Nucleic Acids Research, 2005, Vol. 33, Web Server issue ***W311–W314***
doi:10.1093/nar/gki404

MutDB services: interactive structural analysis of mutation data

Jessica Dantzer, Charles Moad[1], Randy Heiland[1] and Sean Mooney*

Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, IN 46202, USA and [1]Scientific Data Analysis Lab, Pervasive Technology Labs, Indiana University, Indianapolis, IN 46202, USA

Received February 14, 2005; Revised and Accepted March 21, 2005

ABSTRACT

Non-synonymous single nucleotide polymorphisms (SNPs) and mutations have been associated with human phenotypes and disease. As more and more SNPs are mapped to phenotypes, understanding how these variations affect the function and expression of genes and gene products becomes an important endeavor. We have developed a set of tools to aid in the understanding of how amino acid substitutions affect protein structures. To do this, we have annotated SNPs in dbSNP and amino acid substitutions in Swiss-Prot with protein structural information, if available. We then developed a novel web interface to this data that allows for visualization of the location of these substitutions. We have also developed a web service interface to the dataset and developed interactive plugins for UCSF's Chimera structural modeling tool and PyMOL that integrate our annotations with these sophisticated structural visualization and modeling tools. The web services portal and plugins can be downloaded from http://www.lifescienceweb.org/ and the web interface is at http://www.mutdb.org/.

INTRODUCTION

Understanding how missense single nucleotide polymorphisms (SNPs) affect the function of proteins is an important research area that is being studied using genetics, biochemistry, evolutionary biology and bioinformatics (1–4). Efficient identification of SNPs would be useful for SNP selection for genetic studies, understanding the molecular basis of disease, and predicting the effects of *in vitro* and *in vivo* mutagenesis experiments. Several web resources are available for the prediction or classification of mutation effects. Two notable examples are SIFT (http://blocks.fhcrc.org/sift/SIFT.html), which utilizes evolutionary information from homologous proteins (5) and PolyPhen (http://www.bork.embl-heidelberg.de/PolyPhen/), which incorporates structural information into classification rules (6). Other resources include SNP3D (http://www.snps3d.org/), SNPeffect (http://snpeffect.vib.be/index.php), PicSNP (http://plaza.umin.ac.jp/~hchang/picsnp/) and TopoSNP (http://gila.bioengr.uic.edu/snp/toposnp/). Additionally, several projects have focused on using machine learning methods to classify deleterious mutations (7–13).

One area of particular interest is in understanding how mutations and missense SNPs affect the structure of the proteins in which they are encoded. We have developed MutDB as a resource for scientists to identify the likely underlying molecular effects of a mutation, and to visualize the location of mutations upon protein structures (14). To enable researchers to investigate whether structural information is available, we are providing a website and novel web services for visualizing mutation sites, as well as an infrastructure for annotating these mutations.

To do this, we have annotated all missense SNPs in dbSNP (15) and the mutations in the Swiss-Prot (16) database with structural information, if available. We also built a website to access and visualize these annotations. In addition, we developed a web service API, using the SOAP protocol, for accessing our annotations, and extended two applications, Chimera (http://www.cgl.ucsf.edu/chimera/) (17) and PyMOL (http://pymol.sourceforge.net/) for interactive visualization of the annotation sets.

METHODS AND USAGE

Database annotations

Structure annotations are determined for the mutation data in two publicly available databases, Swiss-Prot and dbSNP.

*To whom correspondence should be addressed. Tel: +1 317 278 9221; Fax: +1 317 278 9217; Email: sdmooney@iupui.edu

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

© The Author 2005. Published by Oxford University Press. All rights reserved.

The online version of this article has been published under an open access model. Users are entitled to use, reproduce, disseminate, or display the open access version of this article for non-commercial purposes provided that: the original authorship is properly and fully attributed; the Journal and Oxford University Press are attributed as the original place of publication with the correct citation details given; if an article is subsequently reproduced or disseminated not in its entirety but only in part or as a derivative work this must be clearly indicated. For commercial re-use, please contact journals.permissions@oupjournals.org

The flat files containing the data housed in these databases were downloaded to a local machine then parsed using Perl and the BioPerl toolkit (18), when appropriate. Several pieces of information were mined for each SNP: the original source identification number, the associated protein or mRNA sequence, the location of the SNP, the wild-type and mutant amino acids involved in the exchange, and the nucleotides involved in the exchange. For positions found in the Swiss-Prot database, any related PubMed articles were also cataloged.

