# Annotating single amino acid polymorphisms in the UniProt/Swiss-Prot knowledgebase

Yip Sonderegger, Yum Lina; Famiglietti Michel, Maria Liva; Gos, Arnaud; Duek, Paula Debora; David, Fabrice Pierre André; Gateau, Alain; Bairoch, Amos Marc

**How to cite**

YIP SONDEREGGER, Yum Lina et al. Annotating single amino acid polymorphisms in the UniProt/Swiss-Prot knowledgebase. In: Human mutation, 2008, vol. 29, n° 3, p. 361–366. doi: 10.1002/humu.20671

This publication URL: https://archive-ouverte.unige.ch/unige:766

Publication DOI: 10.1002/humu.20671

HUMAN MUTATION 29(3), 361–366, 2008

*DATABASES*

# Annotating Single Amino Acid Polymorphisms in the UniProt/Swiss-Prot Knowledgebase

**Yum L. Yip,[1,2]* Maria Famiglietti,[1] Arnaud Gos,[1] Paula D. Duek,[1] Fabrice P.A. David,[1,2] Alain Gateau,[1] and Amos Bairoch[1,2]**

*[1]Swiss-Prot Group, Swiss Institute of Bioinformatics, Centre Médical Universitaire, Geneva, Switzerland; [2]Department of Structural Biology and Bioinformatics, University of Geneva, Centre Médical Universitaire, Geneva, Switzerland*

*Communicated by Alastair F. Brown*

**UniProtKB/Swiss-Prot (http://beta.uniprot.org/uniprot; last accessed: 19 October 2007) is a manually curated knowledgebase providing information on protein sequences and functional annotation. It is part of the Universal Protein Resource (UniProt). The knowledgebase currently records a total of 32,282 single amino acid polymorphisms (SAPs) touching 6,086 human proteins (Release 53.2, 26 June 2007). Nearly all SAPs are derived from literature reports using strict inclusion criteria. For each SAP, the knowledgebase provides, apart from the position of the mutation and the resulting change in amino acid, information on the effects of SAPs on protein structure and function, as well as their potential involvement in diseases. Presently, there are 16,043 disease-related SAPs, 14,266 polymorphisms, and 1,973 unclassified variants recorded in UniProtKB/Swiss-Prot. Relevant information on SAPs can be found in various sections of a UniProtKB/Swiss-Prot entry. In addition to these, cross-references to human disease databases as well as other gene-specific databases, are being added regularly. In 2003, the Swiss-Prot variant pages were created to provide a concise view of the information related to the SAPs recorded in the knowledgebase. When compared to the information on missense variants listed in other mutation databases, UniProtKB/Swiss-Prot further records information on direct protein sequencing and characterization including posttranslational modifications (PTMs). The direct links to the Online Mendelian Inheritance in Man (OMIM) database entries further enhance the integration of phenotype information with data at protein level. In this regard, SAP information in UniProtKB/Swiss-Prot complements nicely those existing in genomic and phenotypic databases, and is valuable for the understanding of SAPs and diseases. Hum Mutat 29(3), 361–366, 2008.** 

KEY WORDS: annotation; single amino acid polymorphisms; mutation resource; proteomic

## INTRODUCTION

The completion of the human genome [International Human Genome Sequencing Consortium, 2001] has provided a large volume of data that creates the basis for the characterization of all human genes and the study of the role of genetic diversity in determining health or disease. Very few traits have a single genetic origin. Most depend on the combination of various genetic factors, together with environmental influences. In the postgenomic era a major challenge is, indeed, to understand the relationship between genetic and phenotypic variation [Ring et al., 2006]. Among genetic variations, single nucleotide polymorphism (SNP) refers to a genetic change in which a nucleotide is replaced by another one. Less than 1% of all SNPs result in a variation in the corresponding protein sequence. Nevertheless, this type of SNP, or single amino acid polymorphism (SAP), is the type of mutation most related to human diseases [Antonarakis and Cooper, 2003]. Currently, numerous central genomic databases record information on SNPs (dbSNP at www.ncbi.nlm.nih.gov/projects/SNP) [Stenson et al., 2003; Fredman et al., 2004]. These databases are mostly gene-centered and provide limited information on the structural and functional consequences of SAPs.

The UniProt Knowledgebase (UniProtKB; http://beta.uniprot.org/uniprot) is a comprehensive, freely accessible central resource on protein sequences and functional annotation [UniProt Consortium, 2007]. UniProtKB is composed of the automatically annotated UniProtKB/TrEMBL section and the manually annotated UniProtKB/Swiss-Prot section [Boeckmann et al., 2003]. Although it is not a mutation-oriented database, UniProtKB/Swiss-Prot currently records 32,698 protein variants in 6,086 human proteins, 32,282 of which are single amino acid polymorphisms or SAPs (Release 53.2, 26 June 2007). Apart from information on protein variants, a number of sequence features are also present and are specifically intended for researchers working on human genetic diseases. In this article, we report on the annotation of protein variation and human genetic diseases in the UniProtKB/Swiss-Prot knowledgebase, as well as its information content.

