Ročník 2012 electroscope Číslo V

System for Functional Annotation of Single Nucleotide Polymorphisms

J. Bendl [1], J. Zendulka[1]
[1] Department of Information Systems, Brno University of Technology,
Božetěchova 1, Brno

Abstract:
Single nucleotide polymorphisms are the substitution of one nucleotide in the DNA sequence that may or may not have phenotypic consequences. Here we describe a new system for ranking non-synonymous protein substitutions by their deleterious effects. The computational core of the proposed system is based on a rational combination of the results from the selected subset of publicly available tools. The weight coefficients for the individual tools are calculated on the basis of their confidence score and their reliabilities are assigned accordingly to their performance measured on the extensive dataset. The validation of the performance on the dataset consisting of 5 000 substitutions shows that overall accuracy of the system was improved by 6% in comparison to the simple majority vote.

INTRODUCTION

Human genetic variation occurs primarily as a result of single nucleotide polymorphism (SNPs) [8]. SNP is the substitution of one nucleotide in the DNA sequence for another with the frequency about 0.1%. Although the most of these substitutions are considered as neutral, some substitutions can affect a gene expression or the function of a translated protein. SNPs can have drastic phenotypic consequences leading to the development of various diseases. Approximately half of the known disease-causing mutations are the result of amino acid substitutions [8]. Thus it is very important to distinguish non-neutral substitutions that affect protein function from those that are functionally neutral. There are many computational methods for predicting the effect of amino acid substitution on a protein function, however, these methods show limited reliability and accuracy. The main reason for their limited reliability lies in the fact that they were trained on datasets which were not sufficiently diverse. They also employ different principles of decision making, some of which work well on one type of dataset, but fail on another [8]. Today, there are many tools for predicting the effect of amino acid substitution on protein function. Most of these tools are designed to predict whether the substitution is neutral or deleterious [8]. Decision about the effect on function is made on the basis of parameters derived from the evolutionary information, e.g., MAPP [12], Panther [13], PhD-SNP , SIFT [9], or from combination of sequence with structural or functional characteristics, e.g., MutPred [7], nsSNPAnalyzer [4], PolyPhen-1 [10], PolyPhen-2 [1] and SNAP [2].

The prediction using the methods based on the sequence information assumes that the amino acids important for the function are conserved within the protein family. The algorithms find related sequences in the databases and create multiple sequence alignment. Then the rate of conservation on individual positions is determined [3]. The properties of the amino acids can also be taken into account, e.g., if there is only hydrophobic amino acid on one specific position, its change to polar amino acid is mostly considered as deleterious. Structure-based prediction methods find the best match of the input sequence with the sequences available in the database of protein structures. These prediction methods use the general structural features surrounding the site of substitution, and thus do not require specific information at the atomic level. For this reason, they can model the substitution in the structure of a homologous protein without a need of the exact structure of the input sequence [3]. They often take into account several structural factors of substituted amino acid, e.g., solvent accessibility, crystallographic B-factor, or the difference in the free energy after introduction of substituting amino acid. Some of these prediction methods use annotations to refine the prediction. Annotations provide information about function of a particular position in the protein. Amino acids on positions belonging to the binding site, active site or forming a disulfide bond are considered as deleterious [8].

PREDICTION METHODS

As the current prediction tools show limited accuracy, the main purpose of the methodology presented here is to combine the existing tools to obtain more reliable results. The idea of improving accuracy by applying consensus was proposed in previous study [11], which successfully combined methods employing only conservation analysis, with the method employing only structural parameters. The most important criterion for the selection of tools is their performance on testing dataset. Other significant factors include the number of citations of the article describing a given method, or the average speed of the tool. Finally, the algorithm used for the prediction and the level of its description is also taken into

consideration as the diversity of used techniques is the cornerstone for obtaining more accurate results.

