Improving the prediction of disease-related variants using protein three-dimensional structure.

Emidio Capriotti[*§] *and Russ B. Altman*[*‡]

Departments of Bioengineering[*] and Genetics[‡], Stanford University, Stanford (CA), United States of America; [§]Department of Mathematics and Computer Sciences, University of Balearic Islands, Palma de Mallorca, Spain.

{emidio, russ.altman}@stanford.edu

Abstract

Background:
Single Nucleotide Polymorphisms (SNPs) are an important source of human genome variability. The non-synonymous SNPs occurring in coding regions resulting in single amino acid polymorphisms (SAPs) may affect protein function and lead to pathology. Several methods attempt to estimate the impact of SAPs using different sources of information. Although sequence-based predictors have shown good performances, the quality of the prediction can be further improved introducing new features derived from the protein three-dimensional structure.
Results:
In this paper, we present a structure-based machine learning approach to predict disease-related SAPs. We have trained a Support Vector Machine (SVM) on a set of 3,342 disease-related mutations and 1,644 neutral polymorphisms from 784 protein chains. We use SVM input features from the protein sequence, structure and function information. After dataset balancing, the structure-based method reaches an overall accuracy of 84%, a correlation coefficient of 0.67, and an area under the receiving operating characteristic curve (AUC) of 0.91. When compared with a similar sequence-based predictor, structure-based method results in an increase of the overall accuracy and the AUC ~3%, and 0.06 for the correlation coefficient.
Conclusion:
This work demonstrates that structural information can increase the accuracy of detecting of disease-related SAPs. Our results also quantify the magnitude of the improvement on a large data. This improvement is in agreement with the previously observed results in the prediction of the protein stability change upon mutation.

Background

Currently the number of validated Single Nucleotide Polymorphisms (SNPs) is larger than 14 millions [1]. In general, mutations occurring in coding regions may have a greater impact on the gene functionality than those occurring in non-coding regions [2]. Only a small fraction of SNPs (~61,000) corresponds to the subset of annotated missense coding SNPs [3]. For this subset of Single Amino acid Polymorphisms (SAPs), curators of the Swiss Institute of Bioinformatics provide a classification dividing SAPs in disease-related and neutral according to peer-reviewed bibliography. In the last few year several methods have been developed to predict the impact of a given single point

protein mutation [4-16]. These algorithms are able to predict the protein stability change [10, 11, 16], the variation in protein functional activity [6] and the insurgence of human pathologies [4, 5, 7-9, 12-15]. The majority of the methods rely on information derived from protein sequence [4, 8, 9, 14], others use protein structure data [12, 17] and knowledge-based information [7, 13, 15]. In this paper we focus our attention on SAPs presenting a new machine learning based method to predict disease-related SAPs using together protein sequence, structural and functional information. We quantified the improvement of the performance resulting from the use of protein structure information.

Results

Performance of the method

In the last decades machine learning approaches have been successfully used to address several biological problems and develop new prediction methods. We modified a previously developed predictor introducing new three-dimensional structure information. In particular we use new features to describe the structural environment of the mutation considering a radius shell of 6 Å around the C-α. To quantify the improvement of the accuracy resulting from the use of 3D structure information, we compare the performances of a structure-based method (SVM-3D) with a sequence-based one (SVM-SEQ). In Tab. 1 different accuracy measures for both predictors are reported. The structure-based method results in 3% better overall accuracy and 0.06 better correlation. Comparing the ROC curves (Fig. 1 A), SVM-3D results in 0.02 better Area Under the Curve (AUC) with respect to SVM-SEQ. If 10% of wrong predictions are accepted SVM-3D has 6% more true positive. The output returned by the SVM has been used to calculate the Reliability Index (RI) and filter prediction. If predictions with RI>5 are selected the SVM-3D method results in 90% overall accuracy, 0.81 correlation coefficient on 74% of the whole dataset (see Fig 1 B). Analyzing the predictions of SVM-SEQ and SVM-3D methods we found that outputs agree in the 88% of the cases. On this subset the overall accuracy is 86% and the correlation coefficient of the method is 0.73. For the remaining 12% of the predictions, SVM-SEQ method results in a very poor overall accuracy and correlation respectively 37% and -0.25. SVM-3D performs slightly better than a random predictor resulting in 63% overall accuracy and a 0.25 correlation (see Tab 2).

