Published in final edited form as:
Nat Methods. 2010 April ; 7(4): 248–249. doi:10.1038/nmeth0410-248.

A method and server for predicting damaging missense mutations

Ivan A. Adzhubei[1,7], **Steffen Schmidt**[2,7], **Leonid Peshkin**[3,7], **Vasily E. Ramensky**[4], **Anna Gerasimova**[5], **Peer Bork**[6], **Alexey S. Kondrashov**[5], and **Shamil R. Sunyaev**[1]

[1]Division of Genetics, Brigham & Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA

[2]Department of Biochemistry, Max Planck Institute for Developmental Biology, Tübingen, Germany

[3]Department of Systems Biology, Harvard Medical School, Boston, Massachusetts, USA

[4]Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow, Russia

[5]Life Sciences Institute and Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, USA

[6]European Molecular Biology Laboratory, Heidelberg, Germany

To the Editor:

Applications of rapidly advancing sequencing technologies exacerbate the need to interpret individual sequence variants. Sequencing of phenotyped clinical subjects will soon become a method of choice in studies of the genetic causes of Mendelian and complex diseases. New exon capture techniques will direct sequencing efforts towards the most informative and easily interpretable protein-coding fraction of the genome. Thus, the demand for computational predictions of the impact of protein sequence variants will continue to grow.

Here we present a new method and the corresponding software tool, PolyPhen-2 (http://genetics.bwh.harvard.edu/pph2/), which is different from the early tool PolyPhen1 in the set of predictive features, alignment pipeline, and the method of classification (Fig. 1a). PolyPhen-2 uses eight sequence-based and three structure-based predictive features (Supplementary Table 1) which were selected automatically by an iterative greedy algorithm (Supplementary Methods). Majority of these features involve comparison of a property of the wild-type (ancestral, normal) allele and the corresponding property of the mutant (derived, disease-causing) allele, which together define an amino acid replacement. Most informative features characterize how well the two human alleles fit into the pattern of amino acid replacements within the multiple sequence alignment of homologous proteins, how distant the protein harboring the first deviation from the human wild-type allele is from the human protein, and whether the mutant allele originated at a hypermutable site2. The alignment pipeline selects the set of homologous sequences for the analysis using a clustering algorithm and then constructs and refines their multiple alignment (Supplementary Fig. 1). The functional significance of an allele replacement is predicted

Correspondence to: Shamil R. Sunyaev[1] ssunyaev@rics.bwh.harvard.edu.
[7]These authors contributed equally to this work

from its individual features (Supplementary Figs. 2–4) by Naïve Bayes classifier (Supplementary Methods).

We used two pairs of datasets to train and test PolyPhen-2. We compiled the first pair, HumDiv, from all 3,155 damaging alleles with known effects on the molecular function causing human Mendelian diseases, present in the UniProt database, together with 6,321 differences between human proteins and their closely related mammalian homologs, assumed to be non-damaging (Supplementary Methods). The second pair, HumVar3, consists of all the 13,032 human disease-causing mutations from UniProt, together with 8,946 human nsSNPs without annotated involvement in disease, which were treated as non-damaging.

We found that PolyPhen-2 performance, as presented by its receiver operating characteristic curves, was consistently superior compared to PolyPhen (Fig. 1b) and it also compared favorably with the three other popular prediction tools4¯6 (Fig. 1c). For a false positive rate of 20%, PolyPhen-2 achieves the rate of true positive predictions of 92% and 73% on HumDiv and HumVar, respectively (Supplementary Table 2).

One reason for a lower accuracy of predictions on HumVar is that nsSNPs assumed to be non-damaging in HumVar contain a sizable fraction of mildly deleterious alleles. In contrast, most of amino acid replacements assumed non-damaging in HumDiv must be close to selective neutrality. Because alleles that are even mildly but unconditionally deleterious cannot be fixed in the evolving lineage, no method based on comparative sequence analysis is ideal for discriminating between drastically and mildly deleterious mutations, which are assigned to the opposite categories in HumVar. Another reason is that HumDiv uses an extra criterion to avoid possible erroneous annotations of damaging mutations.

For a mutation, PolyPhen-2 calculates Naïve Bayes posterior probability that this mutation is damaging and reports estimates of false positive (the chance that the mutation is classified as damaging when it is in fact non-damaging) and true positive (the chance that the mutation is classified as damaging when it is indeed damaging) rates. A mutation is also appraised qualitatively, as benign, possibly damaging, or probably damaging (Supplementary Methods).

The user can choose between HumDiv- and HumVar-trained PolyPhen-2. Diagnostics of Mendelian diseases requires distinguishing mutations with drastic effects from all the remaining human variation, including abundant mildly deleterious alleles. Thus, HumVar-trained PolyPhen-2 should be used for this task. In contrast, HumDiv-trained PolyPhen-2 should be used for evaluating rare alleles at loci potentially involved in complex phenotypes, dense mapping of regions identified by genome-wide association studies, and analysis of natural selection from sequence data, where even mildly deleterious alleles must be treated as damaging.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by National Institutes of Health Grant R01 GM078598.

