INTEGRATION SYSTEM FOR FUNCTIONAL ANNOTATION OF SINGLE NUCLEOTIDE POLYMORPHISM

Jaroslav Bendl
Doctoral Degree Programme (1), FIT BUT
E-mail: ibendl@stud.fit.vutbr.cz

Supervised by: Jaroslav Zendulka
E-mail: zendulka@vutbr.cz

Abstract: This paper describes a new integrative system for ranking non-synonymous protein substitutions by their deleterious effects. The computational core of the proposed system is based on a sophisticated combination of results from the selected subset of existing tools. The weight coefficients for individual tools are calculated on the basis of their confidence score and the tool reliability which are assigned accordingly to the tool performance measured on the extensive dataset. The performance validation on the dataset consisting of 5 000 substitutions shows that overall accuracy of the system was improved by 8% in comparison to the best of the integrated methods.

Keywords: SNP, nsSNP, pathogenicity prediction, protein substitutions

1 INTRODUCTION

Human genetic variation occurs primarily as a result of single nucleotide polymorphism (SNPs) [3]. SNP is the substitution of one nucleotide in the DNA sequence for another with the frequency about 0.1%. Although most of these substitutions are considered as neutral, some substitutions in and around gene can affect genes expression or the function of the produced proteins. 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 [3]. 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 effects of amino acid substitution on protein function, however, these methods are still not reliable and accurate enough. The main reason for their unreliability lies in the fact that they were trained on datasets which were not diverse enough. They also employ different principles of decision making, some of which work well on one type of dataset but fail on another [3]. Today, there are many tools for predicting the effect of amino acid substitution on protein function and stability. Most of these tools are designed to predict whether the substitution is benign or deleterious [5]. Decision about pathogenicity is made on the basis of parameters derived from evolutionary information (MAPP, Panther, PhD-SNP, SIFT) or from combination of sequence with structural or functional characteristics (MutPred, nsSNPAnalyzer, PolyPhen, PolyPhen2, SNAP, SNPs&GO).

The computation of prediction in methods concerning sequence information is based on the idea that amino acids which are important for the correct function of protein are conserved sequences belonging to the same protein family. Suitable algorithm finds related sequences in databases, and creates multiple sequence alignment. Then the rate of conservation on individual positions is determined [1]. Also properties of amino acids can be taken into account, e.g., if there is only hydrophobic amino acid on one concrete position, its change to polar amino acid is mostly considered as deleterious. Structure- based prediction methods find the best match of input sequence against protein structure database. These prediction methods use 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 onto the structure of a homologous protein without need of the exact structure of the input sequence [1]. They often take into account several structural factors of substituted amino acid such as solvent accessibility, crystallographic B-factor, or the difference in the free energy after introduction of new amino acid instead of the original one. Some of the prediction methods use annotations to refine the prediction. Annotations provide information about function of 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.

2 PREDICTION METHODS

As the current tools, which employ the previously described techniques in many different ways, are not accurate, the main purpose of the presented integrative tool is to combine the existing tools to obtain more reliable results. The idea of improving accuracy by applying consensus was proposed in previous study [4] which sucessfully combined methods employing only conservation analysis with method employing only structural parameters to provide better results. The most important criterion for the selection of tools for final subset 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 of prediction and the level of its description is also taken into consideration as the diversity of the used techniques is the cornerstone for obtaining more accurate results. The list of selected tools with their short descriptions is shown in table 1. If the final decision about pathogenicity is based on the conservation analysis, the quality of multiple sequence alignment (MSA) is crucial. Therefore, it would be desirable to use the same MSA for all methods which employ MSA to assure objectivity of overall results. Since only MAPP and SIFT enable the insertion of user defined MSA, the MSAs provided by individual tools were used. When tools offered additional parameters (e.g. choice of structural database for finding of homologs), default setting was applied. All integrated tools were queried remotely with the exception of MAPP which was installed locally.

3 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 (this is 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 (this is called *tool reliability*, which is unique for each tool). The multiplication of these two parameters defines the weight coefficient for each tool. This weight coefficient is applied on prediction for a given mutation (neutral / deleterious) in the process that is further described in details using mathematic notation. Generally speaking, the method of calculation of the overall prediction is based on the commonly used variant of the weighted average trainable combiner [2].