This data was stored in a local MySQL database. Tables in the database were created to hold the most recent updates for the data in MutDB, as well as to hold the annotations of associated PDB (19) structures mapped to each gene. All of these tables are maintained and updated by several Perl scripts.

For each gene containing an SNP, it was desirable to find any associated protein structures. Using the protein sequences recorded from the Swiss-Prot and dbSNP databases, a BLAST search was run for both the wild-type and mutant amino acid sequences. Results from these searches were only kept if they were identical to the query sequence. Local copies of PDB files were then used to find the exactly matching positions within the protein sequence, using a pattern-matching algorithm and a pairwise alignment.

Web interface

MutDB is a web-based tool, built using several Perl CGI scripts. Navigation is possible through browsing the list of genes or searching by several different parameters, including the NCBI protein ID, Swiss-Prot gene ID, gene symbol and Refseq mRNA ID. A keyword search is also available, to find genes as they relate to particular diseases or by their full name. The genes listed are from the UCSC human genome database and are displayed in alphabetic order.

Each gene link takes the user to a page listing all available SNP data. A map of all SNPs to the gene in question is shown, as well as a catalog of each SNP. A few other pieces of information that may be useful are also displayed, such as the chromosome that the gene can be found on and links to pages in other databases, including the many parts of NCBI's website and Swiss-Prot. The SNPs are divided into three categories, mutations, synonymous mutations and non-coding SNPs. For each SNP in the list, the source ID, wild-type and mutant amino acids, sequence location and any PubMed document ID numbers are displayed. The source ID values also serve as links to separate pages about each SNP.

On each SNP page, the same information as was displayed on the relevant gene page is shown, plus some additional data. Relevant PDB structures are presented, along with a way to display them via the Jmol visualization tool (http://jmol.sourceforge.net/). The mutations are highlighted in the structure so that they are more easily recognized. Also for each SNP, the amino acid sequence, if it is known, is given, with the actual mutation shown in red. These pages also include links to the original source data from Swiss-Prot or dbSNP.

Web service

Web service is a standard for providing application-to-application communication over the Internet. A web service is any service that is available over the Internet, uses messages encoded in XML (eXtensible Markup Language) and is independent of any operating system and programming language. SOAP is the current standard protocol (XML-based) for communication, and the XML-based web services Description Language is the mechanism for describing services. Web services are becoming more prevalent in the bioinformatics community. Some examples include the new (beta) RCSB PDB (19), KEGG (20) and BioMOBY (21).

The web services for MutDB represent the contents of our resources and can easily be interfaced with other existing biological databases and applications. Our core services allow access to mutation information from either a protein structure or a gene-based perspective. This facilitates visualization clients by allowing them to easily map mutations of interest to the protein structure being viewed. Our service interface is published on our distribution site and may be extended as new feature annotations are included. Attributes are declared to represent mutations, irrespective of whether it is an SNP or a Swiss-Prot mutation. These attributes include, but are note limited to, the source identification number, amino acid position, wild-type amino acid, mutant-type amino acid, phenotype and PubMed references if they exist. Structural attributes are encoded as the PDB code, chain and the starting/ending residue index. These attributes allow for a trivial mapping to an object-oriented representation of the mutations.

In addition to the MutDB services, several other utility services are offered. This includes a PDB to Gene (and *vice versa*) mapping service that has been developed to better facilitate the data. For example, a researcher could use this functionality to map mutations involving their gene of interest to a structure.

Visualization clients

To readily take advantage of the MutDB services, we provide clients that are plugins for use in two well-developed molecular graphics tools: PyMOL and UCSF's Chimera. Upon installation, a user has full access to the contents of MutDB, and all structures and annotations are dynamically displayed based on the user's input. The user is not required to type advanced display commands, because the client leverages the power of all the services previously mentioned.

