BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 22 2006, pages 2729–2734
doi:10.1093/bioinformatics/btl423

Sequence analysis

Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information

E. Capriotti[1], R. Calabrese[1] and R. Casadio[1,*]

[1]Laboratory of Biocomputing, CIRB/Department of Biology, University of Bologna, via Irnerio 42, 40126 Bologna, Italy

Received on June 21, 2006; revised and accepted on July 28, 2006
Advance Access publication August 7, 2006
Associate Editor: Alex Bateman

ABSTRACT

Motivation: Human single nucleotide polymorphisms (SNPs) are the most frequent type of genetic variation in human population. One of the most important goals of SNP projects is to understand which human genotype variations are related to Mendelian and complex diseases. Great interest is focused on non-synonymous coding SNPs (nsSNPs) that are responsible of protein single point mutation. nsSNPs can be neutral or disease associated. It is known that the mutation of only one residue in a protein sequence can be related to a number of pathological conditions of dramatic social impact such as Altzheimer's, Parkinson's and Creutzfeldt-Jakob's diseases. The quality and completeness of presently available SNPs databases allows the application of machine learning techniques to predict the insurgence of human diseases due to single point protein mutation starting from the protein sequence.

Results: In this paper, we develop a method based on support vector machines (SVMs) that starting from the protein sequence information can predict whether a new phenotype derived from a nsSNP can be related to a genetic disease in humans. Using a dataset of 21 185 single point mutations, 61% of which are disease-related, out of 3587 proteins, we show that our predictor can reach more than 74% accuracy in the specific task of predicting whether a single point mutation can be disease related or not. Our method, although based on less information, outperforms other web-available predictors implementing different approaches.

Availability: A beta version of the web tool is available at http://gpcr.biocomp.unibo.it/cgi/predictors/PhD-SNP/PhD-SNP.cgi

Contact: casadio@alma.unibo.it

1 INTRODUCTION

Single nucleotide polymorphisms (SNPs) are the most common type of genetic variations in humans, accounting for ~90% of sequence differences (Collins *et al*., 1998). 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 characterize human evolution (Goldstein and Cavalleri, 2005). Studying SNPs distribution in different human populations can indeed lead to important considerations about the history of our species (Barbujani and Goldstein, 2004; Edmonds *et al*., 2004). Finally SNPs can also be responsible of genetic diseases (Ng and Henikoff, 2002; Bell, 2004). New experimental techniques for large-scale identification of SNPs in the human population (Wang *et al*., 1998) have increased exponentially the consistence of the dbSNP database (http://www.ncbi.nlm.nih.gov/SNP) (Sherry *et al*., 2001) that presently contains ~6 millions of validated cases (dbSNP 126). Recently several databases, servers and tools have been developed in order to study the effects of SNPs in *Homo sapiens* (Wang and Moult, 2001; Ramensky *et al*., 2002; Riva and Kohane, 2002; Ng and Henikoff, 2003; Stenson *et al*., 2003; Conde *et al*., 2004; Reumers *et al*., 2005; Karchin *et al*., 2005; Yue and Moult, 2006).

An important goal is the understanding of which variants are disease-related. As a rule of thumb, mutations occurring in coding regions may have a larger effect on the gene functionality. In this paper we analyzed a particular class of SNPs that cause changes in the deduced aminoacid sequence. These kinds of SNPs are called non-synonymous coding SNPs (nsSNPs). The paper describes a method to predict whether a given single point protein mutation is related to a human disease or not.

2 MATERIALS AND METHODS

2.1 The mutation datasets

Our dataset is derived from the release 48 (Dec 2005) of the Swiss-Prot database (Boeckmann *et al*., 2003). The classification of neutral and deleterious polymorphisms was taken from Swiss-Prot. For each variant Swiss-Prot lists, in a dedicated and OMIM-linked web page, the effect of nsSNPs; in particular for the deleterious ones, pathological effects are also described. We considered three datasets: the first for training/testing our SVM system based on sequence information (HumVar), the second for training/testing our SVM system based on profile information (HumVarProf) and the third, to be used when testing the robustness of our predictor (NewHumVar).