Suppose there are q different integrated prediction tools and p non-synonymous amino acid substitutions. Each of them is expressed as a discrete variable $X_i (i = 1, \ldots, p)$ 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 $\delta_{ij} (i = 1, \ldots, p; j = 1, \ldots, q)$ 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 $\langle 0, 1 \rangle$. The normalized confidence score $\overline{S_{ij}}$ for the given tool is calculated on the basis of corresponded equation from 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

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

Method	Principle	Training dataset	Inputs for predictor
MAPP http://mendel.stanford.edu /SidowLab/downloads/MAPP	Alignment score	-	Conservation analysis (with using own alignment and phylogenetic tree)
nsSNPAnalyzer http://snpanalyzer.uthsc.edu	Decision tree (random forest)	SwissProt 3 512 deleterious 503 neutral	Conservation analysis (with using SIFT) Structural parameters (derived from homologous structure)
Panther http://www.pantherdb.org	Alignment score	-	Conservation analysis (with using Panther library and Hidden Markov Model)
PolyPhen http://genetics.bwh. harvard.edu/pph	Rule-based classifier	HGVbase, hsSWALL 11 152 deleterious 9 310 neutral	Conservation analysis (with using PSIC profiles) Structural parameters (derived from homologous structure + predicted by known methods) Annotation generated from SwissProt
PolyPhen-2 http://genetics.bwh. harvard.edu/pph2	Naive bayes classifier	UniProtKB, dbSNP HumDiv: 3 155 deleterious, 6 321 neutral HumVar: 13 032 deleterious, 8 946 neutral	Conservation analysis (with using PSIC profiles) Structural parameters (derived from homologous structure + predicted by known methods) Annotation generated from SwissProt
SIFT http://sift.bii.a-star.edu.sg	Alignment score	-	Conservation analysis (with using own / generated alignment)
SNAP http://rostlab.org/ services/snap	Feed-forward neural network	Swiss-Prot, PMD 39 987 deleterious 40 830 neutral	Conservation analysis (with using Pfam, PSIC profiles and PSI-BLAST) Annotation generated from SwissProt
SNPs&GO http://snps.uib.es/snps-and-go	Support vector machine	Swiss-Prot 19 234 deleterious 19 234 neutral	Conservation analysis (with using sequence environment, sequence profiles and Panther) Annotation generated from Gene Ontology

Explanatory notes: HGDM - Human Gene Mutation Database, hsSWALL - homo sapiens subset of the SWALL database (SWALL - SwissProt + TrEMBL), HGVbase - Human Genome Variation database, dbSNP - Single Nucleotide Polymorphism Database, PMD - Protein Mutant Database, PSIC - position-specific independent counts, HumDiv and HumVar - different datasets for training neural network (HumVar is suitable for distinguishing mutations which causes Mendelian diseases, HumVar is suitable for distinguishing mutations which causes complex diseases).

result prediction of pathogenicity (neutral / deleterious). If there is not any tool with the same result prediction, default value 0.5 is used.

While S_{ij} expresses confidence of the tool for its own decision, continuous variable $TR_j (j = 1, \ldots, p)$, belonging to the interval $\langle 0, 1 \rangle$, 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 4). MCC allows to handle unbalanced classes and therefore it is regarded as more significant assessment than other performance measures [1]. This coefficient belongs to the interval $\langle -1, 1 \rangle$, where 1 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:

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

The permitted values of the variable PS_i belong to the continuous interval $\langle -1, 1 \rangle$. The substitutions are considered to be neutral for the values from the interval $\langle -1, 0)$ and they are considered to be deleterious for the values from the interval $(0, 1\rangle$. If the PS_i is equal to 0, it is not possible to pre-

dict pathogenicity. The absolute distance of the prediction score from zero expresses confidence of predictor about its own decision.

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

Method	Confidence score	Calculation of the normalization
Panther	derived from the probability score (*pScore*) value from the continuous interval $\langle 0,1\rangle$: $\langle 0,0.5)\ldots$benign, $(0.5,1\rangle\ldots$deleterious, $0.5\ldots$unknown	$\overline{S_{ij}} = \begin{cases} (0.5 - \text{delScore}) * 2 & \text{for delScore} \in \langle 0,0.5) \\ \text{delScore} - 0.5 & \text{otherwise} \end{cases}$
PolyPhen	derived from the assigned cathegory of pathogenicity possible values: possibly damaging, probably damaging, possibly neutral, probably neutral	$\overline{S_{ij}} = \begin{cases} 0.5 & \text{for cathegories } \textit{possibly damaging / neutral} \\ 1 & \text{for cathegories } \textit{probably damaging / neutral} \end{cases}$
PolyPhen-2	derived from the probability score (*pScore*) value from the continuous interval $\langle 0,1\rangle$: $\langle 0,0.5)\ldots$deleterious, $(0.5,1\rangle\ldots$benign, $0.5\ldots$unknown	$\overline{S_{ij}} = \begin{cases} (0.5 - pDeleterious) * 2 & \text{for pScore} \in \langle 0.5,1\rangle \\ pDeleterious - 0.5 & \text{otherwise} \end{cases}$
SIFT	derived from the median of sequence conservation value from the continuous interval $\langle 0,4\rangle$: $median = log2(X)$, where X is number of amino-acids which are not ocurring on the given position in MSA.	$\overline{S_{ij}} = \begin{cases} 1 & \text{for median} > 3.25 \\ 1 - \frac{2^{median} - 10}{10} & \text{otherwise} \end{cases}$
SNAP	derived from the reliability index (*relIndex*) integer value belong to the interval $\langle 1,9\rangle$ where lower value expresses lower confidence	$\overline{S_{ij}} = \frac{(relIndex - toolMinRelIndex+1))}{(toolMaxRelIndex - toolMinRelIndex+1)}$, where toolMinRelIndex=1, toolMaxRelIndex=9
SNPsGO	derived from the reliability index (*relIndex*) integer value belong to the interval $\langle 0,10\rangle$, where lower value expresses lower confidence	$\overline{S_{ij}} = \frac{(relIndex - toolMinRelIndex+1))}{(toolMaxRelIndex - toolMinRelIndex+1)}$, where toolMinRelIndex=0, toolMaxRelIndex=10

