BIOINFORMATICS APPLICATIONS NOTE Vol. 24 no. 20 2008, pages 2397–2398
doi:10.1093/bioinformatics/btn435

Sequence analysis

SNAP predicts effect of mutations on protein function

Yana Bromberg[1,2,*], Guy Yachdav[1,2] and Burkhard Rost[1,2,3]

[1]Department of Biochemistry and Molecular Biophysics, Columbia University, 630 West 168th Street, [2]Columbia University Center for Computational Biology and Bioinformatics (C2B2) and [3]NorthEast Structural Genomics Consortium (NESG) and New York Consortium on Membrane Protein Structure (NYCOMPS), Columbia University, 1130 St Nicholas Ave. Rm. 802, New York, NY 10032, USA

Received on May 29, 2008; revised on August 10, 2008; accepted on August 14, 2008

Advance Access publication August 30, 2008

Associate Editor: Alex Bateman

ABSTRACT

Summary: Many non-synonymous single nucleotide polymorphisms (nsSNPs) in humans are suspected to impact protein function. Here, we present a publicly available server implementation of the method SNAP (screening for non-acceptable polymorphisms) that predicts the functional effects of single amino acid substitutions. SNAP identifies over 80% of the *non-neutral* mutations at 77% accuracy and over 76% of the *neutral* mutations at 80% accuracy at its default threshold. Each prediction is associated with a reliability index that correlates with accuracy and thereby enables experimentalists to zoom into the most promising predictions.

Availability: Web-server: http://www.rostlab.org/services/SNAP; downloadable program available upon request.

Contact: bromberg@rostlab.org

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

1 INTRODUCTION

Non-synonymous SNPs (nsSNPs) are associated with disease: Estimates expect as many as 200 000 nsSNPs in human (Halushka *et al*., 1999) and about 24 000–60 000 in an individual (Cargill *et al*., 1999); this implies about 1–2 mutants per protein. While most of these likely do not alter protein function (Ng and Henikoff, 2006), many non-neutral nsSNPs contribute to individual fitness. Disease studies typically face the challenge finding a needle (SNP yielding particular phenotype) in a haystack (all known SNPs). For example, many of the thousands of mutations associated with cancer do not actually lead to the disease. Evaluating functional effects of known nsSNPs is essential for understanding genotype/phenotype relations and for curing diseases. Computational mutagenesis methods can be useful in this endeavor if they can explain the motivation behind assigning a mutant to *neutral* or *non-neutral* class or if they can provide a measure for the reliability of a particular prediction.

Screening for non-acceptable polymorphisms is accurate and provides a measure of reliability: here, we present the first web-server implementation of SNAP (screening for non-acceptable polymorphisms), a method that combines many sequence analysis tools in a battery of neural networks to predict the functional effects of nsSNPs (Bromberg and Rost, 2007, 2008). SNAP was developed using annotations extracted from PMD, the Protein Mutant Database (Kawabata *et al*., 1999; Nishikawa *et al*., 1994). SNAP needs only sequence as input; it uses sequence-based predictions of solvent accessibility and secondary structure from PROF (Rost, 2000, unpublished data; Rost, 2005; Rost and Sander, 1994), flexibility from PROFbval (Schlessinger *et al.*, 2006), functional effects from SIFT (Ng and Henikoff, 2003), as well as conservation information from PSI-BLAST (Altschul *et al*., 1997) and PSIC (Sunyaev *et al*., 1999), and Pfam annotations (Bateman *et al*., 2004). If available, SNAP can also benefit from SwissProt annotations (Bairoch and Apweiler, 2000). In sustained cross-validation, SNAP correctly identified ∼80% of the non-neutral substitutions at 77% accuracy (often referred to as specificity, i.e. correct non-neutral predictions/all predicted as non-neutral) at its default threshold. When we increase the threshold, accuracy rises at the expense of coverage (fewer of the observed non-neutral nsSNPs are identified). This balance is reflected in a crucial new feature, the reliability index (RI) for each SNAP prediction that ranges from 0 (low) to 9 (high):

$$\mathrm{RI} = \mathrm{INT}(\mathrm{OUT}_{\mathrm{non-neutral}} - \mathrm{OUT}_{\mathrm{neutral}})/10 \qquad (1)$$

where OUT_X is the raw value of one of the two SNAP output units.