Structure environment analysis

Protein three-dimensional structural information is an important feature to predict the effect of SAPs. The analysis of the protein structure provides information about the environment of the mutation. In fact, the effect of the mutation depends on the position of the mutated residue, if it is buried in the hydrophobic core or exposed on the surface of the protein. In Fig. 2 panel A the distributions of the relative solvent accessible area (RSA) for disease-related and neutral variants are plotted. The two distributions have mean RSA values of 20.6 and 35.7 respectively for disease-related and neutral variants (see Fig 2 panel A). They are significantly different and the Kolmogorov-Smirnov test returns a p-value of $2.8*10^{-71}$. We calculated the overall accuracy and correlation coefficient of our method dividing the dataset in 10 bins according to RSA value of the mutated residue. The SVM-3D method shows better performance in the prediction of buried (RSA<20) and highly exposed (RSA>80) residues (see Fig 2 panel B).

Scoring the residue interactions

The protein three-dimensional structure information is important to calculate the interactions between residues far in the sequence but close in the 3D space. We defined two types of interactions: the lost interactions are those missing after the wild-type mutation and the new interactions formed by the mutant residue. In this section we compared the frequency of lost and new interactions related to disease or neutral mutations. We calculated the log odd score for lost and new interactions respectively in panels A and B (see Fig. 3). According to these results, the most deleterious lost contacts are between and Cys-Cys and newly formed interactions between Trp-Trp are the most damaging ones. The missing Cys-Cys interactions could lead to the loss of a disulphide bond and the mutation of a residue into a Tryptophan when close to another Tryptophan could result in stereo-chemical problems.

An example of missing Cys-Cys interaction has been observed in the mutation of Cys163 in the Glycosylasparaginase (Swiss-Prot:ASPG_HUMAN). This mutation is responsible for the insurgence of the Aspartylglucosaminuria (MIM:208400). Looking at the protein structure (Fig 4), we found that the mutation of the Cys163 to Serine results in the loss of the disulfide bridge between Cys163 and Cys179 (respectively Cys140 and Cys156 in the PDB structure 1APY chain A). Interesting example of possible damaging newly formed interaction can be observed in the Thyroid hormone receptor (Swiss-Prot:THB_HUMAN) where the mutation of Arg243 into Tryptophan is cause of the Thyroid hormone resistance (MIM:188570,274300). Analyzing the protein structure (1NAX chain A) we found that the new Tryptophan could be close to another one in position 239. This mutation could result in stereo-chemical problems in the pocket around the position 243 (see Fig 5). Both the examples are correctly predicted by structure-based method and wrongly predicted by the sequence-based algorithm.

Conclusion

We developed a new machine learning approach based on protein structure information to predict the effect of SAPs. The method has been compared to a previously developed sequence-based predictor to quantify the increase of accuracy achieved by protein structure information. Using a balanced set of 6,630 mutations the structure-based method results in about 3% higher accuracy and AUC and 0.06 higher correlation with respect to sequence-based one. Although the increase the accuracy is not extremely high the introduction of structure information can be particularly useful in specific situation providing insight about the disease mechanism like in the cases discussed above. The prediction improvement is in agreement with the previously results observed in the prediction of the protein stability change upon mutation [10].