Received 6 July 2007; accepted revised manuscript 21 September 2007.

*Correspondence to: Yum L. Yip, Swiss-Prot Group, Swiss Institute of Bioinformatics, Centre Médical Universitaire, 1, rue Michel-Servet, 1211 Geneva 4, Switzerland. E-mail: lina.yip@isb-sib.ch

DOI 10.1002/humu.20671
Published online 3 January 2008 in Wiley InterScience (www.interscience.wiley.com).

## The Annotation of SAPs in UniProtKB/Swiss-Prot

In UniProtKB/Swiss-Prot, SAP annotation is embedded in a more general medical annotation effort, which refers to the collection and storage of information on genetic variants, their effects on protein structure and function, and involvement in diseases. Although the main focus is on SAPs, small in-frame deletions and insertions are sometimes recorded (about 416). Frameshift and nonsense mutations are outside the scope of the database and are thus not reported.

Information on genetic variants is mainly derived from literature reports. The annotation process consists in the selection and critical reading of relevant articles. Newly described variants are manually checked before integration in the database. The articles used for annotation are also cited in an entry to allow users to retrieve the original data. A major challenge in the SAP annotation process is to distinguish disease-causing missense mutations from neutral polymorphisms with no clinical relevance. As a general rule, disease-relation is annotated according to the article content, and the following criteria are also taken into account to evaluate the pathogenicity of a variant: de novo appearance of the mutation; segregation of the mutation with the disease within pedigrees; absence of the mutation in control individuals; change of amino acid polarity or size in the protein; occurrence of the change in a domain that is conserved between species and/or shared between proteins belonging to the same family. If the function of the protein is known and the effect of a mutation has been assessed by in vitro mutagenesis and functional assay, this information is also stored in the knowledgebase.

Apart from manual curation of SAPs, the knowledgebase also imports variants from dbSNP (www.ncbi.nlm.nih.gov/SNP), a central repository of genetic variation that includes both high-quality SNPs and candidate SNPs. To ensure the quality of data in UniProtKB/Swiss-Prot, only SNPs responding to the following criteria are imported into the knowledgebase. First, the protein sequence predicted by the NCBI must exactly match one of those in a UniProtKB/Swiss-Prot entry. SNPs that cannot be faithfully mapped into UniProtKB/Swiss-Prot are not integrated. Second, the frequency of each allele must be known, and only SNPs that are validated by frequency or double hit, or that have been described in the literature are retained and integrated. At the moment, 11,076 variants are linked to the corresponding dbSNP entry. It is worth noting that SAPs from dbSNP are considered as simple "variants". Disease association is annotated independently from dbSNP and only according to literature reports. This is also true for those SAPs that dbSNP takes from Online Mendelian Inheritance in Man (OMIM) database.

While the knowledgebase focuses its effort on the annotation of published variants, sequence differences revealed by sequence alignments are also stored in the database, namely in the "Sequence annotation (Features)" section using the "Sequence conflict" key. Although it cannot be excluded that these data might reflect sequencing errors, they might well correspond to yet unidentified or very rare variants and can be considered as the UniProtKB/Swiss-Prot counterpart of in silico detected variants stored in SNPs repositories such as dbSNP and HGVbase [Fredman et al., 2004].