4 EXPERIMENTS AND RESULTS

The presented consensus function was validated on 5 000 randomly chosen substitutions from benchmark database suite Varibench containing information for experimentally verified effects of amino acid substitutions on protein function [5]. The first half of the substitutions were selected from the dataset of disease-causing missense variations (positive dataset) and the second half from the dataset of neutral high frequency SNPs (negative dataset). The efficiency of the proposed predictor has been scored by using the following statistical measures (in following equations, parameters *TP*, *TN*, *FP*, *FN* refer to true positive, true negative, false positive, false negative):

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}, \tag{2}$$

$$\text{Sensitivity} = \frac{TP}{TP+FN}, \tag{3}$$

$$\text{Specificity} = \frac{TN}{TN+FP}, \tag{4}$$

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TN+FN) \times (TN+FP) \times (TP+FN) \times (TP+FP)}}. \tag{5}$$

The 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 and MCC among all integrated tools, and also 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 3.

	TP	TN	FP	FN	Cases+	Cases-	Accuracy	Sensitivity	Specificity	MCC
MAPP	1 629	1 836	659	870	2 499	2 495	0.69	0.65	0.74	0.39
nsSNPAnalyzer	780	729	441	548	1 328	1 170	0.60	0.59	0.62	0.21
Panther	918	740	262	408	1 328	1 002	0.71	0.69	0.73	0.42
PolyPhen	1 397	2 133	340	1 098	2 495	2 473	0.71	0.56	0.82	0.47
PolyPhen-2	2 135	984	727	354	2 489	1 711	0.74	0.86	0.58	0.47
SIFT	1 742	1 927	518	758	2 500	2 445	0.74	0.70	0.79	0.49
SNAP	2 188	1 108	940	311	2 499	2 048	0.72	0.88	0.54	0.47
SNPs&GO	1 528	2 204	265	966	2 494	2 469	0.75	0.61	0.89	0.55
Majority vote	1 911	2 027	473	589	2 500	2 500	0.78	0.76	0.81	0.58
Weighted consensus	2 051	2 123	372	449	2 500	2 500	0.83	0.82	0.85	0.67

Explanatory notes: 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 in consideration).

5 CONCLUSION

The present paper describes a new integrative scoring system for assessment of pathogenicity of non-synonymous protein substitutions which integrates eight existing tools and combines their individual results to obtain more robust prediction. The increased robustness of the system was confirmed in performance validation on the dataset consisting of 5 000 substitutions where both high sensitivity and high specificity was attained at the same time. Moreover, the overall performance of the system was significantly improved by 8% (accuracy) and 0.12 (MCC) in comparison to the best of the integrated methods.

ACKNOWLEDGEMENT

The author thanks 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.

REFERENCES

[1] Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A. F. (2000). Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics, 16(5), 412-424. ISSN: 1367-4803.

[2] Kuncheva, L. I. (2004). Combining Pattern Classifiers: Methods and Algorithms. Wiley-Interscience. ISBN: 0471210781.

[3] Ng, P.C., Henikoff, S (2006). Predicting the effects of amino acid substitutions on protein function. Annual Review of Genomics Human Genetics 7, 61-80. ISSN: 1750-2799.

[4] Saunders, C. T., et at. (2002). Evaluation of structural and evolutionary contributions to deleterious mutation prediction. Journal of Molecular Biology, 322(4), 891-901. ISSN: 0022-2836.

[5] Thusberg, J., Olatubosun, A., Vihinen, M. (2011). Performance of mutation pathogenicity prediction methods on missense variants. Human Mutation, 32(4), 358-368. ISSN: 1098-1004.