Methods

Datasets

The preformaces of machine learning methods strongly depend from the training set. This is the reason why the selection of a representative set of SAPs is a pivotal issue in the development of predictive algorithms. A previous analysis of different SAPs databases has shown that annotated set of variants from Swiss-Var database is the best available one [18]. According to this observation, we selected our set of SAP from Swiss-Var release 57.9 (Oct 2009) and we map all the variants on the protein structures available in the Protein Data Bank (PDB) [19]. To reduce the number of sequence alignments between Swiss-Prot sequences and sequences derived from the PDB, we

use a precompiled list of correspondences between Swiss-Prot and PDB codes available at the ExPASY web site. Using this list we aligned each pair of sequences using Blast algorithm [20] and filtering out alignment with: i) gaps, ii) sequence identity lower than 100% and iii) shorter than 40 residues. The remaining alignments are used to calculate the correspondence between the Swiss-Prot and PDB residue numerations. In case a mutation maps in more than one protein structure, the one with best resolution has been selected. After this filtering procedure we obtain a set of 4,986 mutations from 784 protein chains. The dataset of variants mapped into protein structures is composed by 3,342 disease-related SAPs and 1,644 neutral polymorphisms. To keep the dataset balanced we doubled the number of neutral variants considering their reverse mutation as neutral. The final set results in 6,630 mutations about equally distributed between disease-related and neutral SAPs.

Implemented SVM-based predictors

The proposed task is to predict whether a given single amino acid polymorphism is a neutral or disease-related. The task is treated as a binary classification problem for the protein upon mutation. The Support Vector Machine (SVM) input features for the structural-based predictor include: the amino acid mutation, the mutation structural environment, the sequence-profile derived features, and a functional-based log-odds score calculated considering the GO classification. The final input vector consists 48 elements:

- 20 components encoding for the mutations (Mut)
- 21 local protein structure information (3D)
- 5 inputs features derived from sequence profile (Prof)
- 2 elements encoding for the number of GO term associated to the protein and the GO log-odd score (LGO).

A similar sequence-based SVM predictor has been used to measure the increase of accuracy resulting from the use of protein three-dimensional structure information. The structure-based SVM differs only in the 21 elements vector encoding for the local protein structure environment (3D) that replaces the 20 elements vector encoding for the sequence environment. More details about the SVM input features have been described in supplementary materials.

Interaction score

The residues interactions are defined considering all the residues within a radius shell of 6 Å around the C-α of the mutated residue. According to this we calculate a log odd score dividing the frequency of lost interactions related to disease by the same type of interactions that have no pathological effect.

Although the mutations could be responsible for protein structural changes, as first approximation, we consider the position of the C-α of the new residue will not change significantly after the mutation. Hence, we consider new interactions those between the mutant residue and the residues previously interacting with the wild-type. A score of the possible damaging effect of lost or new interactions are calculated as follow

$$LC_k=\log_2[f(c_k(i,j),D)/f(c_k(i,j),N)] \quad [1]$$

where $f_k(c_k(i,j),D)$ and $f(c_k(i,j),N)$ are the frequencies of contacts between residues i and j respectively for disease-related (D) and neutral (N) variants and *k* is equal to *l* or *n* respectively for lost and new interactions.

Accuracy measures

The performances of our methods are evaluated using a 20-fold cross-validation procedure on the whole SAPs dataset. The dataset has been divided keeping the ratio of the disease-related to the neutral polymorphism mutations similar to the original distribution of the whole set. Furthermore, all the proteins in the datasets are clustered according to their sequence similarity with the *blastclust* program in the BLAST suite [20] by adopting the default value of length coverage equal to 0.9 and the percentage similarity threshold equal to 30%. We kept all the mutations belonging to a protein in the same training set to overestimate the performance. Classical accuracies measures such as the overall accuracy (Q2), the sensitivity (S), the probability of correct predictions (P), the Matthew's correlation coefficient (C), the false and true positive rates (FPR, TPR) and the area under the ROC curve (AUC) are used to score the performance of our predictors. A Reliability Index (RI) score has been calculated to select more reliable predictions. More details about the definition of the statistical index used in this work are provided in the supplementary materials.