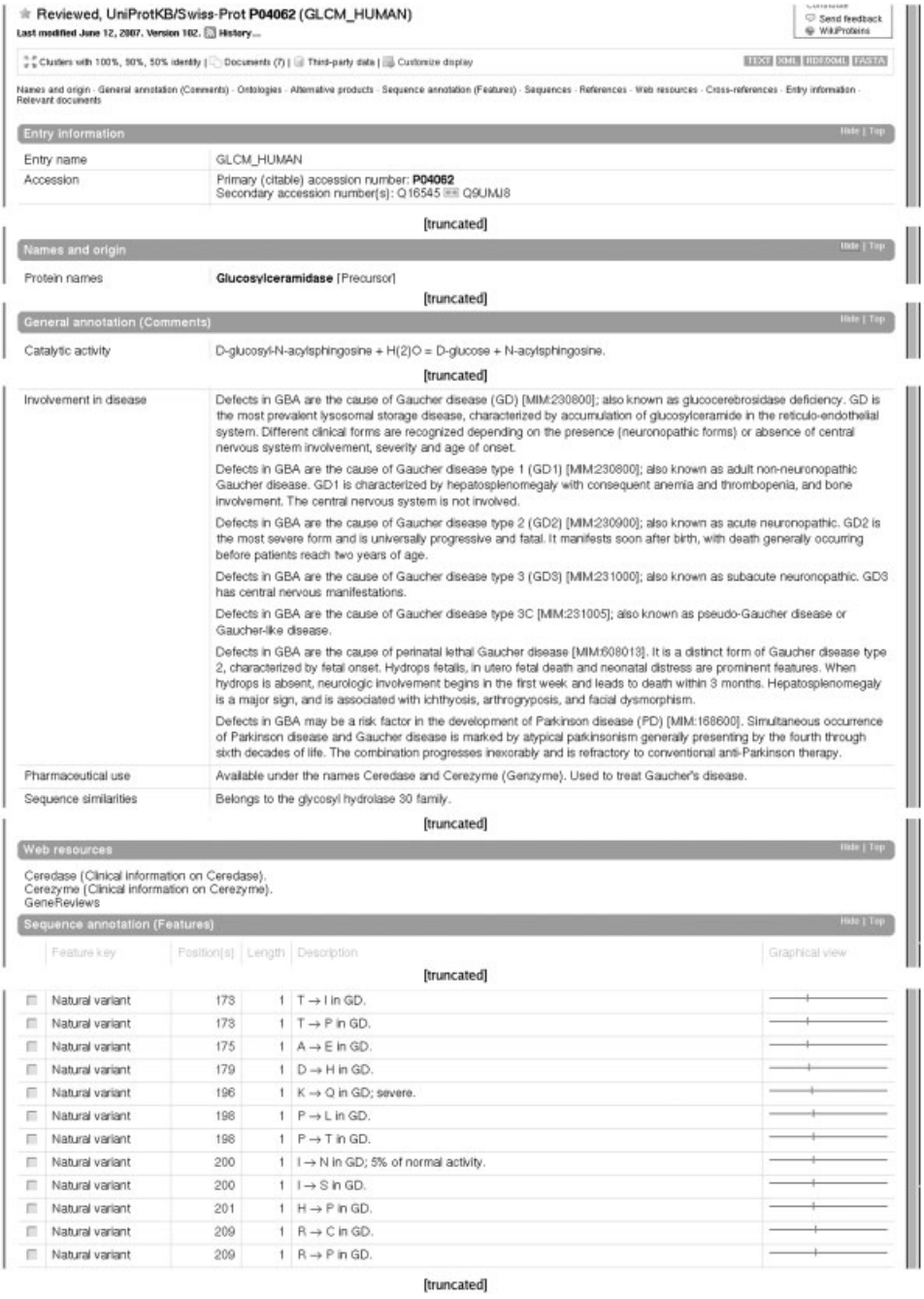
In recent years, the amount of SNP-related data has increased substantially. Consequently, it becomes ever more challenging to retrieve relevant information for SAPs annotation, and to keep the content up to date. For this purpose, we have recently developed an automatic information retrieval method based on text-mining to retrieve relevant articles from PubMed in order to maintain the information (especially the list of published documents) related to SAPs up to date (our unpublished results).

## Information Content

To date, the knowledgebase contains 16,702 manually annotated human entries (release 53.2, 26 June 2007). About 36% (6,086) of these entries contain information on variants. A total number of 32,282 SAPs are recorded, among which 16,043 (49.7%) are associated to diseases, 14,266 (44.2%) are polymorphisms, and 1,973 (6.1%) are still unclassified. It should be noted that in UniProtKB/Swiss-Prot, the term "polymorphism" refers to a variant with no clinical relevance. The frequency at which the variant occurs in the general population is thus not taken into consideration. This term is also used to describe rare variants as well as polymorphisms that have an effect on protein function, but with no resulting clinical phenotype (functional polymorphisms). Up-to-date statistics can be found at http://beta.uniprot.org/docs/humsavar.