The protein dataset was retrieved from Swiss-Prot, with the following constraints:

(1) the protein source is *H.sapiens*;
(2) the mutations are related to diseases or neutral polymorphisms (no unclassified cases are considered);
(3) the data are relative to single point protein mutations (no deletion and insertion mutations are taken into account).

After this filtering procedure, we ended up with a dataset consisting of 21 185 different single point mutations (12 944 of which are disease-related and 8241 are described as neutral polymorphisms), obtained from 3587 protein sequences. These proteins have been grouped in clusters using

*To whom correspondence should be addressed.

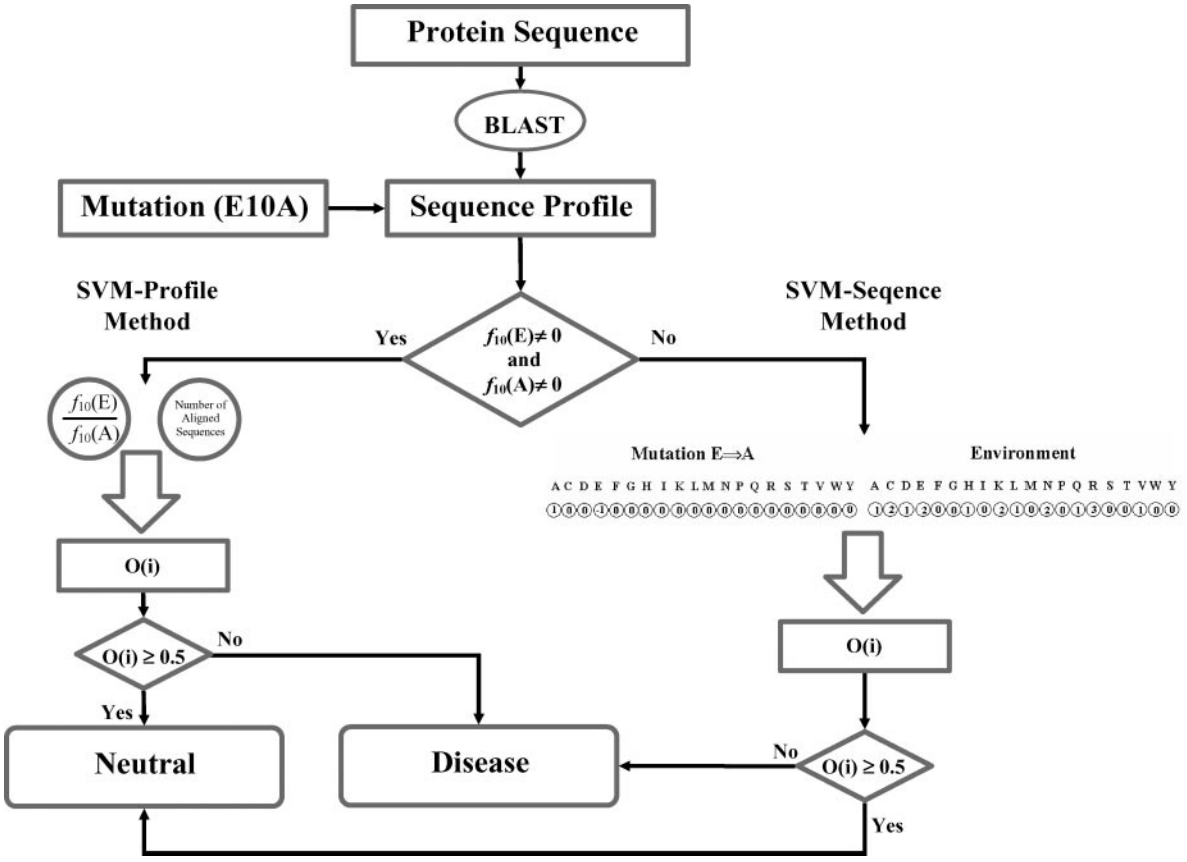

Fig. 1. Flow chart of the hybrid method (HybridMeth). For a protein sequence the method predicts whether a given mutation, (e.g. the residue E in position 10 is mutated to A) is predicted to be related to a human disease or not. In the first step the sequence profile of the protein is built using the BLAST algorithm. In the second step, the value of the sequence profile in the mutated position is evaluated. If both $f_{10}(E) \neq 0$ and $f_{10}(A) \neq 0$, prediction is computed as depicted on the left side of the flow chart, with the SVM-Profile method, taking as input the ratio $f_{10}(E)/f_{10}(A)$ and the number of the aligned sequences in the given position. Otherwise, if $f_{10}(E) = 0$ and/or $f_{10}(A) = 0$ SVM-Sequence discriminates between neutral polymorphism or disease. A representation of the two input vectors of SVM-Sequence are also shown (see Material and methods section for details).

