Distribution of diversities using either the OMA method or the BLAST method. The Y-axis represents the percentage of alignments.

As shown in figure 21, BLAST produced both informative and uninformative alignments. The OMA method, on the other hand, resulted in more alignments with better diversity. The mean diversity of the alignment from sequences of OMA was close to 80%, while the mean diversity of the alignment from BLAST

was only about 50%. This confirms the robustness of the OMA project and thus the good choice for the using of OMA data, in parallel with BLAST data.

We also decided to measure the quality of the additional information concerning all the variants found only by the BLAST method. To address this issue, we extracted all the diversities of the corresponding alignments:

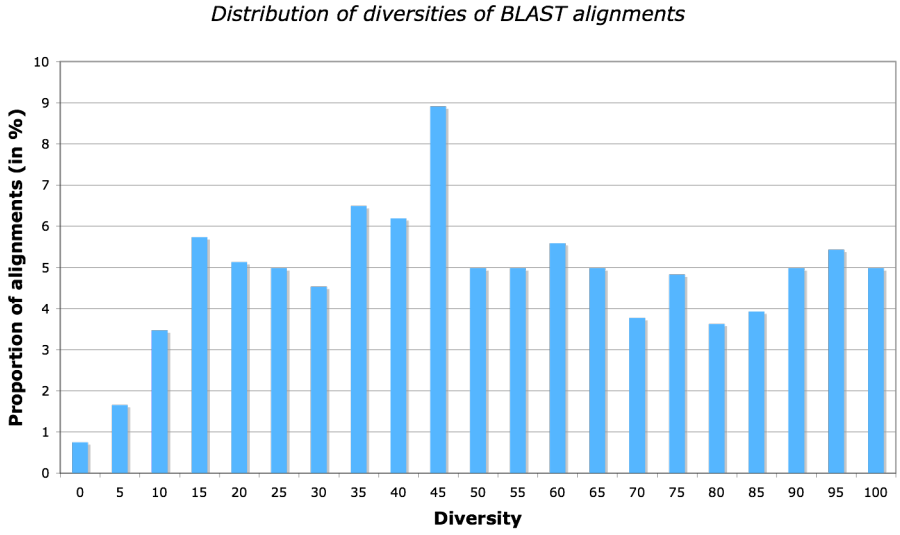


Fig 22 – Distribution of diversities using the BLAST method

The figure 22 showed that the quality of the alignments was variable. It was decided that in order to provide the users with only good quality information, a cut-off of 50% diversity was implemented in the web page. I.e. only conservation score issuing from an alignment of more than 50% diversity was shown.

3.2.3 Distribution of conservation scores using OMA method

We analyzed the distribution of conservation scores among the three categories of variants (disease, polymorphism, unclassified). We chose to use the conservation scores calculated from OMA data as they were of better quality (Fig 21).

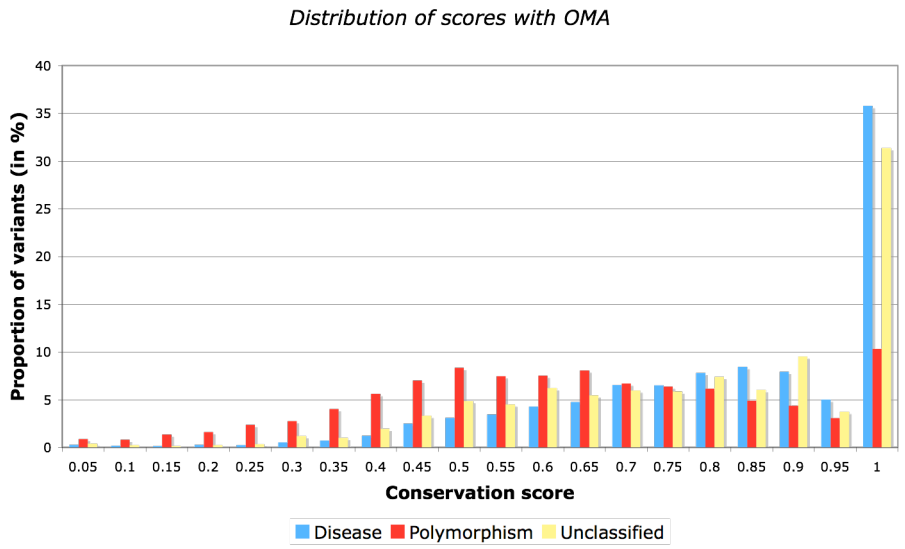


Fig 23 – Distribution of scores for the three categories: Disease, Polymorphism and Unclassified. The Y-axis represents the percentage of variants in the corresponding category.