Since one of the major aims of the knowledgebase is to provide its users with nonredundant data, we chose to display in the knowledgebase the most common isoform of a protein as its "master" sequence. All variations relative to this "master" sequence are annotated in the "Sequence annotation (Features)" table ("FT" lines) using diverse keys according to the origin of the difference (e.g., alternative splicing, polymorphism, or conflicting sequencing results). We thus keep in a single protein entry all the information relative to its variation. This is in contrast to other automatically created databases or repositories, which maintain one sequence entry per unique sequence. Consequently, genetic variation data in UniProtKB/Swiss-Prot are stored in the feature table of the relevant entries using the "Natural variant" key (Fig. 1). Each of these variants is given a unique identifier (FTId) to allow a direct link to the relevant entries in disease mutation databases and to provide these databases with a method to implement reciprocal links. Information on disease association can be found in the FT descriptions (Fig. 1). The detailed description of the disease is recorded in the "General annotation (Comments)" section ("CC" lines), under "Involvement in disease". If a variant is associated with more than one disease phenotype, the variant is recorded only once in the database but information about all diseases is stored in the variant FT description field as well as in the "Involvement in disease" field as described above (see for example, the entry P56539 at http://beta.uniprot.org/uniprot/P56539). However, if at a given position of a sequence, there are more than one type of variant involved in a given phenotype (e.g., in Fig. 1, variants p.Arg170Cys and p.Arg170Leu), they are all annotated in the feature table. On the other hand, the comment "Polymorphism" contains information on repeat expansion polymorphisms (entry P42858 at http://beta.uniprot.org/uniprot/P42858), complex alleles and their nomenclature (entry Q03519 at http://beta.uniprot.org/uniprot/Q03519), blood group systems (entry P29972 at http://beta.uniprot.org/uniprot/P29972) as well as nondisease phenotypes (entry P59533 at http://beta.uniprot.org/uniprot/P59533).

In addition to information on variants, the feature table in UniProtKB/Swiss-Prot also contains description of other sequence-specific features, i.e., relevant protein domains and sites, posttranslational modification (PTM) sites, catalytic sites, which may aid in the understanding of mutations with respect to their phenotype. For example, the mutations p.Asn770Lys and p.Gln776Arg in the protein mineralocorticoid receptor (P08235) were both found to be involved in autosomal dominant



Reviewed, UniProtKB/Swiss-Prot **P04062** (GLCM_HUMAN)
Last modified June 12, 2007. Version 102. History...

Send feedback
WikiProteins

Clusters with 100%, 90%, 50% identity | Documents (7) | Third-party data | Customize display

TEXT XML RDF/XML FASTA

Names and origin · General annotation (Comments) · Ontologies · Alternative products · Sequence annotation (Features) · Sequences · References · Web resources · Cross-references · Entry information · Relevant documents

**Entry information** Hide | Top

| | |
|---|---|
| Entry name | GLCM_HUMAN |
| Accession | Primary (citable) accession number: **P04062**<br>Secondary accession number(s): Q16545 Q9UMJ8 |

[truncated]

**Names and origin** Hide | Top

| | |
|---|---|
| Protein names | **Glucosylceramidase** [Precursor] |

[truncated]

**General annotation (Comments)** Hide | Top

| | |
|---|---|
| Catalytic activity | D-glucosyl-N-acylsphingosine + H(2)O = D-glucose + N-acylsphingosine. |

[truncated]

| | |
|---|---|
| Involvement in disease | Defects in GBA are the cause of Gaucher disease (GD) [MIM:230800]; also known as glucocerebrosidase deficiency. GD is the most prevalent lysosomal storage disease, characterized by accumulation of glucosylceramide in the reticulo-endothelial system. Different clinical forms are recognized depending on the presence (neuronopathic forms) or absence of central nervous system involvement, severity and age of onset.<br>Defects in GBA are the cause of Gaucher disease type 1 (GD1) [MIM:230800]; also known as adult non-neuronopathic Gaucher disease. GD1 is characterized by hepatosplenomegaly with consequent anemia and thrombopenia, and bone involvement. The central nervous system is not involved.<br>Defects in GBA are the cause of Gaucher disease type 2 (GD2) [MIM:230900]; also known as acute neuronopathic. GD2 is the most severe form and is universally progressive and fatal. It manifests soon after birth, with death generally occurring before patients reach two years of age.<br>Defects in GBA are the cause of Gaucher disease type 3 (GD3) [MIM:231000]; also known as subacute neuronopathic. GD3 has central nervous manifestations.<br>Defects in GBA are the cause of Gaucher disease type 3C [MIM:231005]; also known as pseudo-Gaucher disease or Gaucher-like disease.<br>Defects in GBA are the cause of perinatal lethal Gaucher disease [MIM:608013]. It is a distinct form of Gaucher disease type 2, characterized by fetal onset. Hydrops fetalis, in utero fetal death and neonatal distress are prominent features. When hydrops is absent, neurologic involvement begins in the first week and leads to death within 3 months. Hepatosplenomegaly is a major sign, and is associated with ichthyosis, arthrogryposis, and facial dysmorphism.<br>Defects in GBA may be a risk factor in the development of Parkinson disease (PD) [MIM:168600]. Simultaneous occurrence of Parkinson disease and Gaucher disease is marked by atypical parkinsonism generally presenting by the fourth through sixth decades of life. The combination progresses inexorably and is refractory to conventional anti-Parkinson therapy. |
| Pharmaceutical use | Available under the names Ceredase and Cerezyme (Genzyme). Used to treat Gaucher's disease. |
| Sequence similarities | Belongs to the glycosyl hydrolase 30 family. |