the *blastclust* program in the BLAST suite (Altschul *et al*., 1997). Each sequence was then aligned towards the nr95 data base (release June 2005). Considering each sequence profile, we selected those mutations whose frequency at a given position was different from 0 both for the wild-type and mutated residue. This subset is called HumVarProf and comprises 8718 mutations (3852 of which disease-related and 4866 neutral polymorphisms).

The third set of nsSNPs comprises single point protein mutations from sequences of *H.sapiens*, which are reported in the last release of Swiss-Prot (release 50 Jun 2006). This set was extracted from the Swiss-Prot database by considering only new protein sequences that do not belong to any group of the previous HumVar set. This procedure is carried out using *blastclust* (length coverage equal to 0.9 and the score coverage threshold equal to 30) and provided a set of 935 single point protein mutations (149 of which are disease-related and 786 are described as neutral polymorphisms) from a total of 469 different proteins. The final HumVar and NewHumVar sets are available at http://gpcr.biocomp.unibo.it/~emidio/PhD-SNP/.

2.2 The predictors

Our task 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. In this respect the task can be cast as a classification problem for the protein upon mutation. To address this issue we have implemented different methods: a baseline predictor (ProbMeth) as a benchmark to overperform, a single sequence SVM method (SVM-Sequence) that discriminates disease-related mutations based on the local sequence environment of the mutation at hand and a sequence-profile based SVM (SVM-Profile). SVM-Sequence and SVM-Profile are cast in a unique workflow with a decision tree method (HybridMeth) that allows adopting either SVM-Sequence or SVM-Profile depending on the presence or absence of a sequence profile of the sequence at hand, respectively (Fig. 1).

2.2.1 The probability-based method (ProbMeth) The baseline predictor is built by considering the occurrence of a mutation of a pair of residues (wild-type/mutated) in our dataset. For disease (D) and neutral polymorphism (N) cases, we derive the likelihood ratio by computing the $M(D)_{i,j}$ and $M(N)_{i,j}$ matrices, respectively. Each matrix has 20×20 elements; the generic element $M_{i,j}$ scores the occurrence of the mutation of residue i into residue j and is computed as:

$$M_{i,j} = f\ (i,\ j)/[f(i)f(j)], \tag{1}$$

where $f(i,j)$ is the frequency of occurrence of mutation of residue i into residue j; $f(i)$ is the frequency of occurrence of residue i in the data base and $f(j)$ is the frequency of occurrence of any mutation corresponding to residue j in the database. By computing the maximum values of $M(N)_{i,j}$ and $M(D)_{i,j}$ and comparing them, a given mutation of residue i into residue j is predicted to be disease-related or not. When $M(D)_{i,j} > M(N)_{i,j}$ the single

point protein mutation is labeled as disease-related; when $M(D)_{i,j} \leq M(N)_{i,j}$ the mutation is predicted as neutral polymorphism.

2.2.2 The SVM-based method using sequence information (SVM-Sequence) The first SVM classifies mutations into disease-related (desired output set to 0) and neutral polymorphism (desired output set to 1). The decision threshold is set equal to 0.5. 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) with a length of 19 residues (Capriotti *et al*., 2005a). For SVM implementation we use LIBSVM (http://www.csie.ntu.edu.tw/~cjlin/) with a RBF kernel function $K(x_i,x_j) = \exp(-G \|x_i - x_j\|^2)$.

2.2.3 The SVM-based method using profile information (SVM-Profile) The second SVM method (SVM-Profile) classifies mutations into disease and neutral polymorphism taking as input only a vector of 2 elements derived from the sequence profile. This is computed from the output of the BLAST program (Altschul *et al*., 1997), running on the *nr*95 database (*E*-value threshold $= 10^{-9}$, number of runs $= 1$) as obtained with *cd-hit* program available at http://bioinformatics.org/cd-hit/ (Li *et al*., 2001). The first input element is the ratio between the frequencies of wild-type versus that of the mutated residue in the sequence mutated position and the second element is the number of aligned sequences with respect to the mutation at hand. The software and the kernel used for this SVM implementation are as described above.

