3812–3814 *Nucleic Acids Research, 2003, Vol. 31, No. 13*
DOI: 10.1093/nar/gkg509

SIFT: predicting amino acid changes that affect protein function

Pauline C. Ng and Steven Henikoff*

Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue N A1-162, Seattle, WA 98109, USA

Received January 4, 2003; Revised and Accepted February 28, 2003

ABSTRACT

Single nucleotide polymorphism (SNP) studies and random mutagenesis projects identify amino acid substitutions in protein-coding regions. Each substitution has the potential to affect protein function. SIFT (Sorting Intolerant From Tolerant) is a program that predicts whether an amino acid substitution affects protein function so that users can prioritize substitutions for further study. We have shown that SIFT can distinguish between functionally neutral and deleterious amino acid changes in mutagenesis studies and on human polymorphisms. SIFT is available at http://blocks.fhcrc.org/sift/SIFT.html.

INTRODUCTION

Single nucleotide polymorphisms (SNPs) are used as markers in linkage and association studies to detect which regions in the human genome may be involved in disease. SNPs in coding and regulatory regions may be implicated in disease themselves. Non-synonymous SNPs that lead to an amino acid change in the protein product are of major interest, because amino acid substitutions currently account for approximately half of the known gene lesions responsible for human inherited disease (1). SIFT (Sorting Intolerant From Tolerant) uses sequence homology to predict whether an amino acid substitution will affect protein function and hence, potentially alter phenotype (2,3).

SIFT has been applied to human variant databases and was able to distinguish mutations involved in disease from neutral polymorphisms (3). Assuming that disease-causing amino acid substitutions are damaging to protein function, we applied SIFT to a database of missense substitutions associated with or involved in disease (4). SIFT predicted 69% to be damaging. When SIFT was applied to the non-synonymous SNPs in dbSNP (5), a database of putative SNPs, 25% of the variants were predicted to be deleterious. This was similar to SIFT's 20% false positive error which suggested that most non-synonymous SNPs are functionally neutral. Furthermore, a subset of the variants from dbSNP predicted to affect function were involved in disease which confirmed SIFT sensitivity.

The SIFT algorithm relies solely on sequence for prediction, yet performs similarly to tools that use structure (3,6–8). An advantage of not requiring structure is that a larger number of substitutions can be predicted on. Of the non-synonymous SNPs identified by the SNP Consortium, 74% were sufficiently similar to homologs in protein sequence databases for SIFT prediction. The number of substitutions that SIFT can predict on is expected to increase as more genomes are sequenced and more protein sequences become available.

SIFT PREDICTION METHOD

SIFT presumes that important amino acids will be conserved in the protein family, and so changes at well-conserved positions tend to be predicted as deleterious. For example, if a position in an alignment of a protein family only contains the amino acid isoleucine, it is presumed that substitution to any other amino acid is selected against and that isoleucine is necessary for protein function. Therefore, a change to any other amino acid will be predicted to be deleterious to protein function. If a position in an alignment contains the hydrophobic amino acids isoleucine, valine and leucine, then SIFT assumes, in effect, that this position can only contain amino acids with hydrophobic character. At this position, changes to other hydrophobic amino acids are usually predicted to be tolerated but changes to other residues (such as charged or polar) will be predicted to affect protein function.

To predict whether an amino acid substitution in a protein will affect protein function, SIFT considers the position at which the change occurred and the type of amino acid change. Given a protein sequence, SIFT chooses related proteins and obtains an alignment of these proteins with the query. Based on the amino acids appearing at each position in the alignment, SIFT calculates the probability that an amino acid at a position is tolerated conditional on the most frequent amino acid being tolerated. If this normalized value is less than a cutoff, the substitution is predicted to be deleterious (2). The SIFT algorithm and software have been described previously (2,3).

SIFT WEBSITE

Input

Users can obtain predictions for amino acid changes of interest at http://www.blocks.fhcrc.org/sift/SIFT.html. From this page,

*To whom correspondence should be addressed. Tel: +1 2066674515; Fax: +1 2066675889; Email: steveh@fhcrc.org


```
Substitution at pos 1426 from S to P is predicted to AFFECT PROTEIN FUNCTION with a score of 0.02.
    Median sequence conservation: 2.90
    Sequences represented at this position:26

Substitution at pos 1432 from N to K is predicted to be TOLERATED with a score of 0.08.
    Median sequence conservation: 2.90
    Sequences represented at this position:26

Substitution at pos 1445 from D to N is predicted to AFFECT PROTEIN FUNCTION with a score of 0.01.
    Median sequence conservation: 3.66
    Sequences represented at this position:21
WARNING!!  This substitution may have been predicted to affect function just because
 the sequences used were not diverse enough.  There is LOW CONFIDENCE in this prediction.
```

Figure 1. An example of SIFT prediction on amino acid changes in a protein. Substitutions with score less than 0.05 are predicted to affect protein function. In the last prediction, the median conservation of the sequences does not meet the threshold so a warning is issued.

there are links to three submission pages which allow users different levels of involvement in order to control the quality of their predictions.

