SNPeffect 4.0: on-line prediction of molecular and structural effects of protein-coding variants

Greet De Baets1,2,3, Joost Van Durme1,2,3, Joke Reumers4,5, Sebastian Maurer-Stroh6, Peter Vanhee1,3, Joaquin Dopazo7,8, Joost Schymkowitz1,2,* and Frederic Rousseau1,2,*

1VIB Switch Laboratory, 3000 Leuven, 2Department of Molecular Cell Biology, University of Leuven, 3Vrije Universiteit Brussel, 4Vesalius Research Center, VIB, 3000 Leuven, 5Vesalius Research Center, University of Leuven, Belgium, 6Bioinformatics Institute (BII), Agency for Science, Technology and Research (A*STAR), Singapore, 7Department of Bioinformatics, Centro de Investigación Príncipe Felipe (CIPF) and 8Functional Genomics Node (INB) at CIPF, Valencia 46013, Spain

Received September 14, 2011; Revised October 17, 2011; Accepted October 18, 2011

ABSTRACT

Single nucleotide variants (SNVs) are, together with copy number variation, the primary source of variation in the human genome and are associated with phenotypic variation such as altered response to drug treatment and susceptibility to disease. Linking structural effects of non-synonymous SNVs to functional outcomes is a major issue in structural bioinformatics. The SNPeffect database (http://snpeffect.switchlab.org) uses sequence- and structure-based bioinformatics tools to predict the effect of protein-coding SNVs on the structural phenotype of proteins. It integrates aggregation prediction (TANGO), amyloid prediction (WALTZ), chaperone-binding prediction (LIMBO) and protein stability analysis (FoldX) for structural phenotyping. Additionally, SNPeffect holds information on affected catalytic sites and a number of post-translational modifications. The database contains all known human protein variants from UniProt, but users can now also submit custom protein variants for a SNPeffect analysis, including automated structure modeling. The new meta-analysis application allows plotting correlations between phenotypic features for a user-selected set of variants.

INTRODUCTION

Human next-generation sequencing projects currently generate millions of previously unknown single nucleotide variants (SNVs) (1). On average, every newly sequenced genome generates about 300 000 novel SNVs (2). Although it is quite straightforward to annotate these SNVs according to their genomic location (coding, non-coding and regulatory regions), and for coding SNVs to denote their effect on the translated protein (synonymous or non-synonymous), predicting the detailed effect of a coding mutation on the structure and function of a protein is a largely unsolved problem. As these variants can influence drug selection, dosing and adverse effects (3), it is recognized that this genetic information is of great importance for drug development in general (4) and crucial for personalized medicine (5). Most current approaches classify SNVs into neutral or deleterious variants by using either conservation based measures (6) or by using a combination of conservation scores and structural features (7–9). Tools for predicting stability changes upon mutation have also been developed (10,11), however these do not use explicit stability predictions based on a high-resolution structure but rather depend on black-box predictions using intelligent machine-learning approaches such as support-vector machines or neural networks.

Coding non-synonymous SNVs can affect protein structure and function to various degrees (12,13). Although predicting neutral or fully disruptive variants is relatively easy, a large portion of variants will result in more subtle intermediate phenotypic effects that are much more challenging to predict.

To tackle this challenge web servers such as PolyPhen (9) and HOPE (8), for example, base their predictions on a statistical analysis of protein structures extrapolated to the protein under study and do currently not provide quantitative free energy changes of point mutations. SNPeffect on the other hand uses the FoldX (14) force field and aims
SNP effec t PIPELINE FOR MOLECULAR
PHENOTYPING OF HUMAN PROTEIN VARIANTS

The raw data source of the SNP effec t database consists of the UniProt human variation database (http://www
.uniprot.org/docs/humsavar), containing single amino acid polymorphisms, classified either as disease mutations,
polymorphisms or yet unclassified mutations. SNP effec t predicts the impact of these variants on (i) protein aggregation
and amyloid formation (TANGO and WALTZ, respectively), (ii) chaperone binding (LIMBO) and (iii) structural stability (FoldX). The availability of a crystal structure with a minimal resolution of 4Å is required to accurately analyze the effect on protein stability with FoldX. If an exact structural match is not found, the structural stability change upon mutation (ddG).

Furthermore, SNP effec t holds annotations on functional sites, structural features, domain information, cellular
processing and post-translational modifications for each variant.

The effect on functional sites and structural features is analyzed by investigating several properties of the position
of the mutation. Data from the Catalytic Site Atlas is parsed to analyze whether the residue is part of the active site.
Secondary structure information is generated by FoldX and transmembrane topology (extracellular, intracellular, transmembrane) is predicted by TMHMM (20). Domain information is provided by SMART (21) and PFAM (22). PSORT (23) provides a prediction on the sub-cellular localization. SNP effec t also maps changes in post-translational lipid anchor attachment and the peroxisomal targeting signal PTS1 (24). Lipid attachment predictions include myristoylation (25,26), farnesylation (26), GPI-anchor attachment (27) and type-1 and type-2 geranylgeranylation (26).