[truncated]

**Web resources** Hide | Top

Ceredase (Clinical information on Ceredase).
Cerezyme (Clinical information on Cerezyme).
GeneReviews

**Sequence annotation (Features)** Hide | Top

| Feature key | Position(s) | Length | Description | Graphical view |
|---|---|---|---|---|
| [truncated] | | | | |
| Natural variant | 173 | 1 | T → I in GD. | |
| Natural variant | 173 | 1 | T → P in GD. | |
| Natural variant | 175 | 1 | A → E in GD. | |
| Natural variant | 179 | 1 | D → H in GD. | |
| Natural variant | 196 | 1 | K → Q in GD; severe. | |
| Natural variant | 198 | 1 | P → L in GD. | |
| Natural variant | 198 | 1 | P → T in GD. | |
| Natural variant | 200 | 1 | I → N in GD; 5% of normal activity. | |
| Natural variant | 200 | 1 | I → S in GD. | |
| Natural variant | 201 | 1 | H → P in GD. | |
| Natural variant | 209 | 1 | R → C in GD. | |
| Natural variant | 209 | 1 | R → P in GD. | |

[truncated]

FIGURE 1. **Excerpt of the UniProtKB/Swiss-Prot entry P04062 showing the annotation of natural variants in the "Sequence annotation (Features)" section as well as the corresponding disease descriptions in the "General annotation (Comments)" section. The complete entry can be accessed from http://beta.uniprot.org/uniprot/P04062. A detailed description of the structure of a UniProt/Swiss-Prot entry and of the type of information annotated in the different line types, is available from the Swiss-Prot user manual (http://beta.uniprot.org/docs/userman.htm).**

pseudohypoaldosteronism type I. The mutation p.Gln776Arg was further indicated to reduce aldosterone binding. This functional effect can be easily explained by the fact that residues 770, 776, 817, and 945 are involved in the binding of steroid, as annotated in the feature table of the entry (Fig. 2). Most likely, the p.Asn770Lys also induces a similar effect.

Besides its own information content, UniProtKB/Swiss-Prot cross-references more than 100 other databases (http://beta.uniprot.org/docs/dbxref). Human sequence entries have a cross-reference to OMIM (at www.ncbi.nlm.nih.gov/omim), a knowledgebase of human genes and genetic disorders. Links to genetic variation resources such as National Institute of Environmental Health Sciences (NIEHS)-SNPs (http://egp.gs.washington.edu) and Seattle SNPs (http://pga.gs.washington.edu) are also provided in the entry. Moreover, if locus-specific mutation databases (LSDBs) relevant to a given protein exist, a direct link to these databases can be found under the section called "Web resources." The LSDBs represent a complement to the data in UniProtKB/Swiss-Prot because, by focusing on one single gene, they can generally provide more complete and in-depth information on each variant.

Sequence annotation (Features)

| Feature key | Position(s) | Length | Description | Graphical view |
|---|---|---|---|---|
| **Molecule processing** | | | | |
| Chain | 1 – 984 | 984 | Mineralocorticoid receptor | |
| **Regions** | | | | |
| DNA binding | 603 – 668 | 66 | Nuclear receptor | |
| Zinc finger | 603 – 623 | 21 | NR C4-type | |
| Zinc finger | 639 – 663 | 25 | NR C4-type | |
| Region | 1 – 602 | 602 | Modulating | |
| Region | 669 – 732 | 64 | Hinge | |
| Region | 733 – 984 | 252 | Steroid-binding | |
| Region | 782 – 785 | 4 | Important for coactivator binding | |
| **Sites** | | | | |
| Binding site | 770 | 1 | Steroid | |
| Binding site | 776 | 1 | Steroid | |
| Binding site | 817 | 1 | Steroid | |
| Binding site | 945 | 1 | Steroid | |
| **Natural variations** | | | | |
| Alternative sequence | 633 | 1 | G → GKCSW in isoform 3. | |
| Alternative sequence | 672 – 788 | 117 | Missing in isoform 4. | |
| Alternative sequence | 672 – 706 | 35 | ARKSK_QSPEE → ERRCISLPCMNYARGCTKSA FSSFDCSSPLKNTPS in isoform 2. | |
| Alternative sequence | 707 – 984 | 278 | Missing in isoform 2. | |
| Natural variant | 180 | 1 | I → V High frequency in healthy individuals; found in a patient with sporadic pseudohypoaldosteronism type I; increases transcription transactivation at low aldosterone concentrations. dbSNP rs5522. | |
| Natural variant | 241 | 1 | A → V High frequency in healthy individuals; found in a patient with sporadic pseudohypoaldosteronism type I; reduces transcription transactivation upon aldosterone binding. | |
| Natural variant | 444 | 1 | N → T: dbSNP rs5523. | |
| Natural variant | 537 | 1 | R → Q: dbSNP rs5526. | |
| Natural variant | 554 | 1 | N → S: dbSNP rs5527. | |
| Natural variant | 633 | 1 | G → R in PHA1; reduces transcription transactivation upon aldosterone binding. | |
| Natural variant | 645 | 1 | C → S in PHA1. | |
| Natural variant | 659 | 1 | R → S in PHA1. | |
| Natural variant | 759 | 1 | P → S in PHA1. | |
| Natural variant | 769 | 1 | L → P in PHA1. | |
| Natural variant | 770 | 1 | N → K in PHA1. | |
| Natural variant | 776 | 1 | Q → R in PHA1; reduces aldosterone binding. | |
| Natural variant | 805 | 1 | S → P in PHA1. | |
| Natural variant | 810 | 1 | S → L in early onset hypertension; alters receptor specificity and leads to constitutive activation. | |
| Natural variant | 815 | 1 | S → R in PHA1. | |
| Natural variant | 818 | 1 | S → L in PHA1; abolishes translocation to the nucleus and transcription transactivation upon aldosterone binding. | |

