Predicting functional effect of human missense mutations

B.A. van den Berg1,3,4, J.M. Thornton2, M.J.T. Reinders1,3,4, D. de Ridder1,3,4, and T.A.P. de Beer2

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

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 subset mutations from one amino acid to any other amino acid. The subset mutations from one amino acid to any other amino acid were only acquired for part of the mutations.

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 data set.

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).

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.