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. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K: **dbSNP: the NCBI database of genetic variation**. *Nucleic Acids Res* 2001, **29**(1):308-311.
2. Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Shaw N, Lane CR, Lim EP, Kalyanaraman N *et al*: **Characterization of single-nucleotide polymorphisms in coding regions of human genes**. *Nat Genet* 1999, **22**(3):231-238.
3. Yip YL, Famiglietti M, Gos A, Duek PD, David FP, Gateau A, Bairoch A: **Annotating single amino acid polymorphisms in the UniProt/Swiss-Prot knowledgebase**. *Hum Mutat* 2008, **29**(3):361-366.
4. Thomas PD, Kejariwal A, Campbell MJ, Mi H, Diemer K, Guo N, Ladunga I, Ulitsky-Lazareva B, Muruganujan A, Rabkin S *et al*: **PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification**. *Nucleic Acids Res* 2003, **31**(1):334-341.
5. Wang Z, Moult J: **SNPs, protein structure, and disease**. *Hum Mutat* 2001, **17**(4):263-270.
6. Bromberg Y, Yachdav G, Rost B: **SNAP predicts effect of mutations on protein function**. *Bioinformatics* 2008, **24**(20):2397-2398.
7. Calabrese R, Capriotti E, Fariselli P, Martelli PL, Casadio R: **Functional annotations improve the predictive score of human disease-related mutations in proteins**. *Hum Mutat* 2009, **30**(8):1237-1244.
8. Capriotti E, Arbiza L, Casadio R, Dopazo J, Dopazo H, Marti-Renom MA: **Use of estimated evolutionary strength at the codon level improves the prediction of disease-related protein mutations in humans**. *Hum Mutat* 2008, **29**(1):198-204.

9. Capriotti E, Calabrese R, Casadio R: **Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information**. *Bioinformatics* 2006, **22**(22):2729-2734.
10. Capriotti E, Fariselli P, Casadio R: **I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure**. *Nucleic Acids Res* 2005, **33**(Web Server issue):W306-310.
11. Guerois R, Nielsen JE, Serrano L: **Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations**. *J Mol Biol* 2002, **320**(2):369-387.
12. Karchin R, Diekhans M, Kelly L, Thomas DJ, Pieper U, Eswar N, Haussler D, Sali A: **LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources**. *Bioinformatics* 2005, **21**(12):2814-2820.
13. Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, Cooper DN, Mooney SD, Radivojac P: **Automated inference of molecular mechanisms of disease from amino acid substitutions**. *Bioinformatics* 2009, **25**(21):2744-2750.
14. Ng PC, Henikoff S: **Predicting deleterious amino acid substitutions**. *Genome Res* 2001, **11**(5):863-874.
15. Ramensky V, Bork P, Sunyaev S: **Human non-synonymous SNPs: server and survey**. *Nucleic Acids Res* 2002, **30**(17):3894-3900.
16. Capriotti E, Fariselli P, Rossi I, Casadio R: **A three-state prediction of single point mutations on protein stability changes**. *BMC Bioinformatics* 2008, **9 Suppl 2**:S6.
17. Wong WS, Yang Z, Goldman N, Nielsen R: **Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites**. *Genetics* 2004, **168**(2):1041-1051.
18. Care MA, Needham CJ, Bulpitt AJ, Westhead DR: **Deleterious SNP prediction: be mindful of your training data!** *Bioinformatics* 2007, **23**(6):664-672.
19. Berman H, Henrick K, Nakamura H, Markley JL: **The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data**. *Nucleic Acids Res* 2007, **35**(Database issue):D301-303.
20. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res* 1997, **25**(17):3389-3402.