It was found that the number of disease-associated variants having a score greater than 0.5 was 7615. This corresponded to 94% of the disease set (Fig 23). And for 35% of the disease-associated variants (2907 out of 8116), the conservation score was between 0.95 and 1. The result agreed with previous observation that the disease-associated variants tended to have a high conservation score (Yip, *et al.*, 2006).

Concerning the polymorphisms, of the 5604 variants for which a score could be calculated, 73% have a score greater than 0.5 (Fig 23). Here, the distribution followed almost a normal law. However, the proportion of scores comprised between 0.95 and 1 was higher than elsewhere. Actually, it represented 10% of all polymorphisms (580 variants of the 5604 available). While this percentage was still significantly lower than that of the disease-associated set, their number was still non-negligible.

Finally, the variants in the category "unclassified" were analyzed. We found almost the same trend as the disease-associated variants distribution. Following the previous observation, we can speculate that the unclassified variants are mainly disease-associated variants for which no or little annotation are available yet for their classification.

The scores shown here were issued from all the alignments. That means they had a diversity varying between 0 and 100%. We decided to analyze the influence of the diversity on the scores. Indeed, a low diversity would mean that the sequences in the alignment were too closely related. This would show very conserved regions in sequences and could bias the results. We decided thus to discriminate against the low diversity alignments. Only the alignments that had a diversity of greater than 70% were retained. The total number of variants with a conservation score calculated using an alignment of > 70% diversity was shown below:

	SP	***OMA***	***BLAST***	***None of the methods***
Total	28888	10966 (37.96%)	4371 (15.1%)	16020 (55.45%)
Disease	14559	5893 (20.4%)	2592 (8.97%)	7512 (26.1%)
Polymorphism	12217	4084 (14.14%)	1073 (3.71%)	7574 (26.21%)
Unclassified	2112	989 (3.42%)	706 (2.44%)	934 (3.23%)

Fig 24 – Number of variants with a computable conservation score when the diversity of alignment is greater than 70%. The percentages represent the proportion of variants found among the 28888 available

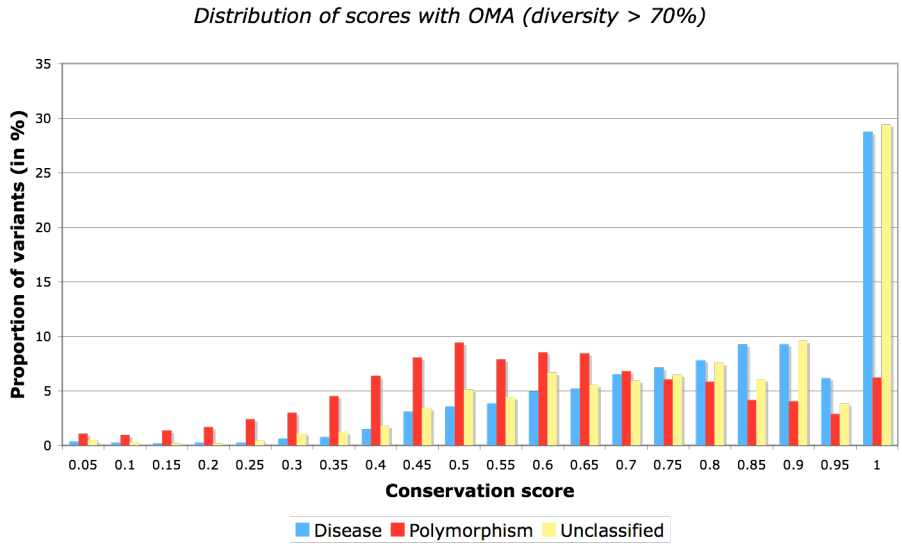


Fig 25 – Distribution of scores after applying the diversity (70%) constraint for the three categories of variants: Disease, Polymorphism and Unclassified. The Y-axis represents the percentage of variants in the corresponding category.