The list of selected tools with their short descriptions is shown in the Table 1.***Table 1****: Summary of the integrated methods for analysis the effect of non-synonymous substitutions.* All of them, except but MAPP, SIFT and Panther, use the machine learning methods for the predictor construction. For these methods it is particularly important to choose a suitable training dataset, since the quality of dataset affects the result to the same extend, as the selection of the decision attributes. The dataset should not only be sufficiently large in the sense of number of entries but it should also be sufficiently diverse to enable the universal prediction. If the final decision about pathogenicity is based on the conservation analysis, the quality of multiple sequence alignment (MSA) is crucial. In terms of objectivity of the overall results, it would be desirable to use the same MSA for all methods which employ it. Unfortunately only MAPP and SIFT enable the insertion of user defined MSA. If the tools offer additional parameters, e.g., choice of structural database for finding homologs, default setting is automatically chosen. With the exception of MAPP, all the tools were queried remotely.

Table 1: Summary of the integrated methods for analysis the effect of non-synonymous substitutions.

Method	Principle	Inputs for predictor
MAPP	Alignment score	Conservation analysis (with using own alignment and phylogenetic tree)
MutPred	Decision tree (random forest)	Conservation analysis Structural parameters Functional parameters
nsSNP Analyzer	Decision tree (random forest)	Conservation analysis Structural parameters
Panther	Alignment score	Conservation analysis (with using Panther library and Hidden Markov models)
PhD-SNP	Support vector machine	Conservation analysis (with using sequence environment, sequence profiles and Panther)
PolyPhen-1	Rule-based classifier	Conservation analysis (with using position-specific independent count profiles) Structural parameters (derived from homologous structure + predicted by known methods) Annotation generated from SwissPort
PolyPhen-2	Naive bayes classifier	Conservation analysis (with using with using position-specific independent count profiles) Structural parameters (derived from homologous structure + predicted by known methods) Annotation generated from SwissPort
SIFT	Alignment score	Conservation analysis (with using own / generated alignment)
SNAP	Feed-forward neural network	Conservation analysis (with using Pfam, with using position-specific independent count profiles and PSI-BLAST) Annotation generated from SwissProt

CONSTRUCTION OF THE CONSENSUS FUNCTION

A key step in the development of the integrative scoring system is the design and implementation of computational framework, which defines the way to combine the results from the individual tools. With the exception of nsSNPAnalyzer, all of the selected tools offer a way to estimate the degree of pathogenicity for evaluated mutation, so called *confidence score*, which it is unique for each pair tool and mutation. Another important parameter of given tool is its performance on testing dataset, so called *tool reliability*, which is unique for each tool. These two values are combined with prediction for a given mutation (neutral/deleterious) in the process that is further described in details using mathematic notation.

Suppose there are q different integrated prediction tools and p non-synonymous amino acid substitutions. Each of them is expressed as a discrete variable X_i $(i=1, ..., p)$ which carries the value of amino acid replacing wild-type at the given position. Then, for each SNP and each tool there is a specific prediction δ_{ij} $(i=1, ..., p; j=1, ..., q)$ which is assigned *1*: if tool prediction for this SNP is be deleterious and *-1* otherwise. Most of the tools also provide confidence score S_{ij} which represents the degree of confidence of the given tool in its own decision where higher value means higher confidence. Because scales of the confidence scores of the individual tools are different, the S_{ij} has to be transformed into $\overline{S_{ij}}$ which carries confidence scores normalized to the continuous interval $<0,1>$. The normalized confidence score $\overline{S_{ij}}$ for the given tool is calculated on the basis of corresponded equation from the Table 2. The tools MAPP and nsSNPAnalyzer, which do not provide confidence score, derive this value according to the weighted arithmetic mean of confidence scores of tools with the same result prediction of the pathogenicity (neutral/deleterious). If there is not any tool with the same result prediction, default value *0.5* is used.

Table 2: Summary of the methods of calculation of normalized confidence score for the integrated tools.