The visualization plugins allow for interactive exploration of the data. For example, in Chimera, once the plugin is installed, a new pull-down submenu will appear under 'Tools' labeled 'LSW Services'. Under LSW Services, a menu item will be listed as 'MutDB'. Clicking on this will bring up the MutDB controller. Several queries are possible here. First, and most easily, the user can enter the gene symbol, such as BRCA1, TP53 or AR. This will query the service and display the PDB IDs that are associated with that gene symbol. Upon selection of a PDB and chain ID, the chain is downloaded by our PDB chain service and the mutation positions are highlighted upon the resulting structure in red, depicted in Figure 1a (Chimera) and Figure 1b (PyMOL). Additionally, a list of the mapped non-synonymous SNPs and mutations are displayed in the mutation list box. When the user selects a position from this box, the mutation is highlighted on the structure, and the source accession number, PubMed ID and comments are displayed in the information box.

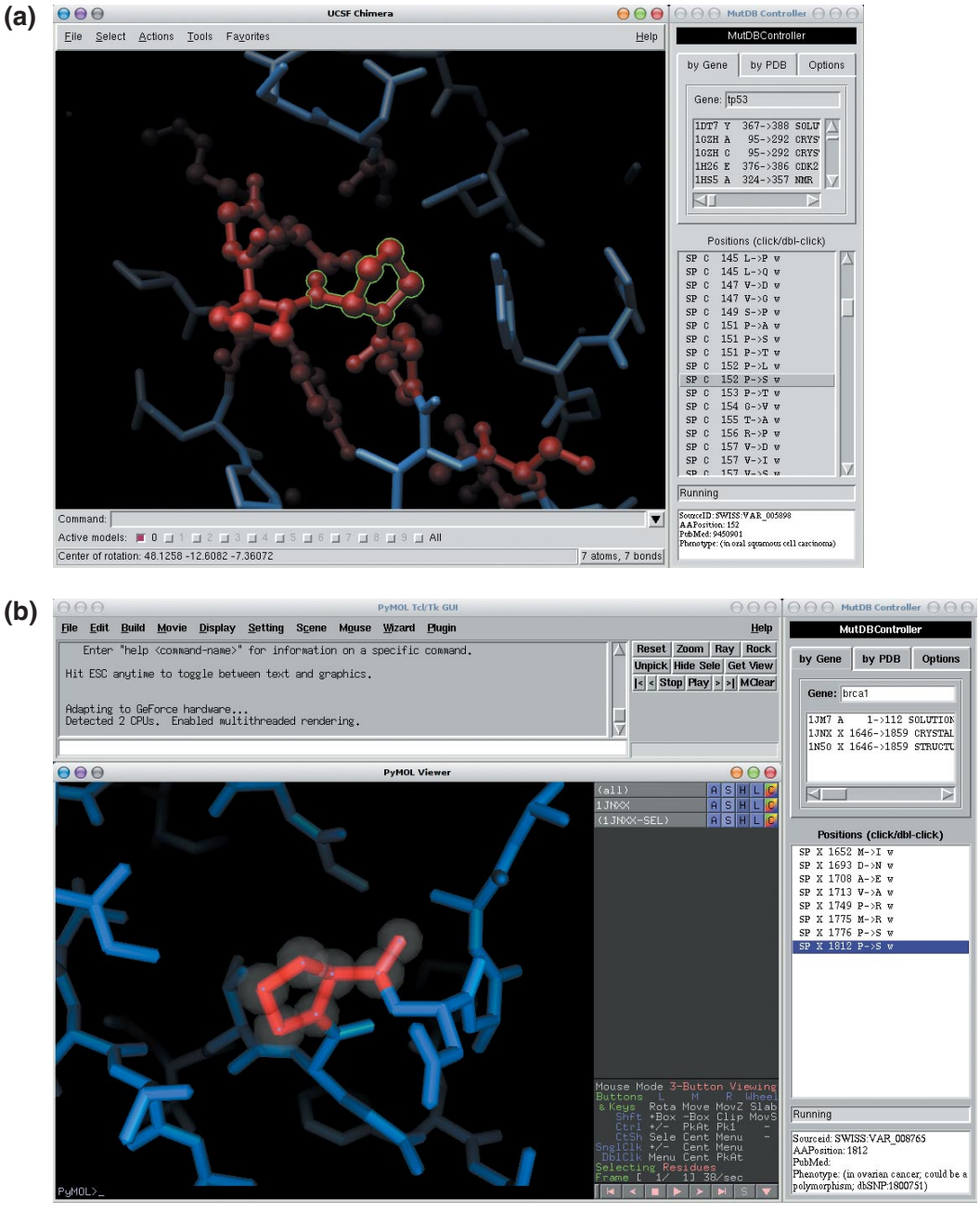

Figure 1. Visualization of mutations through web service-enabled client applications. (**a**) UCSF Chimera is shown with the mutation web service controller window showing a mutation in the TP53 gene. (**b**) PyMOL is shown with the mutation web service controller window showing a mutation in the BRCA1 gene.

DISCUSSION

We have developed a novel tool for the interactive visualization of structurally associated mutation data. We have made the interface intuitive and based on gene symbol, allowing for users not familiar with the PDB to find and visualize structures. Additionally, we have developed an infrastructure for delivering structurally relevant mutation annotations for incorporation into open source visualization tools (Figure 2).

