Discriminating between disease-causing and neutral non-frameshifting micro-INDELs by support vector machines by means of integrated sequence- and structure-based features

Huiying Zhao, Yuedong Yang, Hai Lin, Xinjun Zhang, Matthew Mort, David N. Cooper, Yunlong Liu, and Yaoqi Zhou

1School of Informatics, Indiana University Purdue University, Indianapolis, Indiana 46202, USA
2Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, Indiana 46202, USA
3Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, Indiana 46202, USA
4Institute of Medical Genetics, Cardiff University, Heath Park, Cardiff CF14 4XN, UK

ABSTRACT

Micro-INDELs (insertions or deletions of ≤20 bp) constitute the second most frequent class of human gene mutation after single nucleotide variants. A significant portion of exonic INDELs are non-frameshifting (NFS), serving to insert or delete a discrete number of amino-acid residues. Despite the relative abundance of NFS-INDELs, their damaging effect on protein structure and function has gone largely unstudied whilst bioinformatics tools for discriminating between disease-causing and neutral NFS-INDELs remain to be developed. We have developed such a technique (DDIG-in; Detecting Disease-causing Genetic variations due to INDELs) by comparing the properties of disease-causing NFS-INDELs from the Human Gene Mutation Database (HGMD) with putatively neutral NFS-INDELs from the 1,000 Genomes Project. Having considered 58 different sequence- and structure-based features, we found that predicted disordered regions around the NFS-INDEL region had the highest discriminative capability (disease versus neutral) with an Area Under the receiver-operating characteristic Curve (AUC) of 0.82 and a Matthews Correlation Coefficient (MCC) of 0.56. All features studied were combined by support vector machines (SVM) and selected by a greedy algorithm. The resulting SVM models were trained and tested by ten-fold cross-validation on the microdeletion dataset and independently tested on the microinsertion dataset and vice versa. The final SVM model for determining NFS-INDEL disease-causing probability was built on non-redundant datasets with a protein sequence identity cutoff of 35% and yielded an MCC value of 0.68, an accuracy of 84% and an AUC of 0.89. Predicted disease-causing probabilities exhibited a strong negative correlation with the average minor allele frequency (correlation coefficient, -0.84). DDIG-in, available at http://sparks.informatics.iupui.edu, can be used to estimate the disease-causing probability for a given NFS-INDEL.

Mentor: Yaoqi Zhou and Yunlong Liu.