Communications to SIMAI Congress,
ISSN 1827-9015, Vol. 1 (2006)

DOI: 10.1685/CSC06032

Predicting the insurgence of human genetic diseases due to single point protein mutations with a machine learning approach

Calabrese Remo, Capriotti Emidio and Casadio Rita
Laboratory of Biocomputing, CIRB/Department of Biology, University of Bologna
Via Irnerio 42, 40126 Bologna, Italy
emidio@biocomp.unibo.it

The most common genetic variations in humans are single nucleotide polymorphism (SNPs). The importance of SNPs in genetic studies is due to their association with genetic diseases. In this work we analyzed a particular class of SNPs that cause changes in the corresponding protein sequence. We developed a new method that starting from the sequence information can predict if a new phenotype derived from a SNP is related to a genetic disease. Using a dataset of 21185 protein mutations (3587 human proteins) derived from Swiss-Prot Database, our predictor reaches 70% accuracy for the specific task of predicting if a mutation can be related or not to a genetic disease.

Introduction

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 human evolution. Studying the SNPs statistics in different human populations can lead to important considerations about the history of our species. SNPs can also be responsible of genetic disease. New experimental techniques for a large scale identification of SNPs in the human population have increased exponentially the consistence of the dbSNP database (http://www.ncbi.nlm.nih.gov/SNP) [2] that contains about 5 million of validated cases (dbSNP 125: Sep 29 2005). Recently, several databases, servers and tools have been developed in order to study the effects of SNPs in Homo Sapiens [3,4,5]. One of the most important challenge 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 paper we analyzed a particular class of SNPs that cause changes in the aminoacid sequence. These kind of SNPS are called non-synonymous coding SNPs (nsSNPs).We developed a method based on support vector machines that starting from sequence information can predict if a new phenotype derived from a nsSNP is related to a genetic disease in humans. We present a support vector machine based method, to predict if a given single point protein mutation induces neutral polymorphism or a disease-related mutation.

Methods

The data set here used is derived from the release 48 (Dec 2005) of the Swiss-Prot database [6]. It was retrieved from Swiss-Prot, considering only *Homo sapiens* proteins that have single point mutations related to diseases or neutral polymorphisms.

After this procedure, we ended up with a data set consisting of 21185 different single point mutations (12944 of which are disease related and 8241 are described as neutral polymorphisms), obtained from 3587 protein sequences.

The task here addressed is to predict whether a given single point protein mutation, due to a nsSNP, is a neutral polymorphism or is involved into the insurgence of a human genetic disease. The method is a SVM that classifies mutations into diseases related (desired output set to 0) and neutral polymorphism (desired output set to 1). The decision threshold is set equal to 0.5. Similarly to previous algorithms implemented to predict the protein stability changes upon single point mutation [7] the input vector consists of 40 values: the first 20 (the 20 residue types) explicitly define the mutation by setting to -1 the element corresponding to the wild type residue and to 1 the newly introduced residue (all the remaining elements are kept equal to 0). The last 20 input values encode for the mutation sequence environment (again the 20 elements represent the 20 residue types). Each input is provided with the number of the encoded residue type, to be found inside a window centered at the residue that undergoes the mutation and that symmetrically spans the sequence to the left (N-terminus) and to the right (C-terminus). For SVM implementation we use LIBSVM with a RBF kernel function $K(x_i,x_j)=\exp(-G \|x_i - x_j \|^2)$.

Results

In order to find the most informative sequence environment, different values of window length have been tested. The best accuracy is reached for a window length of 19 residues. A small variation of this value do not change significantly the accuracy of the predictor (data not show).The performance of our SVM method is evaluated using a cross-validation procedure on the whole data set. The reported data for the classification task performed by the SVM are obtained adopting a 20-fold cross-validation procedure in such a way that the disease-related and neutral polymorphism mutation examples respected the original distribution of the whole set. Furthermore, all the proteins in our set are been clustered according to their sequence similarity using the *blastclust* program in the BLAST suite [8], by adopting the default value of length coverage equal to 0.9 and the score coverage threshold equal to 1.75. We kept the mutations detected on the same cluster of protein sequences in the same training set to prevent an overestimation of the results**.**

The score of our predictor on 21185 mutations are reported in table I.

Table I: Performance of the SVM methods

	Q2	P[dis]	Q[dis]	P[pol]	Q[pol]	C
SVM-WIN19	0.70	0.71	0.84	0.65	0.46	0.34

dis and pol : the indexes are evaluated for single point protein mutation related to human disease and neutral polymorphism, respectively; for the definition of the different indexes see reference 6.