Currently, we have annotated 2422 genes with 3587 mutations from Swiss-Prot (release 44) with structure annotations. There are 3339 distinct chains from the PDB in this set. The dbSNP (build 122) has a total of 1487 SNPs annotated with protein structures. A total of 733 structures have been solved for mutations in Swiss-Prot and 789 structures have been solved for non-synonymous SNPs in dbSNP. Currently, there are over 17 000 genes and transcript variants in MutDB, though not all of them have relevant SNP data.

Our web services are all distributed through the lifescienceweb.org domain and the web browser interface can be found at http://www.mutdb.org/.

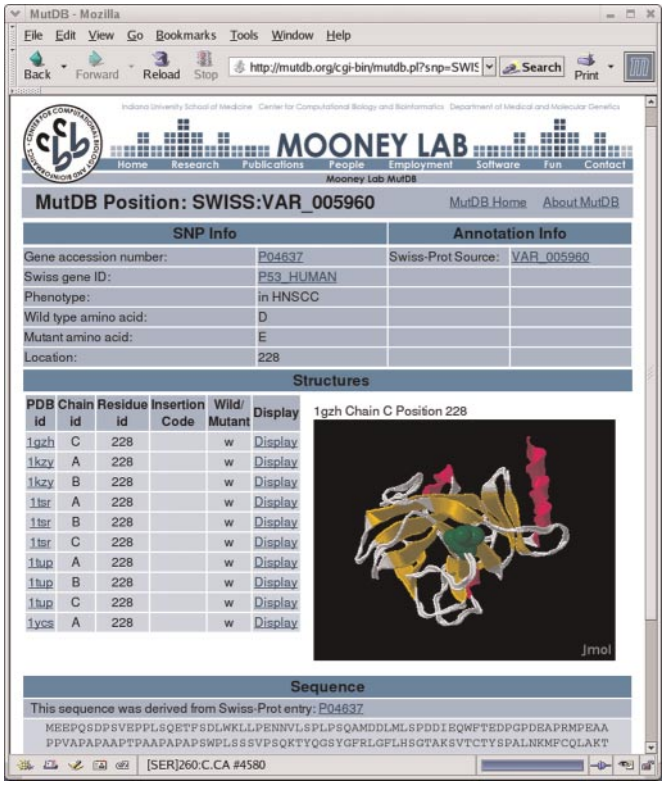

Figure 2. Web interface for structural visualization of mutation data. A mutation in TP53 is highlighted showing an aspartate-to-glutamate substitution.

ACKNOWLEDGEMENTS

The authors would like to thank Burr Fontaine for assistance with dbSNP and Andrew Campen and Eric Pettersen for feedback on the software plugins. S.M. is funded by startup funds provided by an INGEN grant, a Pervasive Technology Labs fellowship and a grant from the Showalter Trust. J.D. is funded in part through a Ronald E. McNair Scholarship. C.M. and R.H. are funded through the IPCRES Initiative grant from the Lilly Endowment. The Indiana Genomics Initiative (INGEN) of Indiana University is supported in part by Lilly Endowment Inc. Funding to pay the Open Access publication charges for this article was provided by the Pervasive Technology Labs at Indiana University.