FIGURE 2. **The feature table of the UniProt/Swiss-Prot protein P08235.**

Other medically relevant databases cross-referenced in the knowledgebase are: DrugBank (http://redpoll.pharmacy.ualberta.ca/drugbank), Orphanet—a free-access website providing information on rare diseases and orphan drugs (www.orpha.net), PharmGKB (www.pharmgkb.org/index.jsp), GeneAtlas (www.dsi.univ-paris5.fr/genatlas), GeneCards (www.genecards.org), and GeneLynx (www.genelynx.org).

The above description shows that while there is a lot of information on SAPs in UniProtKB/Swiss-Prot, this information is scattered in several different sections of a protein entry. To help the users view concise information on variants, the UniProtKB/Swiss-Prot variant web pages were created in 2003 [Yip et al., 2004]. These pages can be accessed via a link provided by each natural variant in the "Sequence annotation (Features)" table. Apart from general information such as the amino acid change, position, and effect of the variant, as well as its association with diseases, the pages provide additional structural information of the variant. In particular, when available, 3D-models generated by an automatic homology modeling method can be obtained so that one can visualize the mutation directly on the protein structure. Apart from these existing features, new data are planned for inclusion in the Swiss-Prot variant pages in the near future. These include: 1) the display of conservation score of the SAP at sequence and structural level; 2) the display of residues involved in protein–protein interactions, and 3) the display of the local structural environment of SAP. More specifically, the presence of other sequence-specific features (e.g., residues involved in ligand binding or posttranslationally modified residues) in the structural neighborhood of the SAP will be shown. This display option will further aid in the understanding of the potential functional effect of SAPs.

Depending on users' interest, there are several ways to retrieve mutation-related information in the knowledgebase. For example, if one wants to retrieve protein entries with data on variants, the keywords "Disease mutation" and "Polymorphism" can be used. Disease-related keywords are also regularly being created to allow easy retrieval of proteins involved in complex disorders and genetically heterogeneous diseases, e.g., deafness (97 entries), obesity (28 entries), retinitis pigmentosa (39 entries), diabetes mellitus (36 entries), cardiomyopathy (36 entries), albinism (13 entries), and Charcot-Marie-Tooth disease (20 entries). Currently there are about 100 "medical" keywords and the list is growing. The complete list of human proteins with variants can be found at http://beta.uniprot.org/docs/humpvar. For users who wish to have

direct access to a particular mutation, they can use the list at http://beta.uniprot.org/docs/humsavar, which, besides stating the position, amino acid change, and the FTId of the variant, further shows its classification, i.e., whether it is disease-related, polymorphism, or unclassified (Fig. 3). It should be noted that the UniProtKB/Swiss-Prot variants, as classified in the above mentioned lists, have been reported to be the best training data set for prediction of human nonsynonymous SNPs [Care et al., 2007], thus reflecting the overall quality and accuracy of the classification.

## Position of a "Protein-Centered" Mutation Resource Compared to Genomic and Phenotypic Information Databases

Proteins are important actors in most cellular activity. They are the essential link for the understanding of genomic variations and phenotypic consequences. By offering a protein-centered view on variant data, UniProtKB/Swiss-Prot provides complementary information to the gene-centered view or disease-based view offered by most SNP-related databases.

