W480–W482 *Nucleic Acids Research, 2005, Vol. 33, Web Server issue*
doi:10.1093/nar/gki372

nsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms

Lei Bao, Mi Zhou and Yan Cui*

Department of Molecular Sciences, Center of Genomics and Bioinformatics, University of Tennessee Health Science Center, 858 Madison Avenue, Memphis, TN 38163, USA

Received January 21, 2005; Revised February 16, 2005; Accepted March 7, 2005

ABSTRACT

Nonsynonymous single nucleotide polymorphisms (nsSNPs) are prevalent in genomes and are closely associated with inherited diseases. To facilitate identifying disease-associated nsSNPs from a large number of neutral nsSNPs, it is important to develop computational tools to predict the nsSNP's phenotypic effect (disease-associated versus neutral). nsSNPAnalyzer, a web-based software developed for this purpose, extracts structural and evolutionary information from a query nsSNP and uses a machine learning method called Random Forest to predict the nsSNP's phenotypic effect. nsSNPAnalyzer server is available at http://snpanalyzer.utmem.edu/.

INTRODUCTION

Assessing susceptibility to diseases based on an individual's genotype has long been a central theme of genetics studies. Among inherited gene variations in humans, nonsynonymous single nucleotide polymorphisms (nsSNPs) that lead to an amino acid change in the protein product are most relevant to human inherited diseases (1). nsSNPs can be classified into two categories according to their phenotypic effects: those that cause deleterious effects on protein functions and are hence disease-associated and those that are functionally neutral. Given the huge number of nsSNPs already discovered (2,3), a major challenge is to predict which of them are potentially disease associated. Computational tools have been developed to predict the nsSNP's phenotypic effect, e.g. the SIFT server (4) and the PolyPhen server (5). Recently, studies have shown that combining information obtained from multiple sequence alignment and three-dimensional protein structure can increase the prediction accuracy (6). nsSNPAnalyzer server integrates multiple sequences alignment and protein structure analysis to identify disease-associated nsSNPs. nsSNPAnalyzer takes a protein sequence and the accompanying nsSNP as inputs and reports whether the nsSNP is likely to be disease-associated or functionally neutral. nsSNPAnalyzer also provides additional useful information about the nsSNP to facilitate the biological interpretation of results, e.g. structural environment class and multiple sequence alignment.

PROGRAM DESCRIPTION

Algorithm and implementation

nsSNPAnalyzer is a web server implementing machine learning methods for nsSNP classification. The program design and data flow are illustrated in Figure 1. Briefly, on receiving

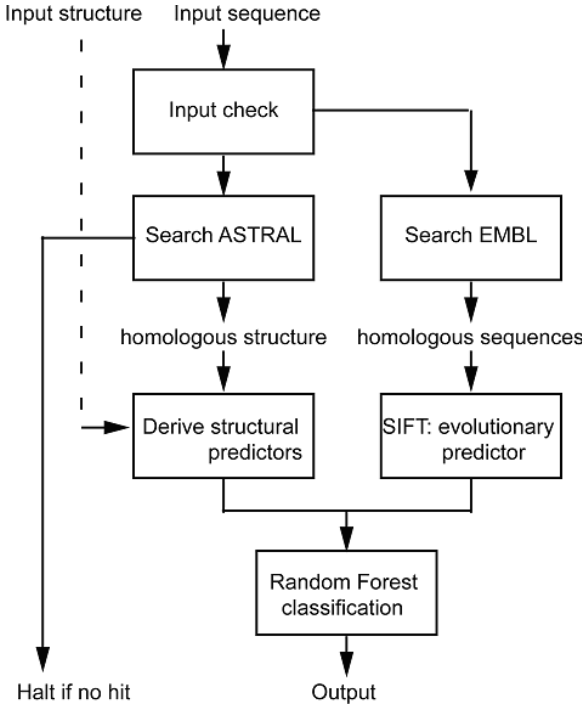

Figure 1. The program design and data flow of nsSNPAnalyzer.