Compared to the previous distribution, the new one follows a similar tendency. Therefore, the hypothesis formulating that a disease could be associated with a high score seems to be verified.

As expected, there were appreciably fewer scores between 0.95 and 1, and more scores next to 0.5 for polymorphisms. The interesting thing would be to characterize the variants that have a very high score. Indeed, if we follow the assumption previously posed, we could say that these variants are polymorphisms that could be potentially linked to diseases. More categorically, these polymorphisms might be putative functional polymorphisms.

4 Discussion

In this project, we have upgraded the Swiss-Prot Variant Web Pages to provide additional structural and sequence information for a given variant. Three new web pages were developed to give detailed information, as well as 3D structural display (when available), on the conservation score, the local environment and the interface involvement of the variant. All these pages are accessible via the hyperlinks provided in the Swiss-Prot Variant Web Pages. Concerning the conservation score page, the principal drawback is the time of generation of the alignment. Indeed, the alignment can be relatively slow to load when a lot of sequences are considered. Moreover, the loading speed is greatly affected by the length of these sequences. The reason is that a large number of HTML tags are generated. To overcome this problem, it would be necessary to better exploit the features of Cascading Style Sheet (CSS) by writing specific classes for each color. Concerning the pages dealing with protein-protein interfaces, only data coming from the Protein Data Bank (PDB) was currently used. It might also be important to incorporate Protein Quaternary Structure (Henrick and Thornton, 1998) data in order to increase the number of physiological chain-chain interactions that might have been missed by analyzing only PDB entries. In the near future, it is expected that the new pages will integrate new structural parameters. The surface accessibility for instance, indicates whether the variant is exposed on the surface of the protein or if it is buried into the core of the protein. The presence of hydrogen bonds or disulfide bridges could be also important features because variations in such sites could modify the stability of proteins and have an effect on their function. It is expected these pages will be made available to public via the ExPASy server. Recent web technologies such as AJAX will be used to significantly cut down the bandwidth consumption and further improve the interactivity between the pages and the end user.

In addition, we have developed and evaluated two automatic methods (OMA and BLAST) to compute the conservation score of human sequences present in UniProtKB/Swiss-Prot. Our results showed that data issued from the OMA project were of good quality and offered a high number of informative alignments. The use of BLAST was however necessary to assure the coverage of the datasets. Nevertheless, BLAST required more computational resources and took almost 10 hours. Considering that the number of sequences significantly increases in ModSNP, it might be judicious to exclude the BLAST method in the future when the OMA project is completed. Another possibility would be to parallelize the BLAST searches on multiple processors in order to guaranty the coverage of the datasets and to accelerate the process. When evaluating the conservation score for the three variant categories (disease-related, polymorphisms, unclassified), we confirmed previous results on the predictive power of this score for disease-associated variants. Our results also suggested that most unclassified variants were probably disease-associated. In the near future, a systematic study of the newly computed structural data should also be carried out so that their relevance and predictive power can be evaluated.

Acknowledgments

I would like to thank all representatives of the Master in Bioinformatics and Proteomics to have given me the opportunity to enlarge my knowledge in the domain of Biology. In particular, I would like to thank Dr. Lina Yip and Fabrice David who were always present to guide me through this project, as well as Dr. Patricia Palagi without whom this master would not have been possible. It has been a real pleasure of working with these people and it will be irrefutably an experience that I won't be able to forget.

Appendix

Table 1 – sp_align

This table contains four fields:

1. The *isoid* of the sequence for which a score was calculated
2. A *methodid* that defines the method employed to compute the score of the sequence

   0 – OMA

   1 – OMA
3. The *diversity* of the alignment
4. The *number of sequences* implicated in the alignment

isoid	methodid	diversity	nb_seq
Q9NQG7-1	0	91.18	16
P55809-1	0	85.06	16
P10323-1	0	53.42	6
P03956-1	0	71.22	11
Q04771-1	0	50.73	14
Q92781-1	0	76.42	13
P05161-1	0	71.05	7
Q8ND56-1	0	63.34	15
P19793-1	0	53.43	13
P57075-1	0	65.81	10
O95755-1	0	51.44	8
Q8NFU7-1	0	72.42	8
Q15032-1	0	48.05	11
O43186-1	0	23.83	9
Q9UBX0-1	0	75.92	12
P01574-1	0	81.22	9
P54710-1	0	52.99	7
Q86VP6-1	0	93.02	28

Table *sp_align*.

