SNPdbe: Constructing an nsSNP functional impacts database

Christian Schaefer¹,²,∗, Alice Meier¹, Burkhard Rost¹,² and Yana Bromberg³
¹TUM, Bioinformatics - I12, Informatik, Boltzmannstr. 3, 85748 Garching, Germany, ²TUM Graduate School of Information Science in Health (GSISH), Boltzmannstr. 11, 85748 Garching, Germany, ³Department of Biochemistry and Microbiology, School of Environmental and Biological Sciences, Rutgers University, New Brunswick, NJ 08901, USA

ABSTRACT

Summary: Many existing databases annotate experimentally characterized single nucleotide polymorphisms (SNPs). Each non-synonymous SNP (nsSNP) changes one amino acid in the gene product (single amino acid substitution; SAAS). This change can either affect protein function or be neutral in that respect. Most polymorphisms lack experimental annotation of their functional impact. Here, we introduce SNPdbe – SNP database of effects, with predictions of computationally annotated functional impacts of SNPs. Database entries represent nsSNPs in dbSNP and 1000 Genomes collection, as well as variants from UniProt and PMD. SAASs come from more than 2600 organisms; “human” being the most prevalent. The impact of each SAAS on protein function is predicted using the SNAP and SIFT algorithms and augmented with experimentally derived function/structure information and disease associations from PMD, OMIM and UniProt. SNPdbe is consistently updated and easily augmented with new sources of information. The DB is available as a MySQL dump and via a web front-end that allows searches with any combination of organism names, sequences and mutation IDs.

Availability: http://www.rostlab.org/services/snpdbe
Contact: snpdbe@rostlab.org

1 INTRODUCTION

Resources like dbSNP (Sherry, et al., 2001) and UniProt (Bairoch, et al., 2005) contain many experimentally determined nsSNPs, but few of these are annotated with respect to function. Some databases (e.g. PMID (Kawabata, et al., 1999)) contain experimental annotations of functional effects of mutants. However, these are sparsely populated and do not directly link to dbSNP or UniProt. For the vast majority of mutations lacking experimental annotation, we can gauge functional impact only via in silico analysis. Proper use of computational methods requires specific skills and resources generally inaccessible to medical researchers or experimental biologists. To help, we created a MySQL database readily usable by non-experts. We collected SAASs from PMID, dbSNP, 1000 Genomes (1000_Genomes_Project_Consortium, 2010) and UniProt “variant”s and “mutant”s. We also store “conflict” records to illustrate how sequencing discrepancies may lead to differing interpretations of the functional significance of a given sequence position. For each SAAS we predict the functional effect using SNAP (Bromberg and Rost, 2007) and SIFT (Ng and Henikoff, 2001). Where available, predictions are augmented by experimental annotations and associated human diseases. We also compute evolutionary conservation of the mutant positions. A web interface provides convenient access to underlying data via organism, sequence and mutation ID queries.

2 DATA SETUP AND RETRIEVAL

Database. SNPdbe mutation data comes from dbSNP, UniProt, 1KG and PMD (Fig. 1A). UniProt and PMD store protein sequences explicitly, while dbSNP links to RefSeq (Pruitt, et al., 2007). dbSNP collects 1KG variants with a time delay, so for SNPdbe we mapped all 1KG nsSNPs to RefSeq using Annovar (Wang, et al., 2010). We keep only one version of redundant protein sequences, referenced by md5 checksums irrespective of origin. Redundancy is assessed at full sequence identity (max one substitution per sequence) over the entire sequence (plus/minus leading Met residue). This allows correlating mutations from different sources referencing the same sequence. We currently store 1,362,793 unique SAASs in 158,004 proteins from 2,684 organisms covering all kingdoms of life; the top five contributors are human, mouse, rice, cow & rat. For each SAAS we provide the following information: (1) SNAP and SIFT binary predictions of functional effects (neutral/non-neutral). (2) Evolutionary conservation information from PSIC (Sunyaev, et al., 1999), PSI-Blast (Altschul, et al., 1997) PSSMs and frequency scores from runs against PDB (Berman, et al., 2000) and UniProt. (3) Functional effects from PMD and UniProt. For human SAASs, also available disease associations from PMID, UniProt and OMIM (Amberger, et al., 2009) (Fig. 1B). (4) dbSNP evidence and average heterozygosity, and (5) interesting functional/structural features (UniProt) at the mutation site. Data is stored in a MySQL database and is downloadable as a dump file.

Web interface. The database is web-accessible allowing gene/protein ID/name, disease, sequence (or its md5 hash) and mutant –based queries. Some queries (e.g. md5, gene ID) are exact. Sequence queries are BLAST similarity based. Keyword searches (e.g. disease) are ‘loose’, i.e. matched to corresponding free text fields. The results page lists all SAASs found within the specified sequence and their functional effect predictions, wild type/mutant conservation scores, information on disease (human...
SNPdbe is designed to fill the annotation gap left by the high cost of experimental testing for functional significance of protein variants. It joins related bits of knowledge, currently distributed throughout various databases, into a consistent, easily accessible, and updatable resource. The major features distinguishing SNPdbe from other databases are (1) the inclusion of a much wider array of organisms and data sources and (2) the explicit differentiation between functional/structural effects and disease-associations. Furthermore, unlike SNPdbe, existing resources (1) lack experimental data in public databases.

3 CONCLUSION
SNPdbe is designed to fill the annotation gap left by the high cost of experimental testing for functional significance of protein variants. It joins related bits of knowledge, currently distributed throughout various databases, into a consistent, easily accessible, and updatable resource. The major features distinguishing SNPdbe from other databases are (1) the inclusion of a much wider array of organisms and data sources and (2) the explicit differentiation between functional/structural effects and disease-associations. Furthermore, unlike SNPdbe, existing resources (1) lack experimental annotation of functional/structural changes or offer only potential to impact disease, Nucleic Acids Res, 35, D61-65.


ACKNOWLEDGEMENTS

Thanks to Laszlo Kajian, Guy Yachdav and Tim Karl (TUM) for maintenance of our compute cluster and to those who deposit their experimental data in public databases. Funding: CS, AM, BR were funded by Alexander von Humboldt Foundation; YB by Rutgers, New Brunswick, start-up funds. Conflict of interest: none declared.

REFERENCES