For minimal involvement, users can simply submit their protein sequences and amino acid substitutions. In its fully automated mode, SIFT will search for protein sequences homologous to the query protein and based on these sequences, calculate probabilities for each possible amino acid change. Users can select from among SWISS-PROT, SWISS-PROT/TrEMBL, or NCBI's non-redundant protein databases for SIFT to search (4,9).

Although SIFT can choose sequences automatically, better prediction results may be obtained when all of the sequences that are provided are orthologous to the query protein. This is because inclusion of paralogous sequences confounds prediction at residues conserved only among the orthologues. If a user already has sequences that are thought to be functionally similar to the protein of interest, these sequences can be directly submitted and SIFT's step for choosing sequences skipped. Given the query protein and homologous sequences, SIFT obtains the alignment.

If regions are misaligned, SIFT will not recognize conserved positions and therefore miss potentially damaging substitutions. For best prediction quality, a third mode of operation allows users to submit their own alignments.

Output

Predictions are given for all 20 possible amino acid changes at each position in the protein. The alignment is also returned so that users can examine the sequences used for prediction and modify them for resubmission. This option is also useful for removing uncertain, erroneous and misaligned sequences from alignment output generated by SIFT in its automatic mode.

For amino acid substitutions submitted by the user, a more detailed synopsis is provided (Fig. 1). The score is the normalized probability that the amino acid change is tolerated. SIFT predicts substitutions with scores less than 0.05 as deleterious. Some SIFT users have found that substitutions with scores less than 0.1 provide better sensitivity for detecting deleterious SNPs (Cornelia Ulrich, personal communication and 10). The quantitative score allows users to prioritize their amino acid changes by ranking them from the lowest scores to the highest.

Confidence in a substitution predicted to be deleterious depends on the diversity of the sequences in the alignment. If the sequences used for prediction are closely related, then many positions will appear conserved and SIFT will predict most substitutions to affect protein function. This leads to a high false positive error where functionally neutral substitutions are predicted to be deleterious.

To alert the user to these situations, SIFT calculates the median conservation value which measures the diversity of the sequences in the alignment. Conservation, as measured by information content (11), is calculated for each position in the alignment and the median of these values is obtained. Conservation ranges from $\log_2 20$ ($=4.32$), when a position is completely conserved and only one amino acid is observed, to zero, when all 20 amino acids are observed at a position. By default, SIFT builds alignments with a median conservation value of 3.0. Predictions based on sequence alignments with higher median conservation values are less diverse and will have a higher false positive error (Fig. 2).

Even if there are few homologous sequences available, SIFT performs better than simply predicting non-conservative amino acid substitutions as deleterious, where non-conservative changes are defined as having negative scores in an amino acid substitution scoring matrix. We have shown that with only one sequence homologous to the test protein, SIFT can predict twice as many neutral substitutions correctly compared to a substitution scoring matrix (2). Even with few homologous sequences, there will be positions that differ between the test protein and the other sequences. Depending on the amino acids appearing at these positions, SIFT may predict these positions to be unimportant for protein function. This additional information can eliminate functionally neutral substitutions and increase selectivity to deleterious substitutions.

In summary, a large number of substitutions can be obtained from mutagenesis projects, SNP datasets, and changes between closely related organisms. When it is not feasible to conduct experiments on all substitutions, SIFT and other similar prediction tools (13) may be useful in prioritizing which