FIGURES

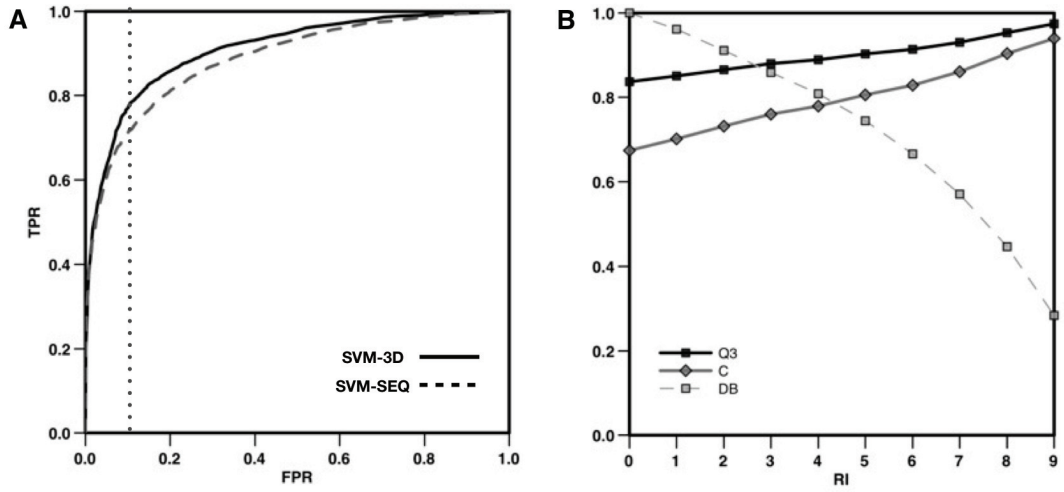

Figure 1. Performance of the SVM-based method. ROC curves of the sequence (SVM-SEQ) and structure-based methods (A) and prediction accuracy of SVM-3D as function of the Reliability Index (RI).

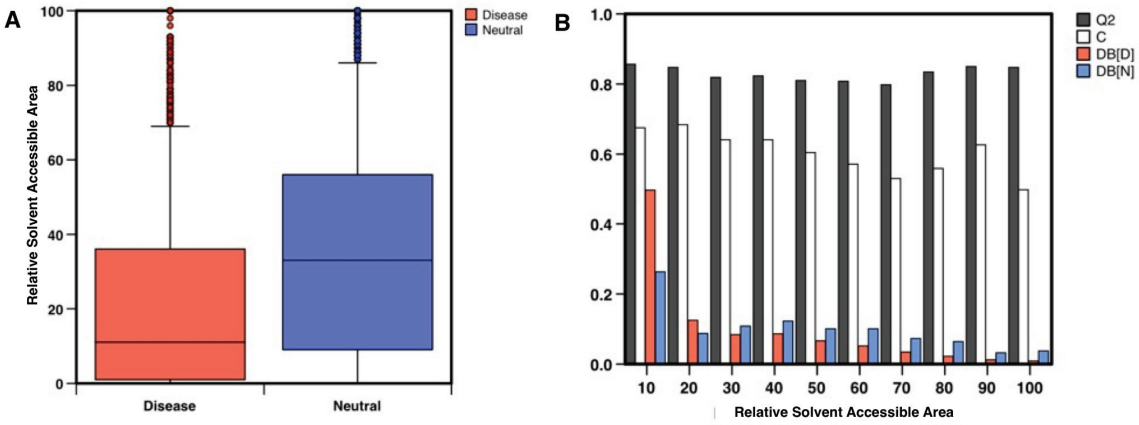

Fig. 2. Analysis of the protein three-dimensional structure environment. Distribution of relative solvent accessible area (RSA) for disease-related and neutral variants (A) and prediction accuracy as a function of the RSA (B). Accuracy measures (Q2, C) are defined in supplementary material. DB is the fraction of the whole dataset for disease-related (D) and neutral (N) mutations.