*To whom correspondence should be addressed. Tel: +1 901 448 3240; Fax: +1 901 4487360; Email: ycui2@utmem.edu

A

Result:

SNP	Phenotype	Environment	AreaBuried	FracPolar	Secondstr	ScopLink	Sift_Score	
D7N	Neutral	P2C	0.185	0.750	C	d1kmaa_	1.00	View Alignment
V176M	Disease	B3S	0.556	0.500	S	d1kmaa_	0.02	View Alignment

B

```
CLUSTAL W(1.81) multiple sequence alignment

QUERY/2-12                                      QTPAFDKPKVE
Q6GP70/1-11                                     ESKAFNKPKVE
Q6DG22/1-9                                      --PAFDKPKVE
Q6IWY7/1-6                                      -----NFPKVE
Q17747/1-7                                      ----LNFPKVE
Q86NI2/1-7                                      ----LNFPKVE
Q90YG3/1-11                                     EQIVFNKPKIE
                                                     : **:*
```

Figure 2. The output of nsSNPAnalyzer. (**A**) The main output page of nsSNPAnalyzer. The user can click the icon to see the interpretation of each field. (**B**) An example of local sequence alignment spanning the nsSNP (D7N). The original amino acid (D) is highlighted in blue, and the mutated amino acid (N) is highlighted in red.

the input sequence, nsSNPAnalyzer searches the ASTRAL database (7) for homologous protein structures. This step is skipped if the users provide the protein structure themselves. nsSNPAnalyzer calculates three types of information from user's input: (i) the structural environment of the SNP, including the solvent accessibility, environmental polarity and secondary structure (8); (ii) the normalized probability of the substitution in the multiple sequence alignment (9); and (iii) the similarity and dissimilarity between the original amino acid and mutated amino acid. nsSNPAnalyzer then uses a machine learning method called Random Forest (10) to classify the nsSNPs. Random Forest is a classifier consisting of an ensemble of tree-structured classifiers. The Random Forest classifier was trained to optimally combine the heterogeneous sources of predictors using a curated training dataset prepared from the SwissProt database (11). Several recent studies have demonstrated the better performance of Random Forest over other machine learning approaches (12–14). For the nsSNP phenotypic effect prediction, we also found that Random Forest gave the best results on this training dataset. In a cross-validation test, the false positive rate is 38% and the false negative rate is 21% (15). The nsSNPAnalyzer web server is implemented on a Linux Redhat 8.0 platform with the Common Gateway Interface scripts written in PHP.

Input

Two inputs are mandatory: protein sequence in FASTA format and the nsSNP identities to be analyzed. An nsSNP is denoted as X#Y, where X is the original amino acid in one letter, # is the position of the substitution (starting from 1), and Y is the mutated amino acid in one letter. Multiple nsSNPs in a protein should be separated by new-line characters. Users may provide the inputs by copy-paste or file uploading. In addition to the two mandatory inputs, users may also upload an accompanying protein structure file in PDB format if they want their own structure to be used. Finally, because the calculation usually takes a while, users may provide their email addresses to avoid waiting online. The results are sent to the email address when the calculations are finished. Users can use the sample data to learn the input format and perform a demo run.

Output

The results of nsSNPAnalyzer are displayed on a web page and stored on the server for a week. A link to the results page can also be sent to the user via email. A sample output is shown in Figure 2. The output includes several calculated features of the nsSNP: (i) predicted phenotypic class (disease-associated versus neutral); (ii) a hyperlink to the homologous structure with a SCOP identifier (7); (iii) the normalized probability of the substitution calculated by the SIFT program (4); (iv) area buried score, a measure of the solvent accessibility; (v) fraction polar score, a measure of environmental polarity related to hydrogen bond formation; (vi) secondary structure (helix, sheet and coil); and (vii) the structural environment class, a discrete environment class definition by combining features (iv)–(vi) (8). The area buried score and fraction polar score are calculated by the ENVIRONMENT program (8), and the secondary structure is calculated by the STRIDE program (16). The user can click the 'View Alignment' button to see the local sequence alignment spanning the substitution sites and get a direct sight on the mutability of the substitution. The original amino acid is highlighted in blue, and the mutated amino acid is highlighted in red.