Conflict of interest statement. None declared.

REFERENCES

1. Mooney,S.D. (2005) Bioinformatics approaches and resources for single nucleotide polymorphism functional analysis. *Brief. Bioinformatics*, **6**, 44–56.
2. Rebbeck,T.R., Spitz,M. and Wu,X. (2004) Assessing the function of genetic variants in candidate gene association studies. *Nature Rev. Genet.*, **5**, 589–597.
3. Steward,R.E., MacArthur,M.W., Laskowski,R.A. and Thornton,J.M. (2003) Molecular basis of inherited diseases: a structural perspective. *Trends Genet.*, **19**, 505–513.
4. Sunyaev,S., Lathe,W.,III and Bork,P. (2001) Integration of genome data and protein structures: prediction of protein folds, protein interactions and 'molecular phenotypes' of single nucleotide polymorphisms. *Curr. Opin. Struct. Biol.*, **11**, 125–130.
5. Ng,P.C. and Henikoff,S. (2003) SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.*, **31**, 3812–3814.
6. Ramensky,V., Bork,P. and Sunyaev,S. (2002) Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.*, **30**, 3894–3900.
7. Cai,Z., Tsung,E.F., Marinescu,V.D., Ramoni,M.F., Riva,A. and Kohane,I.S. (2004) Bayesian approach to discovering pathogenic SNPs in conserved protein domains. *Hum. Mutat.*, **24**, 178–184.
8. 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.
9. Karchin,R., Kelly,L. and Sali,A. (2005) Improving functional annotation of non-synonymous SNPs with information theory. In Klein,T.E., Hunter,L., Dunker,A.K., Jung,T. and Altman,R.B. (eds), *Proceedings of the Pacific Symposium in Biocomputing 2005 (PBS 2005)*, January 4–8, Hawaii, USA.
10. Krishnan,V.G. and Westhead,D.R. (2003) A comparative study of machine-learning methods to predict the effects of single nucleotide polymorphisms on protein function. *Bioinformatics*, **19**, 2199–2209.
11. Ng,P.C. and Henikoff,S. (2001) Predicting deleterious amino acid substitutions. *Genome Res.*, **11**, 863–874.
12. Saunders,C.T. and Baker,D. (2002) Evaluation of structural and evolutionary contributions to deleterious mutation prediction. *J. Mol. Biol.*, **322**, 891–901.
13. Sunyaev,S., Ramensky,V., Koch,I., Lathe,W.,III, Kondrashov,A.S. and Bork,P. (2001) Prediction of deleterious human alleles. *Hum. Mol. Genet.*, **10**, 591–597.
14. Mooney,S.D. and Altman,R.B. (2003) MutDB: annotating human variation with functionally relevant data. *Bioinformatics*, **19**, 1858–1860.
15. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
16. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al.* (2003) The Swiss-Prot protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
17. Pettersen,E.F., Goddard,T.D., Huang,C.C., Couch,G.S., Greenblatt,D.M., Meng,E.C. and Ferrin,T.E. (2004) UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.*, **25**, 1605–1612.
18. Stajich,J.E., Block,D., Boulez,K., Brenner,S.E., Chervitz,S.A., Dagdigian,C., Fuellen,G., Gilbert,J.G., Korf,I., Lapp,H. *et al.* (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.*, **12**, 1611–1618.
19. Deshpande,N., Addess,K.J., Bluhm,W.F., Merino-Ott,J.C., Townsend-Merino,W., Zhang,Q., Knezevich,C., Xie,L., Chen,L., Feng,Z. *et al.* (2005) The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema. *Nucleic Acids Res.*, **33**, D233–D237.
20. Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, **28**, 27–30.
21. Wilkinson,M.D., Gessler,D., Farmer,A. and Stein,L. (2003) The BioMoBy project explores open-source, simple, extensible protocols for enabling biological database interoperability. *Proc. Virt. Conf. Genom. Bioinf.*, **3**, 16–26.