BIOINFORMATICS APPLICATIONS NOTE Vol. 22 no. 17 2006, pages 2183–2185
doi:10.1093/bioinformatics/btl348

Databases and ontologies

SNPeffect v2.0: a new step in investigating the molecular phenotypic effects of human non-synonymous SNPs

Joke Reumers, Sebastian Maurer-Stroh, Joost Schymkowitz* and Frederic Rousseau*

Switch Laboratory, Flanders Interuniversity Institute of Biotechnology, Vrije Universiteit Brussel, Pleinlaan 2, Brussels, Belgium

Received on April 25, 2006; revised on September 9, 2006; accepted on June 22, 2006
Advance Access publication June 29, 2006
Associate Editor: Jonathan Wren

ABSTRACT

Summary: Single nucleotide polymorphisms (SNPs) constitute the most fundamental type of genetic variation in human populations. About 75 000 of these reported variations cause an amino acid change in the translated protein. An important goal in genomic research is to understand how this variability affects protein function, and whether or not particular SNPs are associated to disease susceptibility. Accordingly, the SNPeffect database uses sequence- and structure-based bioinformatics tools to predict the effect of non-synonymous SNPs on the molecular phenotype of proteins. SNPeffect analyses the effect of SNPs on three categories of functional properties: (1) structural and thermodynamic properties affecting protein dynamics and stability (2) the integrity of functional and binding sites and (3) changes in posttranslational processing and cellular localization of proteins. The search interface of the database can be used to search specifically for polymorphisms that are predicted to cause a change in one of these properties. Now based on the Ensembl human databases, the SNPeffect database has been remodeled to better fit an automatically updatable structure. The current edition holds the molecular phenotype of 74 567 nsSNPs in 23 426 proteins.

Availability: SNPeffect can be accessed through http://snpeffect.vib.be

Supplementary Material: Statistics on the contents of the database, figures on the workflow used to create the database and information on the used sources and tools is available at http://snpeffect.vib.be.

Contact: joost.schymkowitz@vub.ac.be or frederic.rousseau@vub.ac.be

1 INTRODUCTION

In the post-genomic era much attention goes to single nucleotide polymorphisms (SNPs). Such changes of a single nucleotide in the DNA sequence distinguish individuals within a population. Some of these SNPs have been directly linked to differences in susceptibility to disease, to the age of onset and severity of illness, and to drug treatment response. One much cited example is the connection between polymorphisms in the Apolipoprotein E and late-onset Alzheimer's disease (Corder *et al*., 1993). Recently, a glycine to arginine coding transition in the ACTC gene was the first demonstrated example of a polymorphism determining a visible human genetic trait, the dry/wet earwax phenotype (Yoshiura *et al*., 2006). However, in general the effect of SNPs is more subtle and evidence of strong linkage to disease or phenotype is still sporadic. Given the large number of SNPs, an exhaustive experimental study of effects on biological function is a daunting task. A valuable alternative is the use of *in silico* predictions of changes in molecular phenotype allowing prioritization of experimental studies. The SNPeffect server aims to provide a platform for predicting the effect of coding non-synonymous SNPs on the structure and function of the affected protein. It differs in setup from other such services in that it surpasses pure conservation based scores. Although the latter certainly provide some indication of the disruptive nature of a mutation, conservation does not provide insight in the nature of the affected properties. However, not all properties can be accurately predicted or even contemplated. Hence, in the SNPeffect server high resolution structural data is used where available to explicitly model the mutant structures so that changes in protein stability and binding can be assessed. Given the scarcity of high resolution structures and the relatively small set of structure based predictors, the analysis is extended with a sequence based analysis.

