Predicting functional effect of human missense mutations

van den Berg, Bastiaan; Thornton, JM; Reinders, Marcel; de Ridder, Dick; Beer, TAP

Publication date
2013

Document Version
Final published version

Citation (APA)
van den Berg, B., Thornton, JM., Reinders, M., de Ridder, D., & Beer, TAP. (2013). *Predicting functional effect of human missense mutations.* 1.

Predicting functional effect of human missense mutations

B.A. van den Berg[*1,3,4], J.M. Thornton[2], M.J.T. Reinders[1,3,4], D. de Ridder[1,3,4], and T.A.P. de Beer[2]

[1] Delft Bioinformatics Lab, Faculty of Electrical Engineering, Mathematics & Computer Science, Delft University of Technology, Delft, The Netherlands,
[2] European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK,
[3] Netherlands Bioinformatics Centre, Nijmegen, The Netherlands,
[4] Kluyver Centre for Genomics of Industrial Fermentation, Delft, The Netherlands

* b.a.vandenberg@tudelft.nl

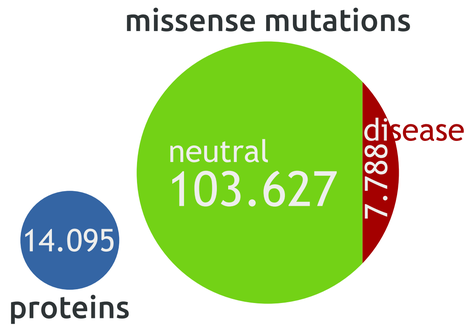
Introduction

Our aim is to prioritize human missense mutations by their probability of being disease causing. Such a computational method could be used to obtain a reduced set of mutations with a relatively large fraction of disease related mutations, thereby aiding in the search for this type of mutation within a large mutation set.

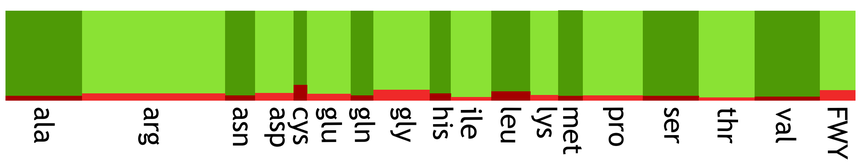
Whereas a range of methods is available for this purpose, only few employ the availability of the 1000G data to obtain a set of neutral mutations. The novelty of our approach is the use of separate classifiers that were trained on a subset of mutations from one amino acid to any other amino acid. The combined performance of these classifiers show an improved performance compared to the often used prediction method PolyPhen2.

Data set

The data set is composed of in total 111.415 mutations in 14.095 proteins. The disease mutations were obtained from the OMIM database and the neutral mutations were obtained from the 1000 Genomes project.



Mutations were split into 20 (non-overlapping) subsets, with in each subset mutations from one amino acid to any other amino acid. The phenylalanine, tryptophan, and tyrosine subsets are combined into one set to increase the set size, resulting in a total of 18 subsets.



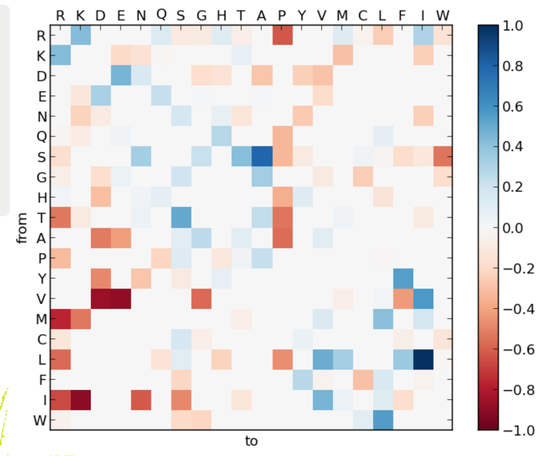
Classification

Separate classifiers were trained on each of the eighteen mutation subsets using the settings below. For comparison, one classifier was trained on the entire dataset.

protocol: 10-fold cross-validation
classifier: linear discriminant analysis (LDA) classifier
measure: area under the receiver operator curve (AUROC)

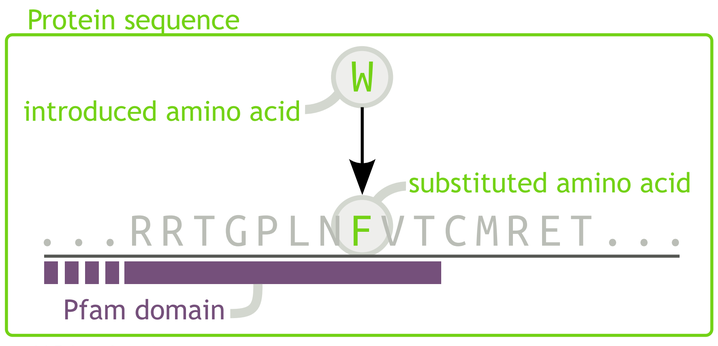
Amino acid counts

Comparison of the occurrences in the neutral and disease set shows which mutations are relatively safe (blue) and dangerous (red).