2.2.4 The hybrid method (HybridMeth) Our hybrid method (HybridMeth) is based on a decision tree with the SVM-based classifier described above (SVM-Sequence) coupled to SVM-Profile trained on sequence profile information (Fig. 1).

HybridMeth comprises the following steps:

(1) for a given protein, its sequence profile is built according to the procedure detailed above. From this we evaluate both the frequency of the wild-type [f_k(wt)] and mutated [f_k(mut)] residues at position k. The normalization factor is the number of sequences in the alignment at a given position;

(2) when the frequency of the wild-type [f_k(wt)] and mutated [f_k(mut)] residues at position k are different from 0, the value of $f_k(wt)/f_k(mut)$ is computed and in conjunction with the total number of aligned sequences in position k is provided to the SVM-Profile method trained on the sequence profile HumVarProf set;

(3) when no profile is returned at a given position for either wild-type or mutated residue, $f_k(wt) = 0$ or $f_k(mut) = 0$. The prediction is performed with the SVM-Sequence method, as described above.

2.3 Scoring the performance

All the results obtained with our SVM methods are evaluated using a cross-validation procedure on the HumVar dataset for SVM-Sequence and on HumVarProf for SVM-Profile. The reported data for the classification task performed by the SVM methods are obtained adopting a 20-fold cross-validation procedure in such a way that the disease-related and neutral polymorphism mutation ratio corresponds to the original distribution of the whole set. Furthermore, all the proteins in the HumVar and HumVarProf sets are clustered according to their sequence similarity using the *blastclust* program in the BLAST suite (Altschul *et al*., 1997), 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.

Performance is scored with several measures. For sake of completeness here we review the ones adopted in this paper. The efficiency of the predictor is scored using the statistical indexes defined in the following. The overall accuracy is:

$$Q2 = p/N, \tag{2}$$

where p is the total number of correctly predicted mutations and n is the total number of mutations.

The correlation coefficient C is defined as:

$$C(s) = [p(s)n(s) - u(s)o(s)]/D, \tag{3}$$

where D is the normalization factor

$$D = \{[p(s) + u(s)][p(s) + o(s)][n(s) + u(s)][n(s) + o(s)]\}^{1/2}. \tag{4}$$

For each class *s* (D and N, for disease and neutral polymorphism, respectively); p(s) and n(s) are the total number of correct predictions and correctly rejected assignments, respectively, and u(s) and o(s) are the numbers of under and over predictions.

The coverage for each discriminated class *s* is evaluated as:

$$Q(s) = p(s)/[p(s) + u(s)], \tag{5}$$

where p(s) and u(s) are the same as in Equation (3). The probability of correct predictions P(s) (or accuracy of s) is computed as:

$$P(s) = p(s)/[p(s) + o(s)], \tag{6}$$

where p(s) and o(s) are the same as in (3) (ranging from 1 to 0).

Finally, it is very important to assign a reliability score to each prediction. With the output O(s) this is obtained by computing:

$$RI(s) = 10*\text{abs}(O(s) - t)*w(s), \tag{7}$$

where *t* is threshold and w(s) is the weight of the set relative to the class *s*.

Other standard scoring measures (Baldi *et al*., 2000), including the area under the ROC curve and the true positive rate [TPR $= Q(s)$] at 5% of false positive rate [FPR$= 1 - P(s)$] are also reported.

3 RESULTS AND DISCUSSION

3.1 The effect of the evolutionary information

Information derived from sequence profile is important for detecting mutations that affects human health (Ramensky *et al*., 2002). Prompted by this observation we undertake a statistical analysis of our data set to capture relevant features in order to discriminate among neutral and disease-related nsSNPs. After a careful search we found that the best scoring discriminating function is the ratio between the frequency of wild-type [f_k(wt)] versus mutated [f_k(mut)] residues in the sequence profile, after alignment towards the nr95 data base. This function as shown in Figure 2 is to some extent discriminative between neutral and disease-related mutations.