Table 2 – conserv_score

The table contains four fields:

1. The *isoid* of the sequence for which a score was calculated
2. A *methodid* that defines the method employed to compute the score of the sequence

   0 – OMA

   1 – BLAST

3. The *position of the variant* in the sequence

4. The *conservation score* of the variant

isoid	methodid	position	score
Q9NQG7-1	0	625	0.578
Q9NQG7-1	0	443	0.138
Q9NQG7-1	0	229	0.304
Q9NQG7-1	0	552	0.8
Q9NQG7-1	0	606	0.906
P55809-1	0	456	0.87
P55809-1	0	324	0.829
P55809-1	0	221	0.612
P55809-1	0	219	1
P55809-1	0	133	0.504
P55809-1	0	58	0.741
P10323-1	0	166	1
P10323-1	0	120	0.523
P03956-1	0	252	1
P03956-1	0	191	0.705
P03956-1	0	405	0.664
P03956-1	0	406	0.817
P03956-1	0	29	0.71
Q04771-1	0	206	0.722
Q92781-1	0	238	1
Q92781-1	0	35	1
Q92781-1	0	33	0.826
Q92781-1	0	280	1

Table *conserv_score* – The sequence with isoid *P55809-1* has six variants, each associated with a score.

Table 3 – mapping

The table contains four fields:

1. The *isoid* of the sequence for which we have a PDB correspondence

2. The *sp_position* that defines a position in the sequence

3. The *chainid* that defines in which chain of which structure the correspondence applies.

4. The *chain_position* that defines the position in the structure

isoid	sp_position	chainid	chain_position
P02766-1	21	1BM7A	-
P02766-1	22	1BM7A	-
P02766-1	23	1BM7A	-
P02766-1	24	1BM7A	-
P02766-1	25	1BM7A	-
P02766-1	26	1BM7A	-
P02766-1	27	1BM7A	-
P02766-1	28	1BM7A	-
P02766-1	29	1BM7A	-
P02766-1	30	1BM7A	10
P02766-1	31	1BM7A	11

P02766-1	32	1BM7A	12
P02766-1	33	1BM7A	13
P02766-1	34	1BM7A	14
P02766-1	35	1BM7A	15
P02766-1	36	1BM7A	16
P02766-1	37	1BM7A	17
P02766-1	38	1BM7A	18
P02766-1	39	1BM7A	19
P02766-1	40	1BM7A	20
P02766-1	41	1BM7A	21

Table *mapping*.

Table 4 – relftstructure

Five fields define this table:

1. The *ftid*
2. The *chainid*
3. The start position *sp_from* of the sequence that covers the structure
4. The end position *sp_to* of the sequence that covers the structure

ftid	chainid	sp_from	sp_to
VAR_004688	1E3GA	657	919
VAR_008996	1IJQA	398	713
VAR_019862	2FXMA	838	963
VAR_025386	2C0YA	17	331
VAR_005852	2B3GB	33	60

Table *relftstructure*.

Table 5 – environment

Seven fields are defined in this table:

1. The *chainid*, which is the PDB code of the 3D structure, concatenated with the chain considered
2. The *position of the variant* in the 3D structure
3. The *attid*. It is an integer that defines the type of the interaction

    0 – in an environment of 3 angstroms

    1 – in an environment of 4 angstroms

    2 – in an environment of 5 angstroms

3 – in an environment of 6 angstroms

4. The *value* representing the interacting residue. Actually, this value is the concatenation of the chain in which the residue is found together with the number of the residue in the PDB file