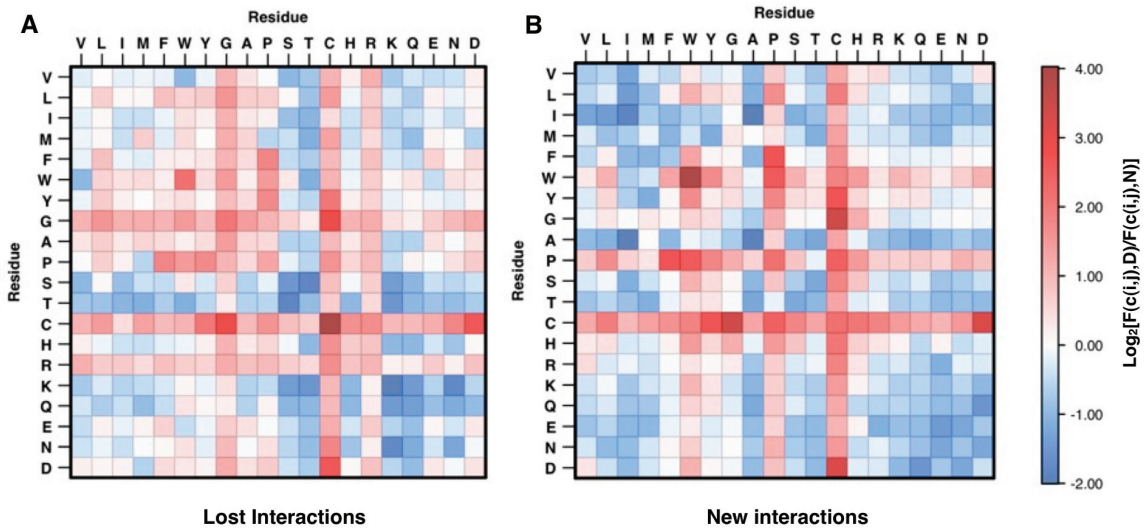

Fig. 3 Log odd score for lost residues interactions (A) and for newly formed interactions (B). The red zones correspond to damaging lost or new interactions. Bleu points correspond to neutral interactions.

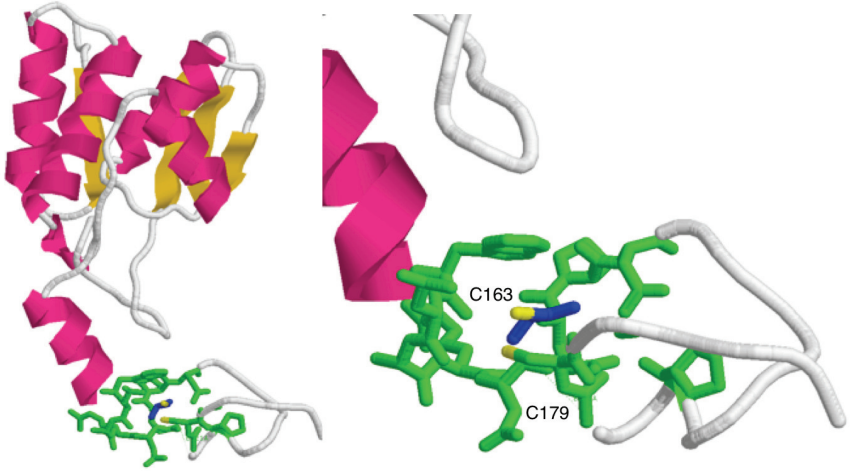

Figure 4. Structure of the Glycosylasparaginase (PDB code 1APY chain A) and details of the interactions around Cys163 (Cys140 in the PDB structure)

