−Table of Contents

Overview

Most of human genetic variation is represented by **SNP**s (**S**ingle-**N**ucleotide **P**olymorphisms) and many of them are believed to cause phenotypic differences between human individuals.

We specifically focus on nonsynonymous SNPs (**nsSNP**s), i.e., SNPs located in coding regions and resulting in amino acid variation in protein products of genes. It was shown in several studies that impact of amino acid allelic variants on protein structure/function can be reliably predicted via analysis of multiple sequence alignments and protein 3D-structures. As we demonstrated in an earlier work, these predictions correlate with the effect of natural selection seen as an excess of rare alleles. Therefore, predictions at the molecular level reveal SNPs affecting actual phenotypes.

PolyPhen-2 is an automatic tool for prediction of possible impact of an amino acid substitution on the structure and function of a human protein. This prediction is based on a number of features comprising the sequence, phylogenetic and structural information characterizing the substitution.

For a given amino acid substitution in a protein, PolyPhen-2 extracts various sequence and structure-based features of the substitution site and feeds them to a probabilistic classifier.

Sequence-based features

A substitution may occur at a specific site, e.g., active or binding, or in a non-globular, e.g., trans-membrane, region. PolyPhen-2 tries to identify a query protein as an entry in the human proteins subset of UniProtKB/Swiss-Prot database and use the feature table (FT) section of the corresponding entry. PolyPhen-2 checks if the amino acid replacement occurs at a site which is annotated as:

- DISULFID, CROSSLNK bond or
- BINDING, ACT_SITE, LIPID, METAL, SITE, MOD_RES, CARBOHYD, NON_STD site

At this step PolyPhen-2 memorizes all positions which are annotated in the query protein as BINDING, ACT_SITE, LIPID, and METAL. At a later stage if the search for a homologous protein with known 3D structure is successful, it is checked whether the substitution site is in spatial contact with these critical for protein function residues.

PolyPhen-2 also checks if the substitution site is located in the region annotated as:

- TRANSMEM, INTRAMEM, COMPBIAS, REPEAT, COILED, SIGNAL, PROPEP

For a substitution in an annotated or predicted trans-membrane region, PolyPhen-2 uses the PHAT trans-membrane specific matrix score to evaluate possible functional effect of a nsSNP.

PSIC profile scores for two amino acid variants

The amino acid replacement may be incompatible with the spectrum of substitutions observed at the position in the family of homologous proteins. PolyPhen-2 identifies homologues of the input sequences via BLAST search in the UniRef100 database. The set of BLAST hits is filtered to retain hits that have:

- sequence identity to the input sequence in the range 30-94%, inclusively, and

- alignment with the query sequence not smaller than 75 residues in length

Sequence identity is defined as the number of matches divided by the complete alignment length.

The resulting multiple alignment is used by the **PSIC** software (**P**osition-**S**pecific **I**ndependent **C**ounts) to calculate the so-called profile matrix. Elements of the matrix (profile scores) are logarithmic ratios of the likelihood of given amino acid occurring at a particular position to the likelihood of this amino acid occurring at any position (background frequency).

PolyPhen-2 computes the difference between profile scores of both allelic variants in the polymoprphic position. Big positive values of this difference may indicate that the studied substitution is rarely or never observed in the protein family. PolyPhen-2 also shows the number of aligned sequences at the query position. This number may be used to assess the reliability of profile score calculations.

Structural features

Mapping of amino acid replacement to the known 3D structure reveals whether the replacement is likely to destroy the hydrophobic core of a protein, electrostatic interactions, interactions with ligands or other important features of a protein. If the spatial structure of a query protein is unknown, one can use the homologous proteins with known structure.

Mapping of the substitution site to known protein 3D structures

PolyPhen-2 BLASTs query sequence against protein structure database (**PDB**) and by default retains all hits that meet the given criteria:

- sequence identity threshold is set to 50%, since this value guarantees the conservation of basic structural characteristics
- minimal hit length is set to 100
- maximal number of gaps is set to 20

By default, a hit is rejected if its amino acid at the corresponding position differs from the amino acid in the input sequence. The position of the substitution is then mapped onto the corresponding positions in all retained hits. Hits are sorted according to the sequence identity or E-value of the sequence alignment with the query protein.

Structural parameters

Further analysis performed by PolyPhen-2 is based on the use of several structural parameters. Importantly, although all parameters are reported in the output, only some of them are used in the final decision rules.