5. The minimal *distance* between the residue and the variant

6. The *protres* field. It indicates whether the residue is in the PDB sequence.

7. The *description* field indicates the unique atomic number for each atom that composes the interacting residue.

```
select * from modsnp.environment where chainid = '1A1NA' and position = '16 ';

 chainid | position | attid | value | distance | protres |                 description
---------+----------+-------+-------+----------+---------+-------------------------------------------
 1A1NA   | 16       |     3 |  386  |     5.61 | t       | 3229
 1A1NA   | 16       |     3 |  393  |     5.78 | t       | 3236
 1A1NA   | 16       |     0 |  394  |     2.85 | t       | 3237
 1A1NA   | 16       |     2 |  878  |     4.41 | t       | 3415
 1A1NA   | 16       |     2 |  974  |     4.92 | t       | 3466
 1A1NA   | 16       |     3 |  975  |     5.78 | t       | 3467
 1A1NA   | 16       |     1 |  976  |        4 | t       | 3468
 1A1NA   | 16       |     3 |  977  |     5.36 | t       | 3469
 1A1NA   | 16       |     1 | A14   |     3.26 | t       | 114:111:108:115:107:109:112:116:117:110:113
 1A1NA   | 16       |     0 | A15   |     1.33 | t       | 124:123:119:120:118:122:121
 1A1NA   | 16       |     0 | A17   |     1.34 | t       | 131:130:134:139:138:129:135:137:136:132:133
 1A1NA   | 16       |     1 | A18   |     3.75 | t       | 142:141:140:143
```

Table *environment* - The environment of the variant localized at the position *16* of the structure *1A1N* in chain *A*

Table 6 – interface

Nine fields were defined in this table:

1. The *pdb1*, which is the PDB code of the 3D structure

2. The first chainid *chainid1*

3. The second chainid *chainid2*

4. The *sas* representing the solvent accessible surface

5. The *ses* representing the solvent-excluded surface (also known as the molecular surface or Connolly surface)

6. The *chain2_residues* field. It indicates the residues of *chainid2* involved in contacts with *chainid1*

7. The *method* field indicates which method was used to compute the interface (Carbon alpha or Van Der Waals)

8. The *intertypeid* indicates whether the interface is a homomer or a heteromer

9. The *source* of data (PDB or PQS)

pdb1	chainid1	chainid2	sas	ses	chain2_residues	method	intertypeid	source
1A02	1A02N	1A02J	659.642	492.3236	286 .;289 .;287 .;290 .;293	carbone_alpha	0	PDB
1A02	1A02N	1A02J	723.5236	531.1622	286 .;290 .;293 .;289 .;282	van_der_waals	0	PDB

Table *interface* - The interface between the chain *N* and the chain *J* of the 3D structure *1A02*

Table 7 – sp_features

Six fields were defined in this table:

1. The *key*, which is the feature key in UniProtKB/Swiss-Prot

2. The *identifier*

3. The *description* of the feature

4. The start position *seg_start* of the feature

5. The end position *seg_end* of the feature

6. The corresponding *isoid*

key	identifier	description	seg_end	seg_start	isoid
MOD_RES		N-acetylthreonine	2	2	P31946-1
MOD_RES		Phosphoserine	186	186	P31946-1
VAR_SEQ	/FTId=VSP_018632	Missing (in isoform Short)	2	1	P31946-1
MOD_RES		N-acetylmethionine	1	1	P62258-1
MOD_RES		N-acetylglycine	2	2	Q04917-1
MOD_RES		N-acetylvaline	2	2	P61981-1
MOD_RES		Phosphoserine	248	248	P31947-1
VAR_SEQ	/FTId=VSP_021768	Missing (in isoform 2)	116	85	P31947-1

Table *sp_features*.

References

Bairoch, A. *et al.* (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **33**, D154-D159.

Bao, L. and Cui, Y. (2005) Prediction of the phenotypic effects of non-synonymous single nucleotide polymorphisms using structural and evolutionary information. *Bioinformatics*, **21**, 2185-2190.

Bao, L. *et al.* (2005) nsSNPAnalyzer: identifying disease-associated non-synonymous single nucleotide polymorphisms. *Nucleic Acids Res.*, **33**, W480-W482.

Berman, H.M. *et al.* (2000) The Protein data bank. *Nucleic Acids Res.*, **28**, 235-242.

Boeckmann, B. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365-370.

Brookes, AJ. (1999) The essence of SNPs. *Gene*, **234**, 177-186.

Cai, Z.H. *et al.* (2004) Bayesian approach to discovering pathogenic SNPs in conserved protein domains, *Human Mutation*, **24**, 178-184.

Cargill, M. *et al.* (1999) Characterization of single nucleotide polymorphisms in coding regions of human genes. *Nat Genet.*, **22**, 231-238.

Cavallo, A. and Martin, A.C.R. (2005) Mapping SNPs to protein sequence and structure data. *Bioinformatics*, **21**, 1443-1450.