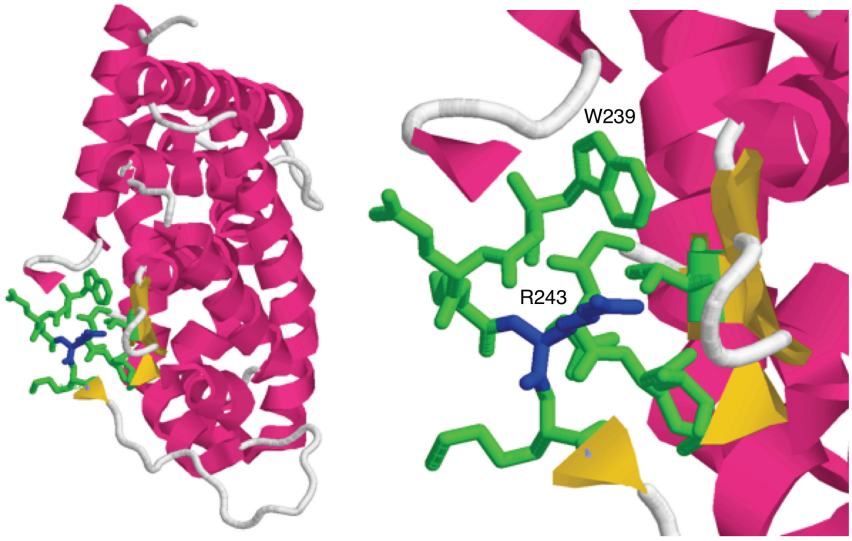

Figure 5. Structure of the Thyroid hormone receptor (PDB code 1NAX chain A) and details of the interactions around the Arg243.

TABLES

Table 1. Performances of the sequence (SVM-SEQ) and structure (SVM-3D) based methods.

	Q2	P[D]	S[D]	P[N]	S[D]	C	AUC
SVM-SEQ	0.81	0.80	0.82	0.81	0.79	0.61	0.89
SVM-3D	0.84	0.82	0.86	0.85	0.81	0.67	0.91

The accuracy measures are defined in supplementary materials. D, N stands for disease-related and neutral variants respectively.

Table 2. Performances of the methods on agree and not agree subset of predictions

	Q2	P[D]	S[D]	P[N]	S[D]	C	AUC	PM
SEQ∩3D	0.86	0.85	0.89	0.88	0.84	0.73	0.92	88
SEQ-3D	0.63	0.66	0.65	0.60	0.60	0.25	0.68	12
3D-SEQ	0.37	0.40	0.35	0.34	0.40	-0.25	0.40	12

SEQ∩3D indicates the subset of agree predictions, SEQ-3D and 3D-SEQ are respectively the predictions of SVM-SEQ and SVM-3D on the not agree prediction subset. The accuracy measures are defined in supplementary materials. PM is the fraction of the dataset. D, N stands for disease-related and neutral variants respectively.

Supplementary Material

Improving the prediction of disease-related variants using protein three-dimensional structure.

Emidio Capriotti[*§] *and Russ B. Altman*[*‡]

Departments of Bioengineering[*] and Genetics[‡], Stanford University, Stanford (CA), United States of America; [§]Department of Mathematics and Computer Sciences, University of Balearic Islands, Palma de Mallorca, Spain.

{emidio, russ.altman}@stanford.edu

Support Vector Machine (SVM) input features

The SVM-based methods developed in this work consider in input the following features: i) residue mutation; ii) protein sequence profile; iii) functional score based on Gene Ontology (GO) terms and iv) either sequence or structure mutation environment.

Encoding residue mutation

The input vector relative to mutation consists of 20 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).

Encoding mutation structure environment

The protein structural environment is encoding with a 21 elements vector. The first 20 elements encode for the number of each residue type, which have at least one heavy atom within a radius shell around the C-α of the mutated residue. After an optimization procedure a shell of 6 Å radius has been considered. The 21st element is the relative solvent accessible area calculated using the DSSP program [1].

Encoding mutation sequence environment

The 20 element input values for the mutation sequence environment (the 20 elements represent the 20 residue types) encode for the number of the each residue type, to be found inside a window centered at the residue that undergoes mutation and that symmetrically spans the sequence to the left (N-terminus) and to the right (C-terminus) with a length of 19 residues [2].

Encoding sequence profile information

We derive for each mutation: the frequency of the wild type, the frequency of the mutated residue, the number of totally and locally aligned sequences and a conservation index (CI) for the position at hand: the more a residue is functionally important the more is conserved over evolution [3]. The conservation index is calculated as:

$$CI(i)=[\Sigma_{a=1}^{20}(f_a(i)-f_a)^2]^{1/2} \quad [1]$$

where $f_a(i)$ is the relative frequency of residue a at mutated position i and f_a is the overall frequency of the same residue in the alignment. The sequence profile is computed from the output of the BLAST program [4] running on the uniref90 database (Oct 2009) (E-value threshold=10^{-9}, number of runs=1).