References

1. Ramensky V, Bork P, Sunyaev S. Nucleic Acids Res. 2002; 30:3894–3900. [PubMed: 12202775]

2. Schmidt S, et al. PLoS Genet. 2008; 4 e1000281.
3. Capriotti E, Calabrese R, Casadio R. Bioinformatics. 2006; 22:2729–2734. [PubMed: 16895930]
4. Ng PC, Henikoff S. Nucleic Acids Res. 2003; 31:3812–3814. [PubMed: 12824425]
5. Bromberg Y, Yachdav G, Rost B. Bioinformatics. 2008; 24:2397–2398. [PubMed: 18757876]
6. Yue P, et al. BMC Bioinformatics. 2006; 7:166. [PubMed: 16551372]

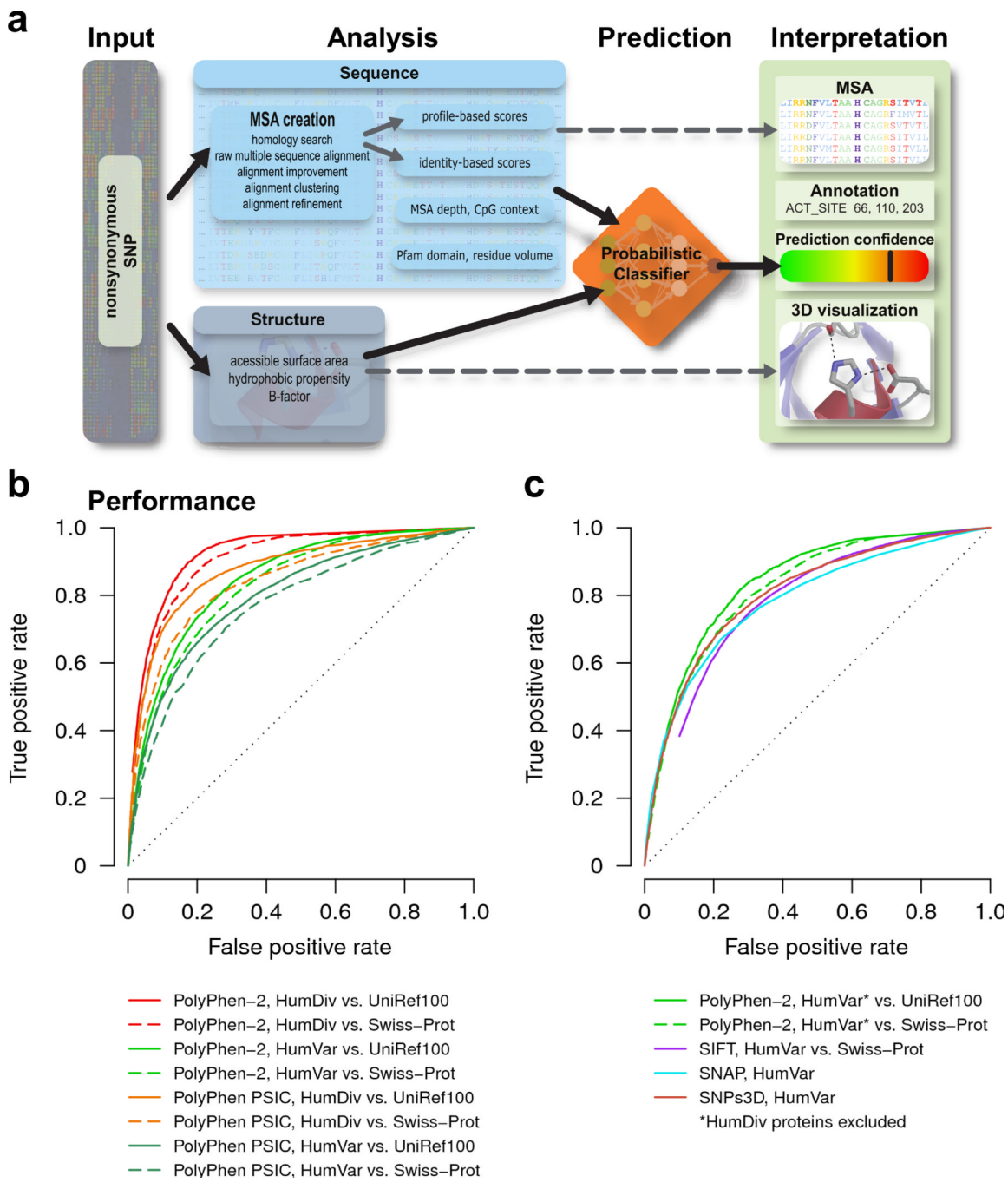

Figure 1.

PolyPhen-2 pipeline and prediction accuracy. (**a**) Overview of the algorithm. (**b**) Receiver operating characteristic (ROC) curves for predictions made by PolyPhen-2 using five-fold cross-validation on HumDiv (red) and HumVar3 (light green). UniRef100 (solid lines) and Swiss-Prot (dashed lines) databases were used for the homology search in the sequence analysis pipeline. Also shown are corresponding ROC curves for PolyPhen on HumDiv (orange) and HumVar (dark green) calculated from the difference between PSIC scores1 of the wild type and the mutant amino acid residues. (**c**) ROC curves for PolyPhen-2 trained on HumDiv and tested on a subset of HumVar non-overlapping with HumDiv (green).

UniRef100 (solid lines) and Swiss-Prot (dashed lines) databases were used for the homology search. Also shown are ROC curves for SIFT4 (blue), SNAP5 (cyan) and SNPs3D6 (brown) on HumVar. Methods other than PolyPhen-2 and PolyPhen could not easily be applied to HumDiv because using the same sequences for obtaining both multiple alignments and non-damaging replacements must be avoided. SIFT was used in conjunction with Swiss-Prot database, SNAP and SNPs3D were used with their corresponding default databases. We used SIFT with Swiss-Prot database for homology search since Swiss-Prot does not contain incomplete sequences, sequences of splice forms and sequences of human allelic variants, making it possible to guarantee that allelic variants used in testing datasets would not appear in multiple sequence alignments used in computing prediction rules by other methods.