It should be noted that the knowledgebase does not simply list amino acid changes predicted from nucleotide variations, but it stores, when available, information on direct protein sequencing and characterization including PTM. This is important as the real effect of missense variants on proteins PTM and/or structural phenotype cannot be deduced from simple translation of single-nucleotide substitutions at the DNA level. By offering numerous links to genomic (e.g., Ensembl; www.ensembl.org/index.html) and proteomic databases (e.g., SWISS-2DPAGE; www.expasy.org/ch2d, and Siena-2D PAGE; www.bio-mol.unisi.it/2d/2d.html), UniProtKB/Swiss-Prot can be regarded as an integration platform between genomics and proteomics.

As for phenotypic data, UniProtKB/Swiss-Prot is extensively linked to OMIM. At present, 2,601 genetic diseases described in the knowledgebase have a direct link to the corresponding OMIM phenotype entries. The users can use these links to retrieve more detailed disease information to complement those recorded in the disease comment lines. Each OMIM entry has a full-text summary of up-to-date knowledge about genetically determined phenotypes and information on inheritance patterns and most representative allelic variants. OMIM provides also the cytogenetic map location of disease genes. It should however be noted that since OMIM is mainly directed to geneticists, it is not specifically concerned with sequence-oriented data. As a consequence, there is a low correspondence between allelic variants and sequence information and it is difficult to faithfully map allelic variants onto a sequence.

This file can be downloaded by ftp.

```
------------------------------------------------------------------------------
        UniProt - Swiss-Prot Protein Knowledgebase
        Swiss Institute of Bioinformatics (SIB); Geneva, Switzerland
        European Bioinformatics Institute (EBI); Hinxton, United Kingdom
        Protein Information Resource (PIR); Washington DC, USA
------------------------------------------------------------------------------

Description: Human polymorphisms and disease mutations: index
Name:        HUMSAVAR.TXT
Release:     53.2 of 26-Jun-2007
------------------------------------------------------------------------------

Statistics for single amino acid variants:

             Disease variants: 16043
             Polymorphisms: 14266
             Unclassified variants: 1973
             Total: 32282
```

| Main gene name | Swiss-Prot Entry name | AC | FTId | Seq pos | AA change | Type of variant | Disease name |
|---|---|---|---|---|---|---|---|
| A1BG | A1BG_HUMAN | P04217 | VAR_018369 | 52 | R -> H | Polymorphism | |
| A1BG | A1BG_HUMAN | P04217 | VAR_018370 | 395 | H -> R | Polymorphism | |
| A2M | A2MG_HUMAN | P01023 | VAR_026820 | 639 | D -> N | Polymorphism | |
| A2M | A2MG_HUMAN | P01023 | VAR_000012 | 704 | R -> H | Polymorphism | |
| A2M | A2MG_HUMAN | P01023 | VAR_026821 | 815 | L -> Q | Polymorphism | |
| A2M | A2MG_HUMAN | P01023 | VAR_000013 | 972 | C -> Y | Polymorphism | |
| A2M | A2MG_HUMAN | P01023 | VAR_000014 | 1000 | V -> I | Polymorphism | |
| A4GALT | A4GAT_HUMAN | Q9NPC4 | VAR_014296 | 37 | M -> V | Polymorphism | |
| A4GALT | A4GAT_HUMAN | Q9NPC4 | VAR_022320 | 163 | Q -> R | Polymorphism | |
| A4GALT | A4GAT_HUMAN | Q9NPC4 | VAR_014297 | 183 | M -> K | Unclassified | |
| A4GALT | A4GAT_HUMAN | Q9NPC4 | VAR_017508 | 187 | G -> D | Polymorphism | |
| A4GALT | A4GAT_HUMAN | Q9NPC4 | VAR_017509 | 251 | P -> L | Polymorphism | |
| A4GNT | A4GCT_HUMAN | Q9UNA3 | VAR_022096 | 218 | A -> D | Polymorphism | |
| AAAS | AAAS_HUMAN | Q9NRG9 | VAR_012804 | 15 | Q -> K | Disease | Achalasia-addisonianism-alacrima syndrome (AAA syndrome) [MIM:231550] |
| AAAS | AAAS_HUMAN | Q9NRG9 | VAR_012805 | 160 | H -> R | Disease | Achalasia-addisonianism-alacrima syndrome (AAA syndrome) [MIM:231550] |
| AAAS | AAAS_HUMAN | Q9NRG9 | VAR_012806 | 263 | S -> P | Disease | Achalasia-addisonianism-alacrima syndrome (AAA syndrome) [MIM:231550] |
| AADAC | AAAD_HUMAN | P22760 | VAR_014798 | 281 | I -> V | Polymorphism | |
| AAK1 | AAK1_HUMAN | Q2M2I8 | VAR_031129 | 509 | K -> Q | Polymorphism | |
| AARS | SYAC_HUMAN | P49588 | VAR_028204 | 275 | G -> D | Polymorphism | |
| AARSL | SYAM_HUMAN | Q5JTZ9 | VAR_027609 | 339 | I -> V | Polymorphism | |
| AARSL | SYAM_HUMAN | Q5JTZ9 | VAR_027610 | 484 | A -> D | Polymorphism | |
| AAT1 | AAT1_HUMAN | Q724T9 | VAR_030243 | 207 | P -> A | Polymorphism | |
| AAT1 | AAT1_HUMAN | Q724T9 | VAR_030244 | 253 | S -> T | Polymorphism | |
| AAT1 | AAT1_HUMAN | Q724T9 | VAR_030245 | 320 | S -> C | Polymorphism | |
| AATK | LMTK1_HUMAN | Q6ZMQ8 | VAR_032679 | 81 | S -> F | Disease | An ovarian mucinous carcinoma sample |
| AATK | LMTK1_HUMAN | Q6ZMQ8 | VAR_032680 | 97 | L -> V | Disease | A lung adenocarcinoma sample |
| AATK | LMTK1_HUMAN | Q6ZMQ8 | VAR_032681 | 104 | M -> V | Disease | An ovarian mucinous carcinoma sample |
| AATK | LMTK1_HUMAN | Q6ZMQ8 | VAR_027267 | 118 | T -> M | Polymorphism | |
| AATK | LMTK1_HUMAN | Q6ZMQ8 | VAR_032682 | 703 | G -> C | Polymorphism | |
| AATK | LMTK1_HUMAN | Q6ZMQ8 | VAR_032683 | 815 | S -> R | Polymorphism | |
| AATK | LMTK1_HUMAN | Q6ZMQ8 | VAR_032684 | 923 | S -> L | Polymorphism | |
| AATK | LMTK1_HUMAN | Q6ZMQ8 | VAR_032685 | 1160 | E -> K | Polymorphism | |
| AATK | LMTK1_HUMAN | Q6ZMQ8 | VAR_032686 | 1192 | P -> S | Polymorphism | |
| AATK | LMTK1_HUMAN | Q6ZMQ8 | VAR_032687 | 1266 | F -> S | Polymorphism | |
| AATK | LMTK1_HUMAN | Q6ZMQ8 | VAR_032688 | 1332 | A -> T | Polymorphism | |
| ABAT | GABT_HUMAN | P80404 | VAR_018979 | 56 | Q -> R | Polymorphism | |
| ABAT | GABT_HUMAN | P80404 | VAR_008883 | 220 | R -> K | Disease | GABA-AT deficiency [MIM:137150] |
| ABCA1 | ABCA1_HUMAN | O95477 | VAR_017529 | 85 | P -> L | Disease | High density lipoprotein deficiency type 2 (HDLD2) [MIM:604091] |

