54. Prediction of the insurgence of human genetic diseases due to single point protein mutation.

Calabrese Remo, Capriotti Emidio, Fariselli Piero and Casadio Rita

Laboratory of Biocomputing, CIRB/Department of Biology, University of Bologna, Italy.

Single nucleotide polymorphisms (SNPs) are the most common type of genetic variations in humans accounting for about 90% of sequence differences [1]. It is estimated that SNPs occur approximately every 1000 bases in the overall human population. The importance of SNPs in genetic studies is due to different reasons. First, since most of SNPs are inherited from one generation to the next, they are useful to study the human evolution. SNPs can also be responsible for genetic diseases. Finally, SNPs are at the basis of the relationship between a given phenotype and one or more regions of the genome. Due to several experimental efforts, the consistence of the dbSNP database is increasing exponentially (http://www.ncbi.nlm.nih.gov/SNP) [2] and presently amounts to about 5 million of validated cases (dbSNP 124: Jan 6 2005).

Recently, several databases, servers and tools have been developed in order to study the effects of SNPs in the human genome [3, 4,5]. One of the most important challenges is the understanding of which variants cause diseases. Generally speaking the mutations that occur in coding regions have a larger effect on the gene functionality.
In this work we analyzed a particular class of SNPs that cause changes in the corresponding protein sequence. These kinds of SNPS are called non-synonymous coding SNPs (nsSNPs). We developed a method based on support vector machines that starting from the sequence information can predict if a new phenotype derived from a nsSNP is related to a genetic disease. Our predictor discriminates whether a given single point protein mutation induces neutral polymorphism or is disease-related. Using a dataset of 14264 protein mutations (from 1288 human protein sequences) derived from Swiss-Prot Database (release 45, October 2004), we show that the accuracy of our predictor is as high as 79% for the specific task of predicting if a single point mutation can be related to a genetic disease.

REFERENCES

[1] Collins FS, Brooks LD, Charkraverti A (1998). A DNA polymorphism discovery resource for research on human genetic variation. Genome Res 8, 1229-1231.
[2] Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K (2001). dbSNP: the NCBI database of genetic variation. Nucleic Acids Research 29(1), 308-311.
[3] Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NS, Abeysinghe S, Krawczak M, Cooper DN. Human Gene Mutation Database (HGMD): 2003 update. Human Mutatation 2003 Jun;21 (6):577-81.
[4] Wang Z, Moult J (2001). SNPs, protein structure, and disease. Human Mutation, 17:263-270.
[5] Ramensky V, Bork P, Sunyaev S (2002). Human non-synonymous SNPs: server and survey. Nucleic Acids Research, 30: 3894-3900.