All entries are additionally linked to the OMIM genetic disorder database (Online Mendelian Inheritance in Man, OMIM). McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD), 2011. World Wide Web URL: http://omim.org/) and the Gene Ontology database (28).

SNP effec t DATABASE

SNP effec t currently contains data on 63,410 human non-synonymous SNVs. Automatic updates from the UniProt human variation database are scheduled every 6 months.

The database interface (Figure 1, left) allows users to search SNVs by filtering on molecular phenotypic effects, mutation type, disease, UniProt identifier, dbSNP identifier and gene name. Molecular phenotypic effects include changes in aggregation tendency (dTANGO), amyloid formation (dWALTZ), chaperone binding (dLIMBO) and structural stability change upon mutation (ddG). Applying the filter settings results in a set of variants that can be analyzed in a protein-centered or variant-centered view (Figure 1, right).

This SNP effec t update focuses primarily on the scientist’s ability to quickly retrieve and rapidly analyze the effect of protein variants. Moreover, the wild-type protein of each variant is also fully analyzed and directly linked from the variant webpage. The effects are visualized by self-explanatory barplots and histograms. Structural data is retrieved from the Protein Data Bank (PDB) (29). When an exact match to the wild-type sequence is not found, a homology model is built from a template structure that has at least 90% sequence identity to the original sequence. If structural information is retrieved, we offer visualization of both the wild-type and mutant residue environment in the protein structure. Additionally, every phenotypic analysis is accompanied by a graphical and textual comparison to the wild-type protein. Figure 1 illustrates the summary of a variant that meets the criteria set in the filter.

META-ANALYSIS

A new feature in SNP effec t 4.0 is the ability to analyze and plot phenotypic features of a specific subset (or all) of the
SNP effec t database. The meta-analysis tool enables scientists to carry out large-scale data mining of the specified
data and visualize the results in a graphical plot. The data set of variants is primarily chosen on disease associations and the mutation type. Mutation types include disease, polymorphism and unclassified. An additional filter can be applied to limit the results by one or more disease terms that are selected from a list or specified by keywords. SNP effec t will then search for all variants of the selected type and retrieve those that are linked to the selected disease(s). For the disease type, these are solely the mutations annotated with that disease. For the polymorphisms and unclassifieds, SNP effec t retrieves all of
these variants from proteins associated with the selected disease(s). Next, the two phenotypic effects that will be analyzed and plotted can be specified (dTANGO, dWALTZ, dLIMBO or ddG) (Figure 2). For example, one can create an aggregation/stability feature plot of a set of variants to correlate aggregation changes with...
stability changes. If the number of hits for one of the mutation types exceeds 500, the average \( Y \) is plotted for each \( X \) bin, to keep the plots clear and readable. The meta-analysis tool converts phenotypic features of a selected set of variants to comprehensible scatter plots, boxplots and frequency plots (Figure 2).

**JOB SUBMISSION**

Novel to previous versions of SNPeffect (30–32), the current version includes a data submission framework that allows submitting (human or non-human) custom single protein variants for a detailed SNPeffect analysis including TANGO, WALTZ, LIMBO and FoldX. Possible input types are UniProt ID, FASTA sequence, PDB ID, or an uploaded PDB file. If only sequence information but no structural information is provided, SNPeffect will search the PDB for a matching structure to complete the stability analysis with FoldX. When an exact match is not found, a homology filter allows setting the minimum percent sequence identity a structural homolog template should have to build a homology model. The effect on structural stability is then determined by analyzing the homology model. Users receive an e-mail notification when the analysis has finished and can download the results from their SNPeffect account. The results include a PDF file with the complete phenotypic SNPeffect analysis. This file contains figures and extensive life-scientist-friendly text reports with comparison to the wild-type protein. All separate figure files are also available and free to use.

**SUMMARY**

SNPeffect 4.0 offers a detailed and comprehensible molecular and structural phenotypic analysis of all known human protein variants. Major phenotypic features such as aggregation propensity prediction, stability analysis, structural features, post-translational modification and cellular localization are intelligibly visualized and explained for each variant. The meta-analysis tool allows plotting correlations between phenotypic effects concerning a specified set of variants. Custom protein variants can now be submitted for a detailed SNPeffect analysis, including automated structure modeling. SNPeffect 4.0 is available at http://snpeffect.switchlab.org

**FUNDING**

Interuniversity Attraction Poles (IAP Network 6/43) of the Belgian Federal Science Policy Office (BelSPIo) (VIB Switch laboratory); Flanders Institute for Science and Technology (IWT) (to G.D.B.); Fund for Scientific Research (FWO), Flanders (to J.V.D.); MICINN projects BIO2008-04212, and RD06/0020/1019 (RTICC, ISCIII) (Dopazo lab, partial) GVA-FEDER (PROMETEO/2010/001, partial). The CIBER de Enfermedades Raras is an initiative of the ISCIII, MICINN. Funding for open access charge: VIB.

**Conflict of interest statement.** None declared.

**REFERENCES**