FIGURE 3. **List of human proteins with SAPs (http://beta.uniprot.org/docs/humpvar.htm).**

The cross-references to OMIM in UniProtKB/Swiss-Prot entries help overcome this difficulty and allow the integration of phenotype information with data at protein level.

## CONCLUSION

The integration of mutation data, functional data, protein sequence and structural information, protein–protein interaction data, and phenotypic descriptions is essential to elucidate the chain of events leading from a molecular defect to a pathology. UniProtKB/Swiss-Prot responds to this need by providing high quality documentation of protein variety, function and disease, as well as a wealth of cross-links to specialized web resources.

## REFERENCES

Antonarakis SE, Cooper DN. 2003. Mutations in human genetic disease. In: Cooper DN, editor. Encyclopedia of the human genome. London: Nature Publishing Group. p 227–253.

Boeckmann B, Bairoch 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.

Care MA, Needham CJ, Bulpitt AJ, Westhead DR. 2007. Deleterious SNP prediction: be mindful of your training data! Bioinformatics 23:664–672.

Fredman D, Munns G, Rios D, Sjoholm F, Siegfried M, Lenhard B, Lehvaslaiho H, Brookes AJ. 2004. HGVbase: a curated resource describing human DNA variation and phenotype relationships. Nucleic Acids Res 32(Database issue):D516–D519.

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. Nature 409:860–921.

Ring HZ, Kwok PY, Cotton RG. 2006. Human variome project: an international collaboration to catalogue human genetic variation. Pharmacogenomics 7:969–972.

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

UniProt Consortium. 2007. The Universal Protein Resource (UniProt). Nucleic Acids Res 35(Database issue):D193–D197.

Yip YL, Scheib H, Diemand AV, Gattiker A, Famiglietti LM, Gasteiger E, Bairoch A. 2004. The Swiss-Prot variant page and the ModSNP database: a resource for sequence and structure information on human protein variants. Hum Mutat 23:464–470.