Unfortunately our data base of mutations presently contains only 41% of residues for which it is possible to compute a sequence profile for both wild-type and mutated residues. The remaining portion either cannot be aligned (15%) or has alignments for either residue (44%). This obviously prevents from computing the alignment based-scoring function for any type of mutations. The distribution of Figure 2 is unaffected when all the human proteins from the dataset of sequences are removed. This suggests that our functions are evaluated from sequence profiles computed mainly considering orthologous proteins (data not shown).

Also from Figure 2 it is evident that given our data base of mutations (for which only 8718 out of 21 815 mutations are

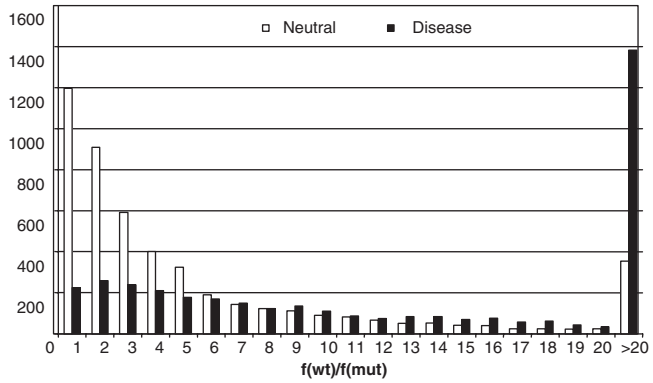

Fig. 2. Distribution of neutral polymorphism and disease-related mutations at different values of the ratio between the frequencies of wild-type and mutated residue in the sequence profile [(f(wt)/f(mut)]. These data are calculated over the HumVarProf set comprising 8718 mutations (3852 of which are disease-related and 4866 are described as neutral polymorphisms).

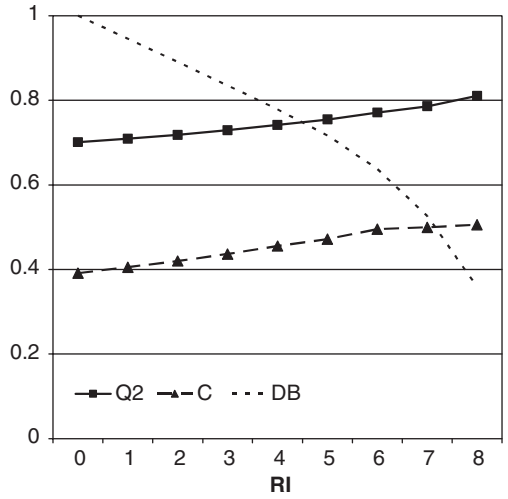

Fig. 3. Accuracy (Q2) and correlation (C) of the SVM-Profile method as a function of the reliability index (RI) of the prediction [Equation (7)]. DB is the fraction of the HumVarProf dataset with RI values higher or equal to a given threshold.

endowed with a sequence profile) both disease-related and neutral mutations are similarly present at a given position with only few exceptions [at very low number of sequences (<50) or very high (950) in the alignment] and that the scoring function is really discriminative at specific values, depending on the number of aligned sequences at a given position.

To overcome these difficulties we implemented a decision tree method that eventually takes advantage of the ratio-based scoring function only when available from sequence profile (Fig. 1). Our method is therefore based on the notion that any mutation starting from the residue sequence of a given protein may by predicted in relation to the possible effect of being or not disease-related. In any case, the mutation can be predicted with a SVM classifier that takes as input a window centered on the specific mutation and a local environment for the mutation at hand (SVM-Sequence). Alternatively, SVM-Profile is activated when the ratio between the frequency of wild-type [f_k(wt)] and mutated [f_k(mut)] residue can be computed. In order to regularize the prediction this arm of the decision tree is trained/tested on the scoring function value as returned after evaluating f_k(wt) and f_k(mut) (both are required to be different from 0) and the number of sequence in the alignment. This procedure is optimized considering sequence profiles with different E-value threshold and more than one run for BLAST program.

The best results that are obtained when the E-value threshold is set to 10^{-9} and for one run of BLAST, reach an overall accuracy of 70% and a Matthew's correlation coefficient of 0.39 (Fig. 3).

