BIOINFORMATICS APPLICATIONS NOTE
Vol. 26 no. 6 2010, pages 851–852
doi:10.1093/bioinformatics/btq028

Databases and ontologies
Advance Access publication January 26, 2010

Easy retrieval of single amino-acid polymorphisms and phenotype information using SwissVar

Anaïs Mottaz[1,2,*], Fabrice P.A. David[1,2], Anne-Lise Veuthey[1] and Yum L. Yip[1,2,*]
[1]Swiss Institute of Bioinformatics and [2]Department of Structural Biology and Bioinformatics, Centre Médical Universitaire, 1, rue Michel-Servet, 1211 Geneva 4, Switzerland

Associate Editor: Dmitrij Frishman

ABSTRACT

Summary: The SwissVar portal provides access to a comprehensive collection of single amino acid polymorphisms and diseases in the UniProtKB/Swiss-Prot database via a unique search engine. In particular, it gives direct access to the newly improved Swiss-Prot variant pages. The key strength of this portal is that it provides a possibility to query for similar diseases, as well as the underlying protein products and the molecular details of each variant. In the context of the recently proposed molecular view on diseases, the SwissVar portal should be in a unique position to provide valuable information for researchers and to advance research in this area.

Availability: The SwissVar portal is available at www.expasy.org/swissvar

Contact: anais.mottaz@isb-sib.ch; lina.yip@isb-sib.ch

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on November 24, 2009; revised on January 18, 2010; accepted on January 19, 2010

1 INTRODUCTION

Human variation data is one of the most valuable information originating from the Human Genome Project (HGP). The current challenge is how to optimally exploit this data to better understand disease association and accelerate the pace towards personalized treatments. Indeed, there are still numerous unanswered questions on the exact relationship between genetic variations, phenotypes and diseases. A plethora of databases or prediction tools exist (Thusberg *et al.*, 2009). Among the databases, only few are central databases covering mutations on all genes. They are mostly gene-centric, with little information related to the proteome. The disease and phenotype information are also currently unstructured, making specific queries difficult. This is a pity, particularly in the context of the recently proposed molecular view on diseases, which emphasizes the relationship between the disease/phenotypic networks and the underlying protein interaction or functional networks (Lage *et al.*, 2007; Oti *et al.*, 2008). Indeed, the possibility to query for similar diseases, as well as the underlying protein products and the molecular details of each variant might prove extremely useful for researchers to study a particular family of disorders or to formulate hypotheses for further research.

In this article, we present the SwissVar portal (www.expasy.org/swissvar), which provides access to a comprehensive collection of single amino acid polymorphisms (SAPs) and diseases in the UniProtKB/Swiss-Prot knowledgebase via a unique search engine. This represents nearly 3300 diseases and 60 000 human protein variations (release 57.10) (Yip *et al.*, 2008). In addition, SwissVar gives direct access to the newly improved Swiss-Prot variant pages that are widely cited by the community but can not be queried, up to now.

2 IMPLEMENTATION

SwissVar accesses two relational databases that store data on variants and diseases. The database UniMed contains disease information extracted from UniProtKB/SwissProt and their mapping to MeSH terms (Mottaz *et al*., 2008). The variant data is found in the ModSNP database (Yip *et al*., 2004). Structural information is calculated through SSMAP, a residue–residue mapping of Protein Data Bank (PDB) structures (David *et al*., 2008). The databases are implemented in PostgreSQL 8.1.9 and are updated at each UniProt release.

The system implementation is based on a three-tier architecture. CGI programs written in Perl query the databases and dynamically generate the web pages. The interface is accessible with the main web browsers.

3 FEATURES

3.1 Query options

Three main search categories are provided: (i) by diseases, (ii) by gene/protein names and (iii) by variant types or functional/structural features.

Query by disease terms enable search using disease names, OMIM identifiers or MeSH terms of the disease category. This query is powerful in that it exploits the mapping between Swiss-Prot disease names and MeSH terms (Mottaz *et al*., 2008), as well as the hierarchy in MeSH to assemble groups of diseases to a granularity defined by users. For example, the users can query for all proteins related to metabolism diseases, and gather in one click proteins and variants related to refsum disease, gout etc. The representation of the MeSH hierarchy further enables the visualization and navigation inside the categories of diseases in which the queried proteins are implicated.

The second axis of query is protein centric. Users can search with a protein or gene name, as well as Swiss-Prot identifiers (AC or ID). Queries with gene names are automatically normalized using a list of

[*]To whom correspondence should be addressed.

synonyms. This option could be particularly useful when analyzing gene or protein expression data.

Finally, variants recorded in Swiss-Prot/UniProtKB can be searched by their molecular characteristics. Several attributes of the amino acid concerned by the mutation can be specified, e.g. the conservation score of the residue, its surrounding environment (both sequential and structural), its surface accessibility as well as its involvement in interfaces are all adjustable parameters. The variants can also be queried using Swiss-Prot feature identifier (FTID), dbSNP rsID, the position of the mutation or the type of amino acid change.