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

Dobson, R.J. *et al.* (2006) Predicting deleterious nsSNPs: an analysis of sequence and structural attributes, *BMC Bioinformatics*, **7**, 217.

Ferrer-Costa, C. *et al.* (2004) Sequence-based prediction of pathological mutations. *Proteins*, **57**, 811-819.

Fredman, D. et al. (2004) HGVbase: a curated resource describing human DNA variation and phenotype relationships, *Nucleic Acids Res.*, **32**, D516-519.

Gibbs, R.A. *et al.* (2003) The International HapMap Project, *Nature*, **426**, 789-796.

Katoh, K. *et al.* (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform, *Nucleic Acids Res.*, **30**, 3059-3066.

Keskin, O. *et al.* (2005) Protein-protein interactions: organization, cooperativity and mapping in a bottom-up Systems Biology approach. *Phys. Biol.*, **2**, S24-S35.

Krawczak, M. *et al.* (2000) Human gene mutation database-a biomedical information and research resource, *Human Mutation*, **15**, 45-51.

Kruglyak, L. and Nickerson, D.A. (2001) Variation is the splice of life, *Nat Genet.*, **27**, 234-236.

Martin, A.C.R. (2005) Mapping PDB chains to UniProtKB entries. *Bioinformatics*, **21**, 4297-4301.

Ng, P.C. and Henikoff, S. (2001) Predicting deleterious amino acid substitutions. *Genome Res.*, **11**, 863-874.

Ng, P.C. and Henikoff, S. (2002) Accounting for human polymorphisms predicted to affect protein function. *Genome Res.*, **12**, 436-446.

Ng, P.C. and Henikoff, S. (2003) SIFT: predicting amino acid changes that affect protein function, *Nucleic Acids Res.*, **31**, 3812-3814.

Ng, P.C. and Henikoff, S. (2006) Predicting the Effects of Amino Acid Substitutions on Protein Function. *Annual Review of Genomics and Human Genetics*, **7**, 61-80.

Pastinien, T. *et al.* (2006) Influence of human genome polymorphism on gene expression, *Hum Mol Genet.*, **15 SpecNo 1**, R9-16.

Ramensky, V. *et al.* (2002) Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.*, **30**, 3894-3900.

Reich, D.E. *et al.* (2003) Quality and completeness of SNP databases, *Nat Genet.*, **33**, 457-458.

Schneider, A., Dessimoz, C. and H. Gonnet, G. (2007) OMA Browser – Exploring orthologous relations across 352 complete genomes, *Bioinformatics.*

Sherry, S.T. *et al.* (2001) dbSNP: the NCBI database of genetic variation, *Nucleic Acids Res.*, **30**, 3894-3900.

Sunyaev, S. *et al.* (2000) Towards a structural basis of human non-synonymous single nucleotide polymorphisms. *Trends Genet.*, **16**, 198-200.

Sunyaev, S. *et al.* (2001) Prediction of deleterious human alleles, *Hum Mol Genet.*, **10**, 591-597.

Valdar, WS. and Thornton, JM. (2001) Protein-protein interfaces: analysis of amino acid conservation in homodimers. *Proteins*, **42**, 108-124.

Valdar, WS. (2002) Scoring residue conservation. *Proteins*, **48**, 227-241.

Wang, Z. and Moult, J. (2001) SNPs, Protein Structure, and Disease. *Human Mutation*, **17**, 263-270.

Ye, ZQ. *et al.* (2007) Finding New Structural and Sequence Attributes to Predict Possible Disease-Association of Single Amino Acid Polymorphism (SAP). *Bioinformatics*, **23**, 1444-1450.

Yip, Y.L. *et al.* (2004) The Swiss-Prot Variant Page and the ModSNP Database: A Resource for Sequence and Structure Information on Human Protein Variants. *Human Mutation*, **23**, 464-470.

Yip, Y.L. *et al.* (2006) Structural Assessment of Single Amino Acid Mutations: Application to TP53 Function. *Human Mutation*, **27**, 926-937.

Yip, Y.L et al. (2007) Retrieving Mutation-Specific Information For Human Proteins In UniProt/Swiss-Prot Knowledgebase. (submitted).

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

Unpublished document

Bairoch, A., Yip, L., Famiglietti, L. The UniProtKB/Swiss-Prot protein knowledgebase in the context of human molecular medical research. Version of august 2005.