3.2 The predictors at work

Our previous work indicated the role of sequence environment in improving classification with SVM based methods (Capriotti *et al*., 2005a,b). SVM-Sequence takes advantage of information as derived from the environment (see Materials and methods) where the mutation occurs. Coupling the predictions provided by SVM-Sequence and SVM-Profile we can predict any mutation endowed or not with sequence profile (HybridMeth).

The prediction of SVM-Sequence and HybridMeth are compared with a simple method based only on probabilistic rules (ProbMeth). In Table 1 the performance of the three different methods is reported on the whole dataset of mutations (HumVar) and adopting a cross validation procedure.

Table 1. Performance of our different methods on the HumVar set

Method	Q2	P(D)	Q(D)	P(N)	Q(N)	C
ProbMeth	0.62	0.63	0.91	0.56	0.18	0.13
SVM-Sequence	0.70	0.71	0.84	0.65	0.46	0.34
HybridMeth	0.74	0.80	0.76	0.65	0.70	0.46

D and N: indexes are evaluated for single point protein mutation related to human disease (D) and neutral polymorphism (N), respectively; for the definition of the different indexes see the Material and methods section.

SVM-Sequence is, as expected, more accurate than ProbMeth. The overall accuracy of prediction (Q2) increases of 8% points when SVM-Sequence is compared with ProbMeth (Table 1). Also and more importantly the correlation coefficient C increases up to 0.21. In conclusion, the first two methods scored in Table 1 show a good accuracy in the prediction of the disease-related mutations [Q(D)] and are less accurate when predicting neutral polymorphisms [Q(N)].

Merging the prediction of SVM-Sequence with that obtained with SVM-Profile (the HybridMeth predictor) results into an overall accuracy of 0.74 and a correlation coefficient of 0.46. The main reason of this improvement is related to the increasing of the accuracy in the prediction of neutral polymorphisms [Q(N)]. The overall Q2 accuracy and the correlation coefficient are computed as a function of the reliability index (RI) in Figure 4.

This identifies a relationship between the reliability value and the HybridMeth predictor accuracy. The value of the reliability index and its relationship with the prediction accuracy may help in selecting which mutations are more dangerous for human health. The accuracy of our predictor is also represented in Figure 5. For both SVM-Sequence and HybridMeth, the value of the TPR is reported

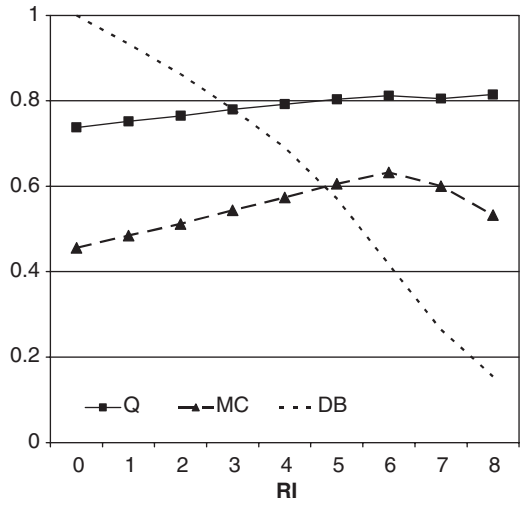

Fig. 4. Accuracy (Q2) and correlation (C) of the hybrid method (HybridMeth) as a function of the reliability index (RI) of the prediction [Equation (7)]. DB is the fraction of the HumVar dataset with RI values higher or equal to a given threshold.

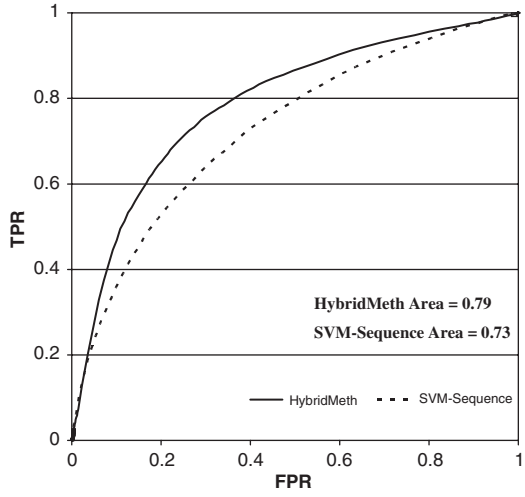

Fig. 5. ROC curve of HybridMeth and SVM-Sequence obtained plotting the False Positive Rate [FPR = 1 − P(s)] versus the True Positive Rate [TPR = Q(s)].

as a function of the FPR (ROC curve). It is evident that the decision tree method by including profile information increases its ROC curve area from 0.73 to 0.79. According to our results when we accept a FPR equal to 5% the decision tree method overperforms SVM-sequence by 10% points (Figure 5).