Method	Calculation of the norm. confidence score
MutPred	derived from the probability score (*general score*) $\overline{S_{ij}} = (0.5 - delScore)2\ldots for\ delScore \in \langle 0,0.5 \rangle$ $\overline{S_{ij}} = (delScore - 0.5)\ldots otherwise$
Panther	derived from the probability score (*pScore*) $\overline{S_{ij}} = (0.5 - delScore) \cdot 2\ldots for\ delScore \in \langle 0,0.5 \rangle$ $\overline{S_{ij}} = (delScore - 0.5)\ldots otherwise$
PhD-SNP	derived from the reliability index (*relIndex*): integer value belongs to the interval *<1, 9>* where lower value express lower confidence $\overline{S_{ij}} = \frac{(relIndex - toolMinRelIndex + 1)}{(toolMaxRelIndex - toolMinRelIndex + 1)}$ $where\ toolMinRelIndex = 0, toolMaxRelIndex = 10$
PolyPhen1	derived from the assigned pathogenic cathegory: possible values: possibly damaging, probably damaging, possibly neutral, probably neutral $\overline{S_{ij}} = 0.5\ldots for\ possibly\ damag./neutral$ $\overline{S_{ij}} = 1 for\ cathegories\ probably\ damag,/neutral$
PolyPhen2	derived from the probability score (*pScore*): value from the continuous interval *<0, 1>*: <0, 0.5> … deleterious, *(0.5, 1>* … benign, *0.5* … neutral $\overline{S_{ij}} = \frac{(relIndex - toolMinRelIndex + 1)}{(toolMaxRelIndex - toolMinRelIndex + 1)}$ $where\ toolMinRelIndex = 0, toolMaxRelIndex = 10$
SIFT	derived from the median of sequence conservation: value from the continuous interval <0,4>: *median = log2(X)*, where *X* is number of amino acids which are not occurring on the given position in MSA. $\overline{S_{ij}} = 1\ldots for\ median > 3.25$ $\overline{S_{ij}} = 1 - \frac{2^{median} - 10}{10}\ldots otherwise$
SNAP	derived from the reliability index (*relIndex*) integer value belongs to the interval *<1, 9>* where lower value express lower confidence $\overline{S_{ij}} = \frac{(relIndex - toolMinRelIndex + 1)}{(toolMaxRelIndex - toolMinRelIndex + 1)}$ $where\ toolMinRelIndex = 1, toolMaxRelIndex = 9$

While S_{ij} expresses confidence of the tool for its own decision, continuous variable TR_j *(j=1, ..., ...p)*, belonging to the interval *<0,1>*, expresses the overall tool reliability. TR_j was assigned to individual tools according to their Matthews correlation coefficient (MCC) obtained from the tools performance evaluation on the extensive dataset (see section Experiments and results). MCC allows to handle unbalanced classes and therefore it is regarded as more significant assessment than other performance measures [3]. This coefficient belongs to the interval *<-1,1>*, where *i* means perfect prediction, *0* means average random prediction and *-1* means an inverse prediction. Finally, using the introduced mathematical notation, the prediction score is defined as follows:

$$PS_i = \frac{\sum_{j=1}^{q} TR_j \cdot \delta_{ij} \cdot \overline{S_{ij}}}{\sum_{j=1}^{q} TRj} \qquad (1)$$

The permitted values of the variable PS_i belong to the continuous interval *<-1,1>*. The substitutions are considered to be neutral for the values from the interval *<-1,0)* and they are considered to be deleterious for the values from the interval *(0,1>*. If the PS_i is equal to *0*, it is not possible to predict pathogenicity. The absolute distance of the prediction score from zero expresses confidence of predictor about its own decision.

EXPERIMENTS AND RESULTS

The presented consensus function was validated with the subset of dbSNP database containing 5 000 mutations. This database is a free public archive for genetic variation within and across different species [15] and it was filtered exclusively for single nucleotide polymorphisms (SNPs). The distinction between disease-causing missense variants and neutral variants was performed on the basis of activity code, associated with the given record (the activity code [=] was considered as disease-causing and all the others as neutral). The efficiency of the proposed predictor has been scored by using the following statistical measures (in the following equations, parameters *TP*, *TN*, *FP*, *FN* refer to true positive, true negative, false positive, false negative):

- $Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$,
- $Precision = \frac{TP}{TP + FP}$,
- $Sensitivity = \frac{TP}{TP + FN}$,
- $Specificity = \frac{TN}{TN + FP}$,
- $NPV = \frac{TN}{TN + FN}$,
- $MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FN) \times (TP + FP)}}$.