The combination of all search parameters is possible. This combination strongly enhances the query power and the information content of the tool. For example, it is possible to retrieve all variants implicated in metabolic brain diseases, which are within 4 Å of a metal binding site (Supplementary Fig. 1).

3.2 Result pages

The result of the search is presented in a table (Supplementary Fig. 2), from which the users can have direct access to the original UniProtKB/Swiss-Prot entry, the MeSH descriptor data, the Swiss-Prot variant pages and the mapped PDB structure when available. The Swiss-Prot variant pages concisely present a complete outline of known information on each variant (Supplementary Figs 3 and 4). They were recently improved by newly added features which include the display of conservation score of the mutated residue at sequence and structural level; the display of protein features in the local structural environment of the variant (e.g. residues involved in ligand binding or post-translational modifications) as well as residues involved in protein–protein interaction when experimentally resolved 3D information is available. It is hoped that these information will further aid the users in understanding or evaluating the potential functional effect of SAPs. New articles on variants automatically retrieved through text-mining methods are also proposed on the pages (Yip *et al*., 2007).

Results can be downloaded as lists (e.g. a list of the protein accession numbers, a list of variant FTIDs or rsID) or in a tab-delimited or XML format containing all the information.

4 DISCUSSION

With the completion of the Human proteome, the UniProtKB/Swiss-Prot database has a complete collection of 20 330 human proteins with increasingly detailed functional annotation (The UniProt Consortium, 2009). The SwissVar portal gives access to this wealth of data by further providing the possibility to gather proteins/variants related to similar diseases, and allowing queries on variants using a range of sequence and structural parameters.

Further improvement of the portal and the information content is planned. First, data coverage: the current SAPs coverage is clearly not exhaustive. However, as a partner of the GEN2PHEN consortium (www.gen2phen.org), it is anticipated that data related to SAPs from consortium members will be made visible via UniProtKB and the Swiss-Prot variants pages. As such, the SwissVar portal will continue to gain its value as the amount of data grows. Second, disease terminology/phenotype information: the portal currently relies on MeSH classification that offers a reasonably broad coverage of diseases including genetic diseases. The classification is nevertheless not entirely based on phenotypic similarities. Incorporating comprehensive structured phenotype information could enhance the disease query. New resources, such as Human Phenotype Ontology (Robinson *et al*., 2009), are currently being studied for this purpose. Finally, it is planned that pathway information will be incorporated in the near future to allow seamless integration and search between diseases, phenotypes, pathways and detailed sequence and structural information of the variants.

5 CONCLUSION

In summary, the SwissVar portal provides a unique environment and search facility to investigate the relationship between human variants and phenotypes, with a particular focus on human proteome. To the knowledge of the authors, no online servers offer this kind of search possibilities that directly link molecular details of SAPs to disease classification. The current application also illustrates our ongoing effort in bridging biological and medical information. The SwissVar portal can be accessed via www.expasy.org/swissvar.

ACKNOWLEDGEMENTS

We would like to acknowledge Harris Procopiou, Gregory Loichot and Nathalie Lachenal who have contributed to the development of the Swiss-Prot variant pages.

Funding: Swiss National Science Foundation (3100A0-113970); European Community's Seventh Framework Programme under grant agreement 200754 (the GEN2PHEN project).

Conflict of Interest: none declared.

REFERENCES

David,F.P. and Yip,Y.L. (2008) SSMap: a new UniProt-PDB mapping resource for the curation of structural-related information in the UniProt/Swiss-Prot Knowledgebase. *BMC Bioinformatics*, **9**, 391.

Lage,K. *et al*. (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat. Biotechnol.*, **25**, 309–316.

Mottaz,A. *et al.* (2008) Mapping proteins to disease terminologies: from UniProt to MeSH. *BMC Bioinformatics,* **9**(Suppl. 5), S3.

Oti,M. *et al.* (2008) Phenome connections. *Trends Genet.,* **24**, 103–106.

Robinson,P.N. *et al.* (2009) The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *Am. J. Hum. Genet.,* **83** 610–615.

The UniProt Consortium (2009) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **37**, D169–D174.

Thusberg,J. and Vihinen,M. (2009) Pathogenic or not? And if so, then how? Studying the effects of missense mutations using bioinformatics methods. *Hum. Mutat.,* **30**, 703–714.

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. *Hum. Mutat.,* **23**, 464–470.

Yip,Y.L. *et al.* (2007) Retrieving mutation-specific information for human proteins in UniProt/Swiss-Prot Knowledgebase. *J. Bioinform. Comput. Biol.,* **5**, 1215–1231.

Yip,Y.L. *et al.* (2008) Annotating single amino acid polymorphisms in the UniProt/Swiss-Prot knowledgebase. *Hum. Mutat.*, **29**, 361–366.