The analysis of the accuracy of HybridMeth as a function of the chemico-physical properties indicates that the mutations involving charged/charged, apolar/apolar mutations score lower than those involving other swaps (Table 2). This suggests that changes in the class of apolar and charged residues are more difficult to predict and that more information than the local sequence environment and profile is necessary for a high predictive score.

Table 2. Q2 accuracy as a function of the mutated residue type

wt\mut	Polar			Apolar			Charged		
	Q2	C	DB	Q2	MC	DB	Q2	MC	DB
Polar	0.76	0.47	19	0.73	0.44	12	0.81	0.52	14
Apolar	0.76	0.46	12	0.67	0.33	14	0.82	0.48	3
Charged	0.73	0.37	19	0.73	0.39	2	0.66	0.33	5

The accuracy (Q2) and the correlation (C) of a given mutation are reported as a function of the mutated residue type, classified according to chemico-physical properties. Rows account for the wild-type residue (wt), while the column positions define the mutated residues in the mutant proteins (mut). DB is the percentage of a given mutation type in the HumVar dataset.

Table 3. Comparison of our HybridMeth with other web available methods

Method	Q2	P(D)	Q(D)	P(N)	Q(N)	C	PM%
PolyPhen[a]	0.72	0.62	0.72	0.80	0.73	0.44	93
SIFT[b]	0.67	0.76	0.67	0.56	0.66	0.33	94
HybridMeth[c]	0.74	0.80	0.76	0.65	0.70	0.46	100

Taken from web server:
[a]http://www.bork.embl-heidelberg.de/PolyPhen/
[b]downloaded from http://blocks.fhcrc.org/sift/SIFT.html and run locally. Performance is scored on HumVar.
[c]Only HybridMeth is scored with a 20-fold cross-validation procedure on 21 185 mutations. For legend see also Table 1. PM is the percentage of predicted mutations.

3.3 Comparison with other methods

In this section, we compare our method with other web available predictors. The web tools considered are PolyPhen (Ramensky *et al*., 2002) and SIFT (Ng and Henikoff, 2003). The first one is also based on a decision tree and takes into account several information as derived by structural parameters, functional annotations and evolutionary information; the second one is based only on sequence homology considering residues conserved in a given protein family. Differently from HybridMeth these predictors sometimes cannot provide results. This fact occurs when there is no functional or evolutionary information for a given protein. In our case we always give a prediction, thank to the combined action of both methods included in the decision tree.

In Table 3 we report scoring indexes for the methods as compared to the one described in this paper. It should be noticed that only HybridMeth is scored by adopting a real cross validation procedure on our HumVar dataset.

The performances reported indicate that PolyPhen and HybridMeth are more accurate than SIFT. The HybridMeth method reaches the maximum value of accuracy (0.74) among the three methods, gaining 7% with respect to SIFT. PolyPhen is scoring similarly to our method, although only HybridMeth is the only performing predictions for every mutations of dataset, being the simplest in terms of amount of information required starting from the protein sequence. PolyPhen and SIFT do not predict about 1500 mutations since more specific functional or evolutionary information is requested for the mutation at hand.