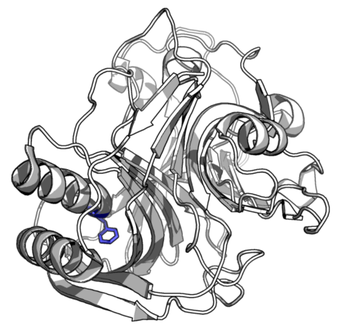


Feature data sources

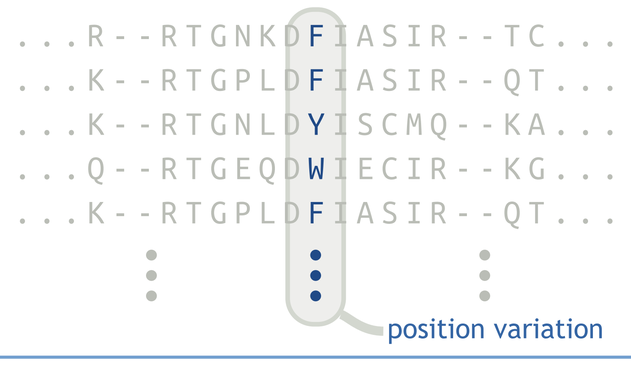
The data sources shown in the figures were employed for feature extraction. Because the availability of structure data is limited, structure-based features were only acquired for part of the mutations.



Protein structure



MSA with homologous sequences



Features

Missense mutation feature vector

Twenty features encode the missense mutation, each column representing one amino acid. The substituted amino acid is set to -1 and the introduced amino acid to 1. All other amino acids are 0.

The first feature is a conservation score based on the MSA with homologous sequences as obtained by the Evolutionary Trace Server. The second feature is a binary feature that indicates if the introduced amino acid is in the position variation or not.

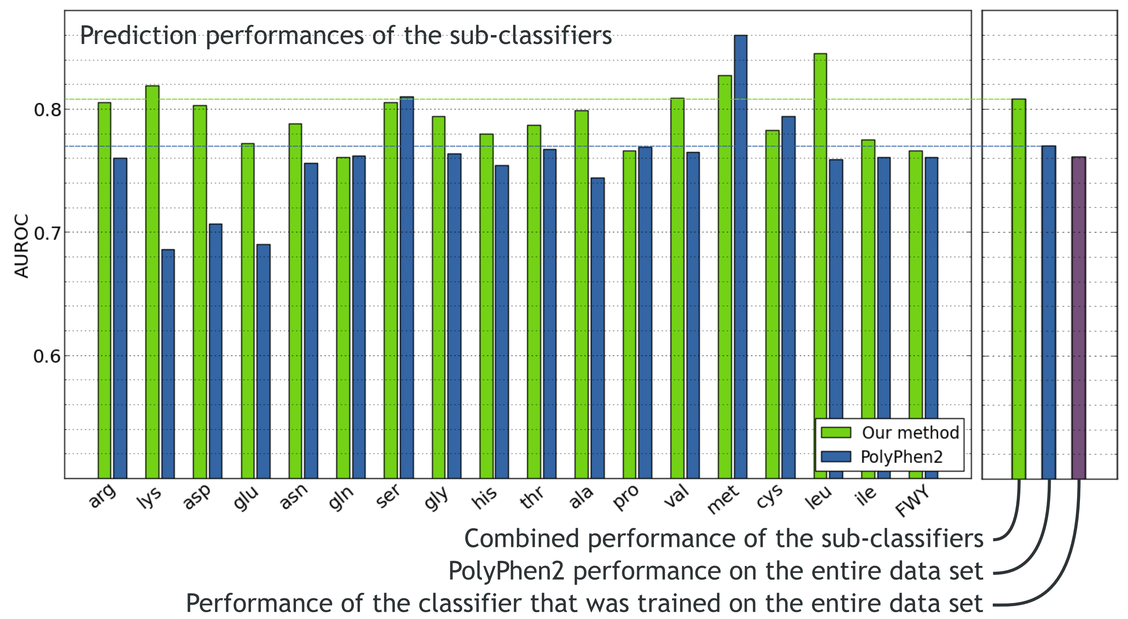
Nineteen features that give the minimal 'characteristic' distance between the introduced amino acid and the amino acids in the position variation. The used characteristics are, for example, hydrophobicity, size, and isoelectric point.

Protein structure features: solvent exposed area and the three backbone angles.

Binary feature that indicates if the mutation falls within a Pfam domain or not.

Results: classification performance

Most of the sub-classifiers as well as their combined result (green) show an improved performance compared to PolyPhen2 (blue). In particular, a striking improvement is observed for charged (arg, lys, asp, glu) and aliphatic (leu, val) sub-classifiers. The reduced performance of the classifier trained on the entire data set (purple) supports the use of sub-classifiers.



nbic netherlands bioinformatics centre

TU Delft

EMBL-EBI

Kluyver CENTRE | Kluyver Centre for Genomics of Industrial Fermentation