When given alternative prediction methods, investigators often identify a subset of predictions for which methods agree. This approach may increase accuracy over any single method at the expense of coverage. Well-calibrated method-internal reliability indices can be much more efficient than a combination of different methods (Rost and Eyrich, 2001). Simply put: 'A basket of rotten fruit does not make for a good fruit salad' (Chris Sander, CASP1). The SNAP RI has been carefully calibrated.

2 INPUT/OUTPUT

Users submit the wild-type sequence along with their mutants. A comma-separated list gives mutants as: XiY, where X is the wild-type amino acid, Y is the mutant and *i* is the number of the residue (*i* = 1 for N-terminus). X is not required and a star (*) can replace either *i* or Y. Any combination of characters following these rules is acceptable; e.g. X** = replace all residues **X** in **all positions** by **all other amino acids**, *Y = replace **all residues** in **all positions** by **Y**. Users may provide a threshold for the minimal RI [Equation (1)] and/or for the expected accuracy of predictions that will be reported back. These two values correlate; when both are provided, the server chooses the one yielding better predictions. For each

*To whom correspondence should be addressed.

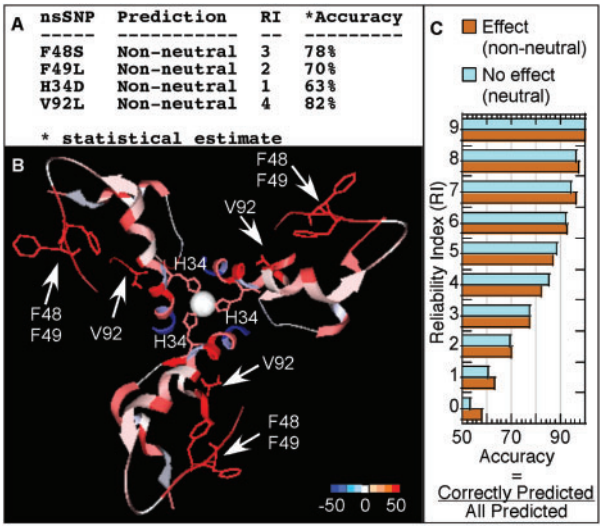

Fig. 1. Examples of SNAP functionality**.** (**A**) SNAP-server predictions for mutations in INS_HUMAN associated with hyperproinsulenemia and diabetes-mellitus type II (Chan *et al*., 1987; Sakura *et al*., 1986; Shoelson *et al*., 1983). (**B**) SNAP predictions for comprehensive *in silico* mutagenesis (all-to-alanine). The crystal structure [PDB 2omg (Norrman *et al*., 2007)] shows an insulin NPH hexamer [insulin co-crystallized with zinc (sphere at the center) in presence of protamine/urea (not highlighted); picture produced by GRASP2 (Petrey and Honig, 2003)]. Red represents mutations predicted as *non-neutral* and blue represents *neutral* predictions. Residues in wire depiction are the same as in (A): V92, H34, F48 and F49 of INS_HUMAN (A chain V3, B chain H10, F24 and F25). SNAP predicts all of these to impact function when mutated to alanine. (**C**) More reliably predicted residues are predicted more accurately: for instance, >90% of the predictions with a reliability index = 6 are expected to be right.

mutant, SNAP returns three values (Fig. 1A): the binary prediction (neutral/non-neutral), the RI (range 0–9) and the *expected accuracy* that estimates accuracy [Equation (1)] on a large dataset at the given RI (i.e. accuracy of test set predictions calculated for each neutral and non-neutral RI; Fig. 1C, Supplementary Online Material Fig. SOM_1).

