PREDICTING DISEASE-RELATED SINGLE AMINO ACID POLYMORPHISMS USING PROTEIN STRUCTURE

E. Capriotti*

Department of Mathematics and Computer Sciences, University of Balearic Islands,
Palma de Mallorca, Spain
Department of Bioengineering, Stanford University
Stanford, California 94305, United States of America
**Email: emidio@stanford.edu*

R. B. Altman

Departments of Bioengineering and Genetics, Stanford University
Stanford, California 94305, United States of America
Email: russ.altman@stanford.edu

Large-scale sequencing and genotyping techniques are allowing to scan the whole human genome providing a huge amount of genetic variation data. Single Nucleotide Polymorphisms (SNPs), which are the main cause of human genome variability, and can also be responsible for the insurgence of human pathologies. In this study, we present a new structure-based machine-learning approach to predict the effect of SNPs.

1. INTRODUCTION

Large-scale sequencing and genotyping techniques are allowing to scan the whole human genome providing a huge amount of genetic variation data. Single Nucleotide Polymorphisms (SNPs), which are the main cause of human genome variability, and can also be responsible for the insurgence of human pathologies. The non-synonymous SNPs occurring in coding regions resulting in single amino acid polymorphisms (SAPs) may affect protein function and lead to a diseased state. In the last years, several methods have been developed to predict disease-related SAPs using protein sequence[1-3] and protein structure[4,5] information. Although sequence-based approaches are reaching a high level of accuracy, they can be further improved introducing new features derived from the protein three-dimensional structure.

2. RESULTS

In this study, we developed a structure-based machine-learning approach to predict if a SAP is disease-related or not. The implemented Support Vector Machine (SVM) method has been trained on a set of 3,342 disease-related mutations and 1,644 neutral polymorphisms from 784 protein chains. The SVM input data are: i) the amino acid substitution, ii) the structure environment, iii) the sequence profile information, and iv) a Gene Ontology (GO) based score. After dataset balancing, the structure-based method, tested by 20-fold cross-validation procedure, results in 85% overall accuracy, a correlation coefficient of 0.70, and an area under the receiving operating characteristic curve (AUC) of 0.92. When compared with a previously developed sequence-based algorithm, the structure-based method results in 3% higher accuracy and AUC, and 0.07 higher correlation coefficient.

Acknowledgments

EC acknowledges support from the Marie Curie International Outgoing Fellowship program (PIOF-GA-2009-237225). RBA would like to acknowledge the following funding sources: NIH LM05652 and GM61374.

References

1. Capriotti E, Calabrese R, Casadio R. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector

[*] Corresponding author.

machines and evolutionary information. *Bioinformatics* 2006; **22**: 2729-2734.

2. Capriotti E, Arbiza L, Casadio R, Dopazo J, Dopazo H, Marti-Renom MA. The use of estimated evolutionary strength at the codon level improves the prediction of disease related protein mutations in human. *Human Mutation* 2008; **29**: 198-204.
3. Calabrese R, Capriotti E, Fariselli P, Martelli PL, Casadio R. Functional annotations improve the predictive score of human disease-related mutations in proteins. *Human Mutation* 2009; **30**; 1237-1244.
4. Yue P, Melamud E, Moult J. SNPs3D: candidate gene and SNP selection for association studies. *BMC Bioinformatics* 2006. **7**: 166.
5. Ye ZQ, Zhao SQ, Gao G, Liu XQ, Langlois RE, Lu H, Wei L. Finding new structural and sequence attributes to predict possible disease association of single amino acid polymorphism (SAP). *Bioinformatics* 2007; **23**: 1444-1450.