The first experiment compares the performance of proposed system on the testing dataset with the results of the individual integrated tools. Weighted consensus obtained the highest scores with respect to accuracy, sensitivity, specificity and MCC among all integrated tools, with exception of MutPred, and also

significantly surpassed simple majority vote (Table 3).

Table 3: Performance evaluation of the integrated methods, simple majority vote and weighted consensus calculated according by the description in section Methods.

Method	MAPP	MutPred	nsSNP Analyzer	Panther	PhD-SNP	Poly-Phen1
TP	1 429	1 927	626	666	1 678	1 429
FN	1 058	81	351	411	330	577
TN	2 029	1 818	175	1 319	1 899	1 733
FP	469	645	86	380	525	689
Cases+	2 487	2 008	977	1 077	2 007	2 006
Cases-	2 498	2 463	261	1 699	2 423	2 422
Acc.	0.694	0.838	0.646	0.715	0.807	0.714
Prec.	0.753	0.749	0.879	0.637	0.762	0.675
Spec.	0.812	0.738	0.670	0.776	0.783	0.716
Sens.	0.575	0.959	0.640	0.619	0.836	0.712
NPV	0.657	0.967	0.332	0.763	0.852	0.750
MCC	0.429	0.812	0.379	0.448	0.700	0.506

Method	Poly-Phen2	SIFT	SNAP	Majority vote	**Weight. consen.**
TP	1 754	1 442	1 289	1 699	**1 770**
FN	254	257	708	309	**238**
TN	1 416	1 231	1 842	1 670	**1 864**
FP	1 044	705	601	793	**599**
Cases+	2 008	1 699	1 998	2 487	**2 487**
Cases-	2 461	1 936	2 442	2 498	**2 498**
Acc.	0.709	0.735	0.705	0.753	**0.813**
Prec.	0.627	0.672	0.682	0.682	**0.747**
Spec.	0.576	0.636	0.754	0.678	**0.757**
Sens.	0.873	0.849	0.645	0.846	**0.881**
NPV	0.848	0.827	0.722	0.844	**0.887**
MCC	0.592	0.608	0.460	0.636	**0.734**

Cases+, Cases- express the absolute number of deleterious mutations, respective benign mutations from the original dataset for which the given tool was able to predict any pathogenicity class (unknown predictions are not taken into consideration). NPV denotes negative predictive value and MCC denotes Matthews correlation coefficient.

The second experiment compares the distribution of prediction score of the proposed weighted consensus function for neutral and pathogenic dataset. The Figure 1 shows significant difference in the distributions of prediction scores. While the median of the neutral dataset is *-0.31*, the median of the pathogenic dataset is *0.49*. With regards to the decision threshold set to *0* only *22.5%* of the neutral substitutions have incorrect prediction score *> 0* and similarly only *17.8%* of pathogenic substitutions have incorrect prediction score *< 0*.

CONCLUSIONS

The present paper describes a new integrative scoring system for assessment of pathogenicity of non-synonymous protein substitutions. The system integrates nine existing tools and combines their individual results to obtain a more robust prediction. The increased robustness of the system was confirmed in the validation of the performance on dataset consisting of 5 000 substitutions, where both high sensitivity and high specificity was attained at the same time. The overall accuracy of the introduced weighted consensus is about 6% better than a simple majority vote.