Functional based score

The Gene Ontology log-odds score (LGO) provides information about the correlation among a given mutation type (disease related and neutral) and the protein function. The annotation data are relative to the GO Database (version Mar 2010) and are retrieved at the web resource hosted at European Bionformatics Institute (EBI). To calculate the LGO, first we derived the GO terms from all the three branches (molecular function, biological process and cellular components) for all our proteins in the dataset. For each annotated term the appropriate ontology tree was traversed upward to retrieve all the parent terms with the GO-TermFinder tool (http://search.cpan.org/dist/GO-TermFinder/) [5] and counting a GO term only once. The log-odds score associated to each protein is calculated as:

$$LGO=\Sigma \log_2[f_{GO}(D)/f_{GO}(N)] \quad [2]$$

where f_{GO} is the frequency of occurrence of a given GO term for the disease-related (D) and neutral mutations (N) adding one pseudo-count to each class. To prevent the overfitting, the LGO scores are evaluated considering f_{GO} values computed over the training sets without including in the GO term counts of the corresponding test set.

Support Vector Machine software

The LIBSVM package (http://www.csie.ntu.edu.tw/~cjlin/libsvm/) has been used for the SVM implementation [6]. The selected SVM kernel is a Radial Basis Function (RBF) kernel $K(x_i,x_j)=\exp(-\gamma||x_i-x_j||^2)$ and γ and C parameters are optimized performing a grid like search. After input rescaling the values of the best parameters are C=8 and γ=0.03125

Statistical indexes for accuracy measure

The prediction accuracy is scored with several measures. In this paper the efficiency of our predictors have been scored using the following statistical indexes.
The overall accuracy is:

$$Q2=P/N \quad [3]$$

where P is the total number of correctly predicted mutations and N is the total number of mutations. The Matthew's correlation coefficient C is defined as:

$$C(s)=[p(s)n(s)-u(s)o(s)] / D \quad [4]$$

where D is the normalization factor:

$$D = [(p(s)+u(s))(p(s)+o(s))(n(s)+u(s))(n(s)+o(s))]^{1/2} \quad [5]$$

for each class s (D and N, stand for disease-related and neutral mutations 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 false negative and false positive for the class s.
The coverage S (sensitivity) for each discriminated class s is evaluated as:

$$S(s)=p(s)/[p(s)+u(s)] \quad [6]$$

where p(s) and u(s) are the same as in Equation 5.
The probability of correct predictions P (or positive predictive values) is computed as:

$$P(s)=p(s) / [p(s) + o(s)] \quad [7]$$

where p(s) and o(s) are the same as in Equation 5 (ranging from 0 to 1).
For each prediction a reliability score (RI) is calculated as follows:

$$RI=20*abs|O(D)-0.5| \quad [8]$$

where O(D) is the SVM output. Other standard scoring measures, such as the area under the ROC curve (AUC) and the true positive rate (TPR= Q(s)) at 10% of False Positive Rate (FPR= 1-P(s)) are also computed [7].

References

1. Kabsch W, Sander C: **Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features**. *Biopolymers* 1983, **22**(12):2577-2637.
2. Capriotti E, Calabrese R, Casadio R: **Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information**. *Bioinformatics* 2006, **22**(22):2729-2734.
3. Pei J, Grishin NV: **AL2CO: calculation of positional conservation in a protein sequence alignment**. *Bioinformatics* 2001, **17**(8):700-712.
4. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res* 1997, **25**(17):3389-3402.
5. Boyle EI, Weng S, Gollub J, Jin H, Botstein D, Cherry JM, Sherlock G: **GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes**. *Bioinformatics* 2004, **20**(18):3710-3715.
6. Chang CC, Lin CJ: **Training nu-support vector classifiers: theory and algorithms**. *Neural Comput* 2001, **13**(9):2119-2147.
7. Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H: **Assessing the accuracy of prediction algorithms for classification: an overview**. *Bioinformatics* 2000, **16**(5):412-424.