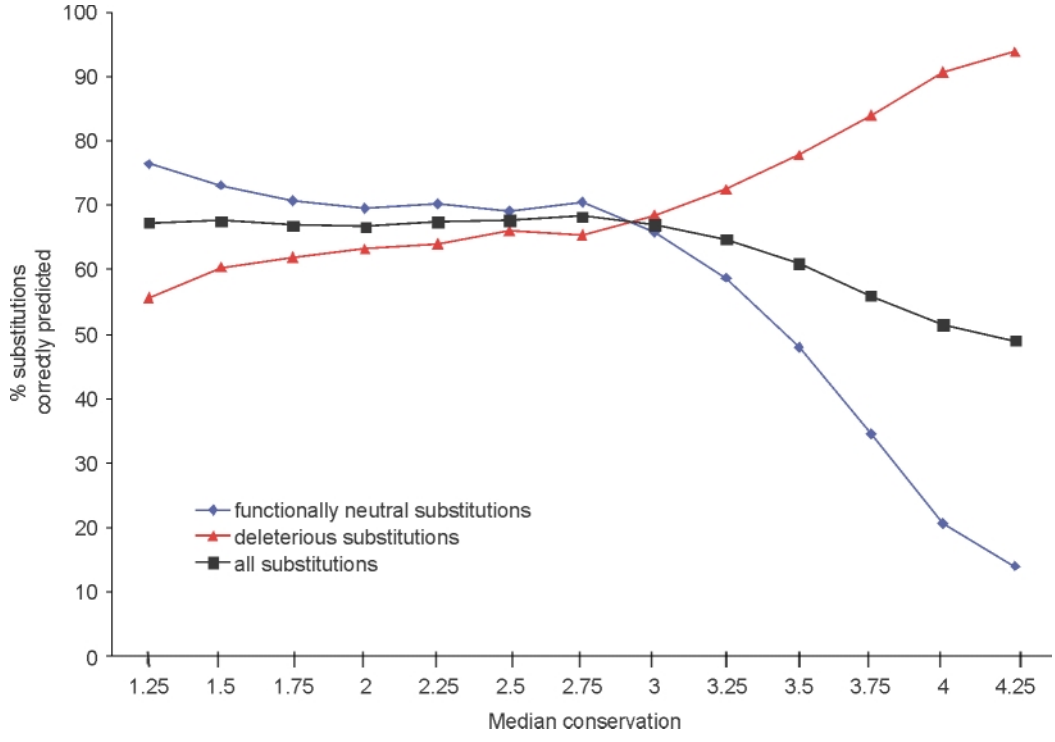

Figure 2. Prediction depends on the diversity of the sequences used in the alignment. Percentage of substitutions correctly predicted is based on over 4000 substitutions that were assayed throughout the LacI protein of *Escherichia coli* (2,12). When the sequences in the alignment used for prediction are closely related (high median conservation) then many positions appear conserved and important for function. In this situation, prediction accuracy on deleterious substitutions is high but many functionally neutral substitutions are erroneously predicted to be deleterious. To obtain an alignment with a specified median conservation, the LacI protein sequence of *E.coli* was submitted to the SIFT website and the median conservation setting adjusted. Because the homologous sequences available are distantly related to *E.coli* LacI, alignments with higher median conservation values could not be obtained. In order to obtain alignments with median conservation values more than 3.25, closely related sequences were simulated by starting with an alignment of identical *E.coli* LacI sequences. A position and a sequence were randomly selected from the LacI alignment with median conservation 2.75. The amino acid corresponding to this location was substituted in the starting alignment. Amino acids continued to be randomly selected and substituted until the desired median conservation was met. The simulated alignment was then evaluated for its performance as previously described (2) and the plotted value is the average performance of 100 simulated alignments.

changes affect protein function and may contribute to phenotypic differences.

ACKNOWLEDGEMENTS

We thank Jorja Henikoff for advice and encouragement. This work was supported by a grant from NIH (GM29009).

REFERENCES

1. Krawczak,M., Ball,E.V., Fenton,I., Stenson,P.D., Abeysinghe,S., Thomas,N. and Cooper,D.N. (2000) Human gene mutation database-a biomedical information and research resource. *Hum. Mutat.*, **15**, 45–51.
2. Ng,P.C. and Henikoff,S. (2001) Predicting deleterious amino acid substitutions. *Genome Res.*, **11**, 863–874.
3. Ng,P.C. and Henikoff,S. (2002) Accounting for human polymorphisms predicted to affect protein function. *Genome Res.*, **12**, 436–446.
4. Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.
5. 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.
6. Sunyaev,S., Ramensky,V., Koch,I., Lathe,W.,III, Kondrashov,A.S. and Bork,P. (2001) Prediction of deleterious human alleles. *Hum. Mol. Genet.*, **10**, 591–597.
7. Chasman,D. and Adams,R.M. (2001) Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. *J. Mol. Biol.*, **307**, 683–706.
8. Saunders,C.T. and Baker,D. (2002) Evaluation of structural and evolutionary contributions to deleterious mutation prediction. *J. Mol. Biol.*, **322**, 891–901.
9. Wheeler,D.L., Church,D.M., Lash,A.E., Leipe,D.D., Madden,T.L, Pontius,J.U., Schuler,G.D., Schriml,L.M., Tatusova,T.A., Wagner,L. and Rapp,B.A. (2002) Database resources of the National Center for Biotechnology Information: 2002 update. *Nucleic Acids Res.*, **30**, 13–16.
10. Leabman,M.K., Huang,C.C., DeYoung,J., Carlson,E.J., Taylor,T., de la Cruz,M., Johns,S.J., Stryke,D., Kawamoto,M., Urban,T.J., *et al.* (2003) Natural variation in human membrane transporter genes reveals evolutionary and functional constraints. *Proc. Natl Acad. Sci. USA*, in press.
11. Schneider,T.D., Stormo,G.D., Gold,L. and Ehrenfeucht,A. (1986) Information content of binding sites on nucleotide sequences. *J. Mol. Biol.*, **188**, 415–431.
12. Pace,H.C., Kercher,M.A., Lu,P., Markiewicz,P., Miller,J.H., Chang,G. and Lewis,M. (1997) Lac repressor genetic map in real space. *Trends Biochem. Sci.*, **22**, 334–339.
13. Ramensky,V., Bork,P. and Sunyaev, S. (2002) Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.*, **30**, 3894–3900.