Table 4. Scoring on the new human mutation set (NewHumVar)

Method	Q2	P(D)	Q(D)	P(N)	Q(N)	C	PM%
PolyPhen[a]	0.72	0.30	0.63	0.92	0.73	0.28	79
SIFT[b]	0.69	0.32	0.55	0.87	0.72	0.22	88
HybridMeth	0.73	0.34	0.74	0.94	0.73	0.36	100

For legend see Tables 1 and 3. Reported data are relative to 935 protein mutations

3.4 The prediction of disease-related mutations on new *H.sapiens* variants

In order to test the robustness of our HybridMeth method and compare it to the other two methods available on the web, we select a set of single protein mutations relative to *H.sapiens*, which have been included in the last version of Swiss-Prot. From this set of proteins, we selected only those sequences that show a low level (<30%) of sequence identity with proteins in HumVar dataset. This procedure is performed using the *blastclust* program (Altschul *et al*., 1997), considering only new sequences that do not fall into any cluster derived by the HumVar dataset. By this we predict whether a given mutation is disease-related or not, using our machine learning approach. When scoring our predictions we also compare with PolyPhen and SIFT (Table 4). Only our method is predicting all the new mutations, with a score that overcomes the best performing one. By this we like to conclude that the decision tree-based new method described in this paper, although far from being perfect, takes advantage of all the possible information only from the protein sequence and by this can discriminate between a disease-related and neutral polymorphism.

ACKNOWLEDGEMENTS

R. Casadio acknowledges the receipt of the following grants: PNR 2001–2003 (FIRB art.8) for a project on Bioinformatics for Genomics and Proteomics, a FIRB 2003 LIBI–International Laboratory of Bioinformatics and the support to the Bologna node of the Biosapiens Network of Excellence project within the European Union's VI Framework Programme. R. Calabrese and E. Capriotti are supported by a FIRB 2003-LIBI grant and a Biosapiens grant, respectively.

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Baldi,P. *et al*. (2000) Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, **16**, 412–424.

Barbujani,G. and Goldstein,D.B. (2004) Africans and Asians abroad: genetic diversity in Europe. *Annu. Rev. Genomics Hum. Genet*., **5**, 119–150.

Bell,J. (2004) Predicting disease using genomics. *Nature*, **429**, 453–456.

Boeckmann,B. *et al*. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res*., **31**, 365–370.

Capriotti,E. *et al*. (2005a) Predicting protein stability changes from sequences using support vector machines. *Bioinformatics*, **21** (Suppl. 2), ii54–ii58.

Capriotti,E. *et al*. (2005b) I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res*., **33**, W306–W310.

Collins,F.S. *et al*. (1998) A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res*., **8**, 1229–1231.

Conde,L. *et al*. (2004) PupaSNP Finder: a web tool for finding SNPs with putative effect at transcriptional level. *Nucleic Acids Res*., **32**, W242–W248.

Edmonds,C.A. *et al*. (2004) Mutations arising in the wave front of an expanding population. *Proc. Natl Acad. Sci. USA*., **101**, 975–979.

Goldstein,D.B. and Cavalleri,G.L. (2005) Genomics: understanding human diversity. *Nature*, **437**, 1241–1242.

Karchin,R. *et al*. (2005) LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources. *Bioinformatics*, **21**, 2814–2820.

Li,W. *et al*. (2001) Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics*, **17**, 282–283.

Ng,P.C. and Henikoff,S. (2002) Accounting for human polymorphisms predicted to affect protein function. *Genome Res*., **12**, 436–446.

Ng,P.C. and Henikoff,S. (2003) SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res*., **31**, 3812–3814.

Ramensky,V. *et al*. (2002) Human non-synonymous SNPs: server and survey. *Nucleic Acids Res*., **30**, 3894–3900.

Reumers,J. *et al*. (2005) SNPeffect: a database mapping molecular phenotypic effects of human non-synonymous coding SNPs. *Nucleic Acids Res*., **33**, D527–D532.

Riva,A. and Kohane,I.S. (2002) SNPper: retrieval and analysis of human SNPs. *Bioinformatics*, **18**, 1681–1685.

Sherry,S.T. *et al*. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*., **29**, 308–311.

Stenson,P.D. *et al*. (2003) Human Gene Mutation Database (HGMD): 2003 update. *Hum. Mutat*., **21**, 577–581.

Wang,D.G. *et al*. (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science*, **280**, 1077–1082.

Wang,Z. and Moult,J. (2001) SNPs, protein structure, and disease. *Hum. Mutat*., **17**, 263–270.

Yue,P. and Moult,J. (2006) Identification and analysis of deleterious human SNPs. *J. Mol. Biol*., **356**, 1263–1274.