2 DATABASE AND WEBSERVER

2.1 Database structure

The previous version of the SNPeffect database (Reumers *et al*., 2005) was based on NCBI's dbSNP database (Sherry *et al*., 2001), holding 31 659 SNPS in 12 717 proteins. In order to automate database updating, a migration was made to the Ensembl human core and variation databases (Birney *et al*., 2006). As the aim of SNPeffect is to follow the effects of amino acid changes resulting from non-synonymous SNPs, only this type of variation is included in the database. The SNPeffect database will follow a bimonthly update schedule. SNPeffect data are also published in the new version of PupaSuite (Conde *et al*., 2005), an interactive web-based SNP analysis tool, which was released in January 2006.

2.2 Webserver implementation and structure

SNPeffect is installed on an Intel server running MySQL version 4.0, Apache 2.0 and PHP 5. Graphics are rendered with Pymol

*To whom correspondence should be addressed.

(Delano, http://www.pymol.org+) and the Perl Graphics module. The analysis pipeline is automatized using the Perl scripting language.

The data are accessible through search and browse interfaces, allowing users to select proteins or SNPs of interest using a number of identifiers and phenotypic effects. Results can be ranked according to the severity of the selected phenotypic change. Selected SNPs or proteins of interest are shown in a card view. The 'SNP view' contains five cards: (1) an overview of SNP properties and phenotypic effects, (2) prediction results for the wild-type protein, (3) and (4) each of the three categories of predicted effects and (5) links to SNP and protein entries in related databases. An additional four cards are shown in the protein view: (1) an overview of the protein identifiers, (2) the full molecular phenotype of the protein, (3) an overview of SNPs in this protein and (4) links to related databases.

3 PHENOTYPIC EFFECTS

The effects of single amino acid substitutions are categorized into three classes: (1) structure and dynamics, (2) functional sites and (3) cellular processing. New tools have been added to the first and third categories in comparison with the first release of the SNPeffect database in 2004.

3.1 Structure and dynamics

For proteins where a high resolution crystal structure is available in the Protein Data Bank (Deshpande *et al*., 2003), mutations are explicitly modelled in a full-atom representation using the empirical force field FoldX (version 2.5) (Schymkowitz *et al*., 2005) and changes in protein stability are calculated using the same package. Requiring a minimal resolution of 2.5 Å as a quality criterion for crystal structures, only a very limited set of PDB structures could be obtained. This set was extended with homology models derived from the SwissModel Repository (Kopp and Schwede, 2004) (>90% sequence identity). In this manner, a set of 266 SNPs in 133 proteins could be mapped to structural data. The neural network system PHD (Rost and Sander, 1993) was used to make predictions on the secondary structure tendencies and expected accessibility of the wild type versus the SNP sequences. The hidden Markov model based algorithm Phobius (Kall *et al*., 2004) was employed to predict transmembrane regions in wild-type and mutated sequences. Protein aggregation and amyloid formation can be predicted using the Tango (Fernandez-Escamilla *et al*., 2004) and AmyScan (López de la Paz and Serrano, 2004) tools, respectively. The Amyscan prediction is based on a position dependent scoring matrix that is derived from extensive mutational studies on the synthetic amylogenic hexapeptide sequence STVIIE. In SNPeffect, two amyloid predictions are calculated for each protein sequence. The first is based on a combinatorial scan of the STVIIE peptide, and the second includes a set of naturally occurring aggregating peptides.

3.2 Functional sites

The Catalytic Site Atlas (Porter *et al*., 2004) contains information on most known enzyme active sites and catalytic residues in enzymes of known 3D structure. In SNPeffect, mutations are checked for disruption of such a site. In addition, a motif scanner based on reported chaperone substrate specificities was developed to detect binding sites for DnaK, trigger factor and ClpB. Although these are bacterial chaperones, the substrate recognition site is highly conserved over a large number of species, including human (Rousseau *et al*., 2006).