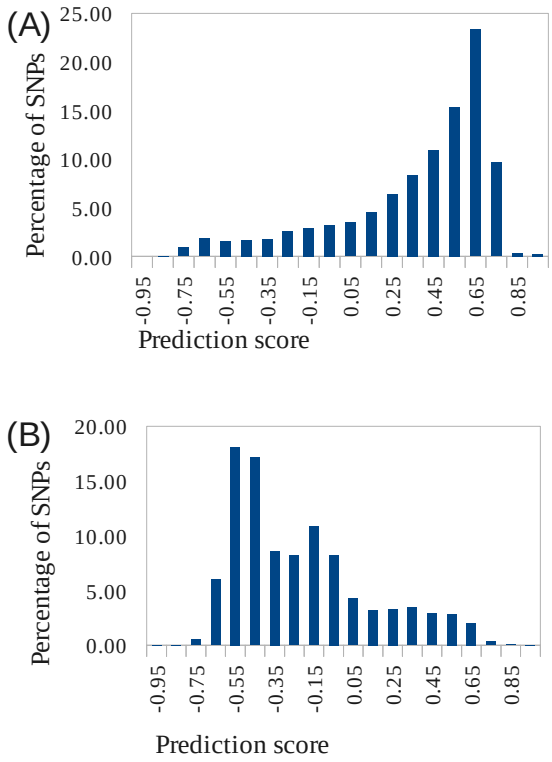

Figure 1: The distribution of prediction scores for (A) pathogenic and (B) neutral dataset.

ACKNOWLEDGEMENTS

The authors thank the colleagues from Loschmidt laboratories for valuable and inspiring consultations, advices and recommendations. MetaCentrum is acknowledged for providing access to computing and data storage facilities, provided under the programme LM2010005 funded by the Ministry of Education of the Czech Republic. This work was partially supported by the European Regional Development Fund CZ.1.05/1.1.00/02.0123, the research plan MSM0021630528, the specific research grant FIT-S-11-2 and the IT4Innovations Centre of Excellence CZ.1.05/1.1.00/02.0070.

CITATIONS

[1] Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., et al. *A method and server for predicting damaging missense mutations*, Nature Methods, volume 7, pp. 248–249, 2010.

[2] Bromberg, Y., Rost, B. *SNAP: predict effect of non-synonymous polymorphisms on function*. Nucleid Acids Research, volume 35, pp. 3823–3835, 2007.

[3] Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A. F., Nielsen, H. *Assessing the accuracy of prediction algorithms for classification: an overview*. Bioinformatics, volume 16, pp.412–424, 2000.

[4] Bao, L., Zhou, M., Cui, Y. *nsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms*. Nucleic Acids Research, volume 33, pp. 480–482, 2005.

[5] Calabrese, R., Capriotti, E., Fariselli, P., Martelli, P. L., Casadio, R. *Functional annotations improve the predictive score of human disease-related mutations in proteins*. Human Mutation, volume 30, 1237–1244, 2009.

[6] 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, volume 22, pp. 2729–2734, 2006.

[7] Li, B., Krishnan, V. G., Mort, M. E., Xin, F., Kamati, K. K., Cooper, D. N., Mooney, S. D., et al. *Automated inference of molecular mechanisms of disease from amino acid substitutions*. Bioinformatics, volume 25, pp. 2744–2750, 2009.

[8] Ng, P.C., Henikoff, S. *Predicting the effects of amino acid substitutions on protein function*. Annual Review of Genomics Human Genetics, volume 7, pp. 61–80, 2006.

[9] Ng, P. C., Henikoff, S. *SIFT: Predicting amino acid changes that affect protein function*. Nucleic Acids Research, volume 31, pp. 3812–3814, 2003.

[10] Ramensky, V., Bork, P., Sunyaev, S. *Human non-synonymous SNPs: server and survey*. Nucleic Acids Research, volume 30, 3894–3900, 2002.

[11] Saunders, C. T., Baker, D. *Evaluation of structural and evolutionary contributions to deleterious mutation prediction*. Journal of Molecular Biology, volume 322, pp. 891–901, 2002.

[12] Stone, E. A., Sidow, A. *Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity*. Genome Research, volume 15, pp. 978–986, 2005.

[13] Thomas, P. D., Campbell, M. J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., et al. *PANTHER: A Library of Protein Families and Subfamilies Indexed by Function.* Genome Research, volume 13, pp. 2129–2141, 2003.

[14] Thusberg, J., Vihinen, M.: *Pathogenic or not? And if so, then how? Studying the effects of missense mutations using bioinformatics methods*. Human Mutation, volume 30, pp. 703–714, 2009.

[15] Smigielski, E. M., Sirotkin, K., Ward, M., Sherry, S. T. *dbSNP: a database of single nucleotide polymorphisms*. Nucl. Acids Res., volume 28, pp. 352–355, 2000.