

Polyphen: prediction of functional effects of human nsSNPs

Kymberleigh A Pagel

Polymorphism Phenotyping

“an automatic tool for prediction of possible impact of an amino acid substitution on the structure and function of a human protein”

- Three main sources of data
 - FT: The sequence annotation on UniProtKB
 - Alignment and PSIC profile scores
 - Structure from PDB or PQS (only 10%)

Train and Test: Choose your own adventure

- HumDiv
 - 3,155 mutations with known effects on molecular function causing **Mendelian** disease, from the UniProtKB database
 - 6,321 differences between human and mammalian homologs, assumed not to be deleterious
- HumVar
 - 13,032 disease-causing mutations from UniProtKB
 - 8,946 common human nsSNPs not annotated to be involved with disease– assumed to be non-damaging.

Step 1: Sequence based

- Characterization of the substitution site.
 - Identify the protein in `hs_swal` – the human protein subset of Uniprot.
 - For every position in the protein, determine the site annotation, if available, and characterize as BINDING, ACT_SITE, LIPID, METAL, TRANSMEM, SIGNAL, or PROPEP. Also makes predictions for TRANSMEM, SIGNAL, PROPEP, and coiled coil regions.
 - If substitution in transmembrane region, The PHAT transmembrane specific matrix score is used to predict functional effect of a nsSNP in the transmembrane region.

Step 2: PSIC Profile Scores

- Use Blast to find homologous sequences that have a SI between 30-94% and are longer than 50 residues.
- This alignment is fed into **Position-Specific Independent Counts** software to generate a profile matrix.
 - Each position is the log ratio of the given amino acid at that position to the background frequency.
- $|\text{score of wildtype amino acid} - \text{score of mutated amino acid}|$
 - Large values indicate rare substitutions.

Step 3 part 1: Mapping to known 3D structures

- Blast the query sequence against either PDB or PQS
 - SI threshold = 50%, min hit length = 100, max number of gaps = 20
- Obtain data from **D**ictionary of **S**econdary **S**tructure in **P**roteins database
 - Secondary structure, solvent accessible surface area, Phi-psi dihedral angles
- But that's not all!

Step 3 Part 2: Calculated structural parameters

- PolyPhen calculates
 - Normed accessible surface area
 - Change in accessible surface propensity
 - Change in residue side chain volume
 - Region of the phi-psi map
 - Normalized B-factor (temperature factor)

Step 3 Part 3: Contacts

1. Contacts with heteroatoms
2. Interchain contacts
 - Using PQS is a good idea here
3. Contacts with functional sites
 - BINDING, ACT_SITE, LIPID, and METAL

Output: Prediction

- Result prediction is one of the following:
 - Probably damaging
 - The mutation affects structure or function with high confidence
 - Possibly damaging
 - “it is supposed to affect protein function or structure”
 - Benign
 - Probably doesn’t have any phenotypic effect.
 - Unknown
 - Not enough data to make a prediction.

How is this determination made:

- A table of rules broken down into three categories: PSIC score difference, substitution site properties, and substitution type properties.
 - Each rule may consist of one or more restrictions on PSIC score difference, substitution site properties, or substitution type properties.
- If the data provided after running is rejected by every rule in the table, the mutation is denoted benign.
- See handout.

Performance:

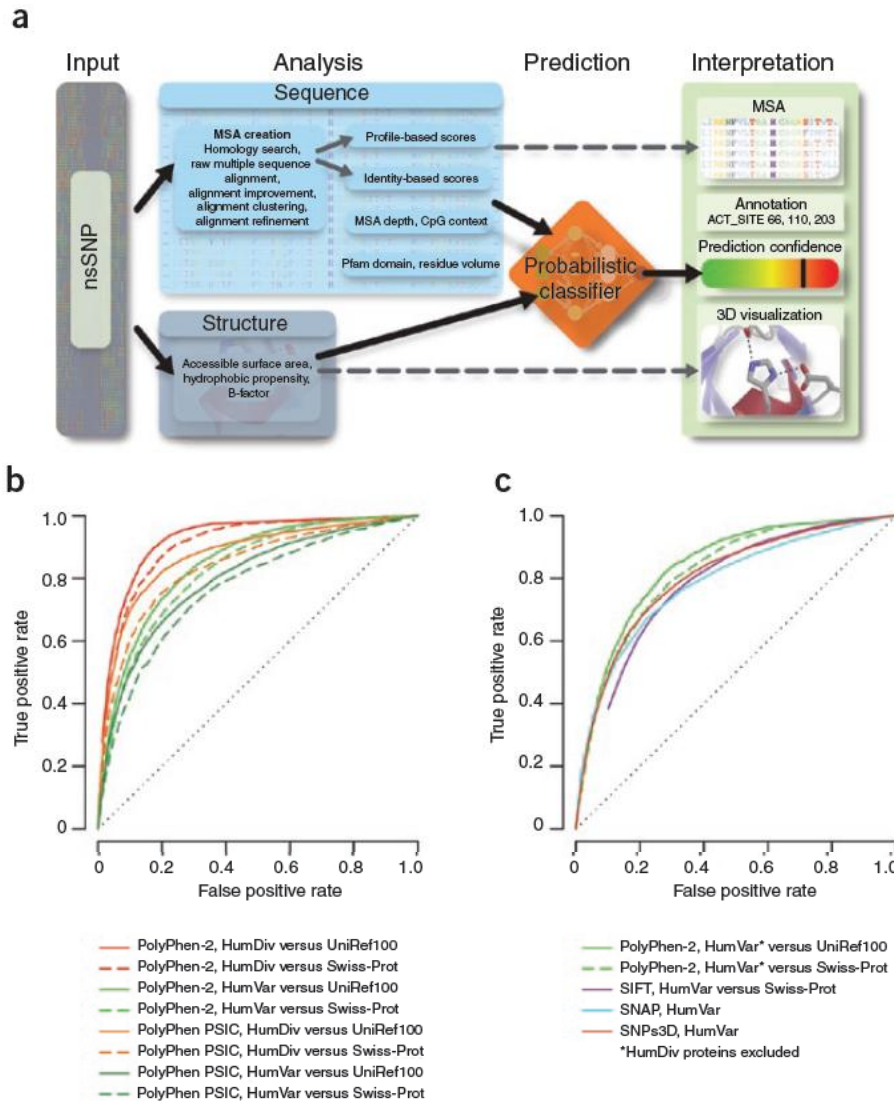


Figure 1 | PolyPhen-2 pipeline and prediction accuracy. (a) Overview of the algorithm. MSA, multiple sequence alignment. (b) Receiver operating characteristic (ROC) curves for predictions made by PolyPhen-2 using fivefold cross-validation on HumDiv and HumVar³ data, using UniRef100 and Swiss-Prot databases for the homology search. Also shown are ROC curves for PolyPhen on HumDiv and HumVar calculated from the difference between position-specific independent counts (PSIC) scores¹ of the wild-type and the mutant amino acids. (c) ROC curves for PolyPhen-2 trained on HumDiv and tested on a subset of HumVar data nonoverlapping with HumDiv data. UniRef100 and Swiss-Prot databases were used for the homology search. Also shown are ROC curves obtained using the programs sorting intolerant from tolerant (SIFT)⁴, screening for nonacceptable polymorphisms (SNAP)⁵ and SNPs3D⁶ on HumVar data. Methods other than PolyPhen-2 and PolyPhen could not easily be applied to HumDiv data because using the same sequences for obtaining both multiple alignments and nondamaging 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 sequences of splice forms, sequences of human allelic variants and incomplete sequences, making it possible to guarantee that allelic variants used in testing datasets would not appear in multiple-sequence alignments.

The end.

Questions?