3.3 Cellular processing

Basic predictions include protein turnover rates based on the N-terminal rule (Bachmair *et al*., 1986), and subcellular localization prediction by Psort II (Nakai and Horton, 1999). In addition, SNPeffect employs several predictors to investigate changes in post-translational modification. PhosphoBase (Kreegipuu *et al*., 1999) and O-GlycBase (Gupta *et al*., 1998) are used to screen for phosphorylated and glycosylated residues, respectively. Additional predictors for posttranslational modifications have been added (Eisenhaber *et al*., 2003; Maurer-Stroh and Eisenhaber, 2005). The five predictors for myristoylation, farnesylation, GPI-anchor attachment and type I and II geranylgeranylation predict lipid anchor attachment motifs in protein sequences with high specificity. The PTS1 signal predictor finds proteins with a C-terminus suited for peroxisomal import.

ACKNOWLEDGEMENTS

The authors are grateful to Frank and Birgit Eisenhaber and Georg Neuberger for making the GPI and PTS1 predictors available. This work was partially supported by a FEBS short term fellowship to Sebastian Maurer-Stroh. Funding to pay the Open Access publication charges was provided by Flanders Interuniversity Institute for Biotechnology (VIB).

Conflict of Interest: none declared.

REFERENCES

Bachmair,A. *et al*. (1986) *In vivo* half-life of a protein is a function of its amino-terminal residue. *Science*, **234**, 179–186.

Birney,E. *et al*. (2006) Ensembl 2006. *Nucleic Acids Res*., **34**, D556–D561.

Conde,L. *et al*. (2005) PupasView: a visual tool for selecting suitable SNPs, with putative pathological effect in genes, for genotyping purposes. *Nucleic Acids Res*., **33**, W501–W505.

Corder,E.H. *et al*. (1993) Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science*, **261**, 921–923.

DeLano,W. (2004) The pymol molecular graphics system.

Deshpande,N. *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.

Eisenhaber,F. *et al*. (2003) Prediction of lipid posttranslational modifications and localization signals from protein sequences: big-pi, nmt and pts1. *Nucleic Acids Res*., **31**, 3631–3634.

Fernandez-Escamilla,A.M. *et al*. (2004) Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat. Biotechnol*., **22**, 1302–1306.

Gupta,R. *et al*. (1999) O-GLYCBASE version 4.0: a revised database of O-glycosylated proteins. *Nucleic Acids Res*., **27**, 370–372.

Kall,L. *et al*. (2004) A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol*., **338**, 1027–1036.

Kopp,A. and Schwede,T. (2004) The SWISS-MODEL repository of annotated three-dimensional protein structure homology models. *Nucleic Acids Res*., **32**, D230–D234.

Kreegipuu,A. *et al*. (1999) PhosphoBase, a database of phosphorylation sites: release 2.0. *Nucleic Acids Res*., **29**, 237–239.

López de la Paz,M. and Serrano,L. (2004) Sequence determinants of amyloid fibril formation. *Proc. Natl Acad. Sci. USA*, **101**, 87–92.

Maurer-Stroh,S. and Eisenhaber,F. (2005) Refinement and prediction of protein prenylation motifs. *Genome Biol*., **6**, R55.

Nakai,K. and Horton,P. (1999) PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *TIBS*, **24**, 34–35.

Porter,C.T. *et al*. (2004) The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res*., **32**, D129–D133.

Reumers,J. *et al*. (2005) SNPeffect: a database mapping molecular phenotypic effects of human non-synonymous coding SNPs. *Nucleic Acids Res*., **33**, D527–D532.

Rost,B. and Sander,C. (1993) Prediction of protein secondary structure at better than 70-percent accuracy. *J. Mol. Biol*., **232**, 584–599.

Rousseau,F. *et al*. (1986) How evolutionary pressure against protein aggregation shaped chaperone specificity. *J. Mol. Biol*., **355**, 1037–1047.

Schymkowitz,J. *et al*. (2005) The FoldX web server: an online force field. *Nucleic Acids Res*., **33**, W382–W388.

Sherry,S.T. *et al*. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*., **29**, 308–311.

Yoshiura,K. *et al*. (2006) A SNP in the ABCC11 gene is the determinant of human earwax type. *Nat. Genet*., **38**, 324–330.