FUTURE PLANS

Considering the remarkable CPU cost of calculation, we are planning to provide precalculated results for all human nsSNPs in the dbSNP (17) with homologous structures available. We will also test the applicability of extracting structural predictors from predicted structures to eliminate the requirement of having experimentally determined structures available.

ACKNOWLEDGEMENTS

We thank Drs James Bowie, Roland Luethy and David Eisenberg for providing the computer program for calculating the structural environments. We thank Drs Pauline Ng and Steven Henikoff for providing access to the SIFT program. We thank Drs Leo Breiman, Andy Liaw and Matthew Wiener for providing access to the Random Forest package. We thank the anonymous reviewers for their helpful suggestions. This work was partly supported by a Phrma Foundation grant to Y.C. Funding to pay the Open Access publication charges for this article was provided by the faculty startup grant from UTHSC.

Conflict of interest statement. None declared.

REFERENCES

1. Stenson,P.D., Ball,E.V., Mort,M., Phillips,A.D., Shiel,J.A, Thomas,N.S., Abeysinghe,S., Krawczak,M. and Cooper,D.N. (2003) Human Gene Mutation Database (HGMD): 2003 update. *Hum. Mutat*., **21**, 577–581.
2. Fredman,D., Siegfried,M., Yuan,Y.P., Bork,P., Lehvaslaiho,H. and Brookes,A.J. (2002) HGVbase: a human sequence variation database emphasizing data quality and a broad spectrum of data sources. *Nucleic Acids Res*., **30**, 387–391.
3. Irizarry,K., Kustanovich,V., Li,C., Brown,N., Nelson,S., Wong,W. and Lee,C.J. (2000) Comprehensive EST analysis of single nucleotide polymorphism across coding regions of the human genome. *Nature Genet*., **26**, 233–236.
4. Ng,P.C. and Henikoff,S. (2003) SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res*., **31**, 3812–3814.
5. Ramensky,V., Bork,P. and Sunyaev,S. (2002) Human non-synonymous SNPs: server and survey. *Nucleic Acids Res*., **30**, 3894–3900.
6. Saunders,C.T. and Baker,D. (2002) Evaluation of structural and evolutionary contributions to deleterious mutation prediction. *J. Mol. Biol*., **322**, 891–901.
7. Chandonia,J.M., Hon,G., Walker,N.S., Lo Conte,L., Koehl,P., Levitt,M. and Brenner,S.E. (2004) The ASTRAL compendium in 2004. *Nucleic Acids Res*., **32**, D189–D192.
8. Bowie,J.U., Luthy,R. and Eisenberg,D. (1991) A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, **253**, 164–170.
9. Ng,P.C. and Henikoff,S. (2001) Predicting deleterious amino acid substitutions. *Genome Res*., **11**, 863–874.
10. Breiman,L. (2001) Random Forest. Technical Report, Stat. Dept. UCB.
11. Yip,Y.L., Scheib,H., Diemand,A.V., Gattiker,A., Famiglietti,L.M., Gasteiger,E. and Bairoch,A. (2003) The Swiss-Prot variant page and the ModSNP database: a resource for sequence and structure information on human protein variants. *Hum. Mutat*., **23**, 464–470.
12. Svetnik,V., Liaw,A., Tong,C., Culberson,J.C., Sheridan,R.P. and Feuston,B.P. (2003) Random Forest: a classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci*., **43**, 1947–1958.
13. Wu,B., Abbott,T., Fishman,D., McMurray,W., Mor,G., Stone,K., Ward,D., Williams,K. and Zhao,H. (2003) Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics*, **19**, 1636–1643.
14. Gunther,E.C., Stone,D.J., Gerwien,R.W., Bento,P. and Heyes,M.P. (2003) Prediction of clinical drug efficacy by classification of drug-induced genomic expression profiles *in vitro*. *Proc. Natl Acad. Sci. USA*, **100**, 9608–9613.
15. Bao,L. and Cui,Y. (2005) Prediction of the phenotypic effects of nonsynonymous single nucleotide polymorphisms using structural and evolutionary information. *Bioinformatics*, doi:10.1093/bioinformatics/bti365.
16. Frishman,D. and Argos,P. (1995) Knowledge-based protein secondary structure assignment. *Proteins*, **23**, 566–579.
17. 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.