At this point, SNAP may take more than an hour to return results (processing status can be tracked on the original submission page). Therefore, most requests will be answered by an email containing a link to the results page. It is also highly recommended to check existing mutant evaluations [available immediately under the 'known variants' tab; referenced by RefSeq id (Pruitt *et al*., 2007) and dbSNP id (Sherry *et al*., 2001)] prior to submitting sequences for processing. In the near future, PredictProtein (Rost *et al*., 2004) that provides the framework for SNAP, will store sequences and retrieve predictions for additional mutants in real time. Full sequence analysis (e.g. *in silico* alanine scans; Fig. 1B) is possible for short proteins ($\leq$150 total mutants/protein) via applicable server query. Analysis of longer sequences and/or local SNAP installation is currently available through the authors.

ACKNOWLEDGEMENTS

Thanks to Jinfeng Liu (Genentech) and Andrew Kernytsky (Columbia) for technical assistance; to Chani Weinreb, Marco Punta, Avner Schlessinger (all Columbia) and Dariusz Przybylski (Broad Inst.) for helpful discussions. Particular thanks to Rudolph L. Leibel (Columbia) for crucial support and discussions.

Funding: National Library of Medicine (grant 5-RO1-LM007 329-04).

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al*. (1997) Gapped Blast and PSI-Blast: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.

Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.

Bateman,A. *et al*. (2004) The Pfam Protein Families Database. *Nucleic Acids Res.*, **32**, D138–D141.

Bromberg,Y. and Rost,B. (2007) SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Res*., **35**, 3823–3835.

Bromberg,Y. and Rost,B. (2008) Comprehensive in silico mutagenesis highlights functionally improtant residues in proteins. *Bioinformatics*, **24**, i207–i212.

Cargill,M. *et al*. (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.*, **22**, 231–238.

Chan,S.J. *et al*. (1987) A mutation in the B chain coding region is associated with impaired proinsulin conversion in a family with hyperproinsulinemia. *Proc. Natl Acad. Sci. USA*, **84**, 2194–2197.

Halushka,M.K. *et al*. (1999) Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nat. Genet.*, **22**, 239–247.

Kawabata,T. *et al*. (1999) The protein mutant database. *Nucleic Acids Res*., **27**, 355–357.

Ng,P.C. and Henikoff,S. (2003) SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.*, **31**, 3812–3814.

Ng,P.C. and Henikoff,S. (2006) Predicting the effects of amino acid substitutions on protein function. *Annu. Rev. Genomics Hum. Genet.*, **7**, 61–80.

Nishikawa,K. *et al*. (1994) Constructing a protein mutant database. *Protein Eng.*, **7**, 773.

Norrman,M. *et al*. (2007) Structural characterization of insulin NPH formulations. *Eur. J. Pharm. Sci.*, **30**, 414–423.

Petrey,D. and Honig,B. (2003) GRASP2: visualization, surface properties, and electrostatics of macromolecular structures and sequences. *Methods Enzymol.*, **374**, 492–509.

Pruitt,K.D. *et al*. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.

Rost,B. (2005) How to use protein 1D structure predicted by PROFphd. In Walker,J.E. (ed.) *The Proteomics Protocols Handbook*. Humana, Totowa, NJ, pp. 875-901.

Rost,B. and Eyrich,V. (2001) EVA: large-scale analysis of secondary structure prediction. *Proteins Struct. Funct. Genet.*, **45** (Suppl. 5), S192–S199.

Rost,B. and Sander,C. (1994) Conservation and prediction of solvent accessibility in protein families. *Proteins Struct. Funct. Genet.*, **20**, 216–226.

Rost,B. *et al*. (2004) The PredictProtein server. *Nucleic Acids Res*., **32**, W321–W326.

Sakura,H. *et al*. (1986) Structurally abnormal insulin in a diabetic patient. Characterization of the mutant insulin A3 (Val----Leu) isolated from the pancreas. *J. Clin. Invest*., **78**, 1666–1672.

Schlessinger,A. *et al*. (2006) PROFbval: predict flexible and rigid residues in proteins. *Bioinformatics*, **22**, 891–893.

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

Shoelson,S. *et al*. (1983) Identification of a mutant human insulin predicted to contain a serine-for-phenylalanine substitution. *Proc. Natl Acad. Sci. USA*, **80**, 7390–7394.

Sunyaev,S.R. *et al*. (1999) PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. *Protein Eng*., **12**, 387–394.