In order to test the reliability of our method we plot the total accuracy (Q2) and the correlation coefficient (C) as a function of the Reliability Index defined in [7] (see Fig 1).

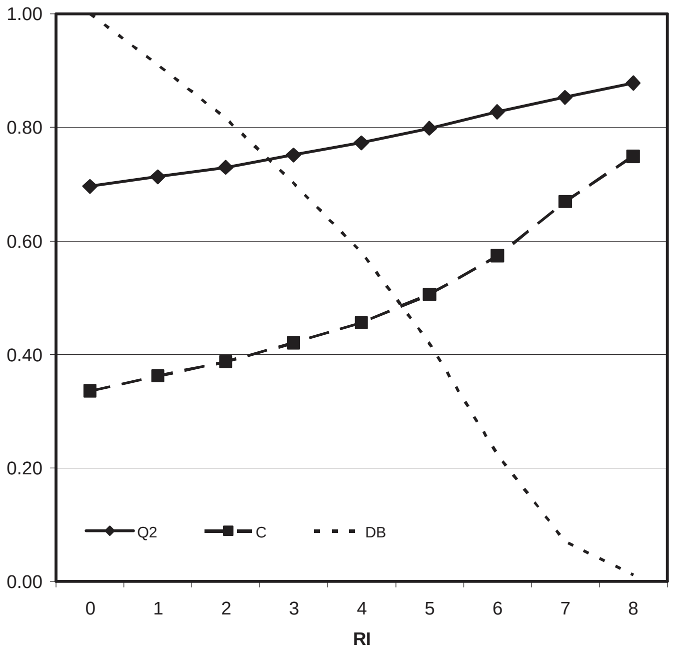

Fig. 1. Accuracy (Q2) and correlation (C) of the SVM method as a function of the reliability index (RI) of the prediction (see [7]). DB is the fraction of the data set with RI values higher or equal to a given threshold.

Another way to show the accuracy of our predictor is represented in figure 2 where we report the ROC curve and its area calculated according to reference [9].

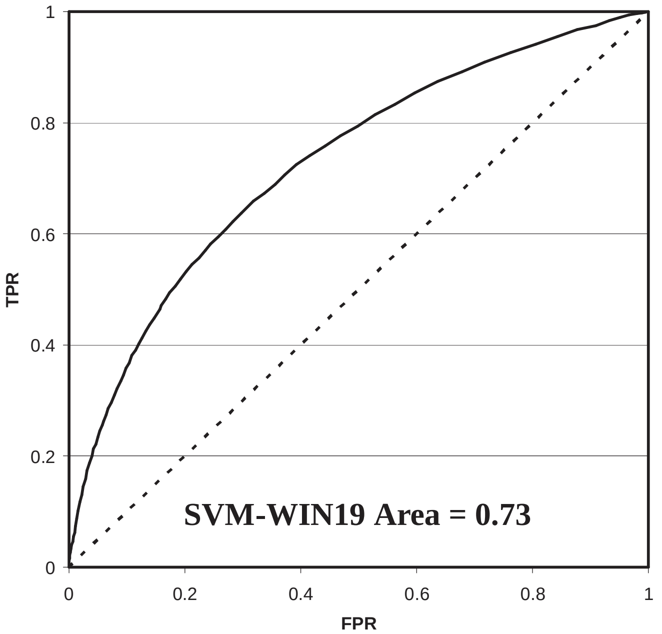

Fig. 2. ROC curve SVM-WIN19 obtained plotting the False Positive Rate vs the True Positive Rate. For the indexes details see reference [9].

We can conclude that our SVM method reaches an overall accuracy (Q2) of 70% and a correlation coefficient (C) of 0.34. When only results with a level of RI ≥5 are selected the accuracy increases to 80% and a correlation coefficient to 0.51 over the 42% of the database. In figure 2 it is show that our predictor reaches a good ROC curve area of 0.73 with respect to the random predictors that have a ROC curve area of 0.5.

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.

6. Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S. and Schneider, M. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003 Nucleic Acids Res, 31, 365-370.

7. Capriotti, E., Fariselli, P., Calabrese, R. and Casadio, R. (2005) Predicting protein stability changes from sequences using support vector machines Bioinformatics, 21 Suppl 2, ii54-ii58.

8. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs Nucleic Acids Res, 25, 3389-3402.

9. Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A. and Nielsen, H. (2000) Assessing the accuracy of prediction algorithms for classification: an overview Bioinformatics, 16, 412-424.