Parameters taken from DSSP

PolyPhen-2 uses **DSSP** (**D**ictionary of **S**econdary **S**tructure in **P**roteins) database to get the following structural parameters for the mapped amino acid residues:

- Secondary structure (according to the DSSP nomenclature)
- Solvent accessible surface area (absolute value in Å^2)
- Phi-psi dihedral angles

Calculated parameters

The following values are calculated by PolyPhen-2:

- Normed accessible surface area: the absolute value divided by the maximal area defined as a 99%-quantile of surface area distribution for this particular amino acid type in PDB

- Change in accessible surface propensity resulting from the substitution. Accessible surface propensities (knowledge-based hydrophobic "potentials") are logarithmic ratios of the likelihood of given amino acid occuring at a site with a particular accessibility to the likelihood of this amino acid occuring at any site (background frequency)
- Change in residue side chain volume measured in Å³. Side chain volumes are here
- Region of the phi-psi map (Ramachandran map) derived from the residue dihedral angles. Ramachandran map is here
- Normalized B-factor (temperature factor) for the residue. B-factor, or temperature factor, is used in crystallographic studies of macromolecules to characterise the "mobility" of an atom. It is believed [Chasman D, Adams RM (2001)] that the values of B-factor of a residue may be correlated with its tolerance to amino acid substitutions.

By default, all parameters above are calculated for the first hit only.

Contacts

The presence of specific spatial contacts of a residue may reveal its role for the protein function. The suggested default threshold for all contacts to be displayed in the output is 6Å. However, the value of 3Å is used in the decision rule. For evaluation of a contact between two atom sets PolyPhen-2 finds the minimal distance amongst all possible between atoms of two sets.

By default, contacts are calculated for all found hits with known structure. This is essential for the cases when several PDB entries correspond to one protein, but carry different information about complexes with other macromolecules and ligands.

PolyPhen-2 checks three types of contacts for a variable amino acid residue:

Contacts with heteroatoms: Contacts with ligands defined as all heteroatoms excluding water and "non-biological" crystallographic ligands that are believed to be related to the structure determination procedure rather than to biological function of a protein.

Interchain contacts: Interactions between subunits of the protein molecule. Technically, they are defined as contacts of a polymorphic residue with residues from other polypeptide chains present in the PDB file.

Contacts with functional sites: Third type of contacts analysed by PolyPhen-2 is represented by contacts with critical for protein function residues (BINDING, ACT_SITE, LIPID, and METAL), where the latter are derived from sequence annotation.

Prediction

PolyPhen-2 predicts the functional significance of an allele replacement from its individual features by Naïve Bayes classifier trained using supervised machine-learning.

Two pairs of datasets were used to train and test PolyPhen-2 prediction models. The first pair, **HumDiv**, was compiled from all damaging alleles with known effects on the molecular function causing human Mendelian diseases, present in the UniProtKB database, together with differences between human proteins and their closely related mammalian homologs, assumed to be non-damaging. The second pair, **HumVar**, consisted of all human disease-causing mutations from UniProtKB, together with common human nsSNPs (MAF>1%) without annotated involvement in disease, which were treated as non-damaging.

The user can choose between HumDiv- and HumVar-trained PolyPhen-2 models. Diagnostics of Mendelian diseases requires distinguishing mutations with drastic effects from all the remaining human variation, including abundant mildly deleterious alleles. Thus, HumVar-trained model should be used for this task. In contrast, HumDiv-trained model 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.

For a mutation, PolyPhen-2 calculates Naïve Bayes posterior probability that this mutation is damaging and reports estimates of false positive rate (FPR, the chance that the mutation is classified as damaging when it is in fact non-damaging) and true positive rate (TPR, the chance that the mutation is classified as damaging when it is indeed damaging). A mutation is also appraised qualitatively, as **benign**, **possibly damaging**, or **probably damaging** based on pairs of false positive rate (FPR) thresholds, optimized separately for each model (e.g., HumDiv and HumVar).

Current version 2.1.0 of the PolyPhen-2 uses 5% / 10% FPR for **HumDiv** model and 10% / 20% FPR for **HumVar** model as the thresholds for this ternary classification. Mutations with their posterior probability scores associated with estimated false positive rates at or below the first (lower) FPR value are predicted to be **probably damaging** (more confident prediction). Mutations with the posterior probabilities associated with false positive rates at or below the second (higher) FPR value are predicted to be **possibly damaging** (less confident prediction). Mutations with estimated false positive rates above the second (higer) FPR value are classified as **benign**.

If the lack of data does not allow to make a prediction then the outcome is reported as **unknown**.

Last modified: 2012/02/15 17:20