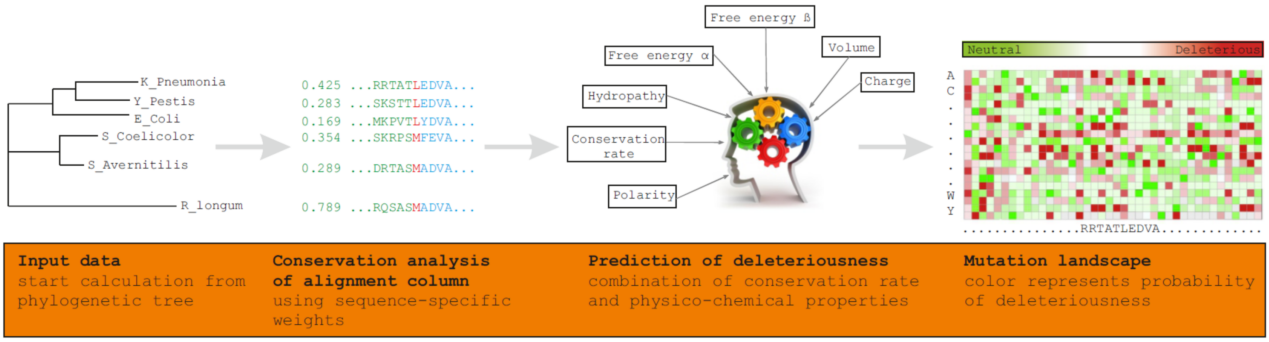
Excel@FIT 2015
http://excel.fit.vutbr.cz

RAPHYD: Predictor of the Effect of Amino Acid Substitutions on Protein Function

Miloš Musil



Abstract
Many genetic mutations are single nucleotide polymorphisms (SNPs), i.e. variations at a single position in a DNA among individuals. Significant number of genetic diseases is caused by non-synonymous SNPs manifested as single point mutations on the protein level. The ability to identify deleterious substitutions could be useful for protein engineering to test whether the proposed mutations do not damage protein function same as for targeting disease causing harmful mutations. However the experimental validation is costly and the need of predictive computation methods has risen. Here we introduce a new *in silico* predictor based on the principles of evolutionary analysis and dissimilarity between original and substituting amino acid physico-chemical properties. Developed algorithm was tested on four datasets with 74,192 mutations from 16,256 sequences in total. The predictor yields up to 72% accuracy and in the comparison with the most existing tools, it is substantially less time consuming. In order to achieve the highest possible efficiency, the optimization process was focused on selection of the most suitable (a) overall decision threshold, (b) third-party software for calculation of a multiple sequence alignment and (c) a set of decision features / physico-chemical properties. To cope with the last mentioned problem, two feature selection methods were tested on the database of 544 possible properties.
Keywords: Amino acid substitutions — Phylogenetic analysis — Mutations — Mutation effect prediction

Supplementary Material: Supplementary materials

*xmusil46@stud.fit.vutbr.cz, *Faculty of Information Technology, Brno University of Technology*

1. Introduction

Non-synonymous single nucleotide polymorphisms (SNPs) can have severe effect on protein functionality. A lot of genetic diseases are caused by single point nucleotide mutation such as cystic fibrosis or sickle cell anemia. The identification of potentially deleterious mutations could also be useful for protein engineering to test whether the proposed mutations do not damage protein function. Since the experimental validation is costly, laborious and time consuming, the application of computational approach is highly desirable.

In recent years, several tools for predicting the ef-

fect of amino acid mutations on protein function were developed. A considerable amount of these prediction tools is based on very complex machine learning methods integrating time demanding calculations of features like secondary structure at the point of mutation or solvent accessibility of original residue. However, the recent study [1] has demonstrated, that the comparable results can be achieved by much faster and more transparent method [2] employing only phylogenetic analysis for calculation of sequence weights from phylogenetic tree and basic amino acid properties. The main aim of this study is to propose and develop a new predictor of deleteriousness of protein mutations utilizing this successful concept. Moreover, optimization in the terms of different multiple-alignment software, p-value[1] thresholds or selected subset of physico-chemical properties will be presented.

2. Methods

2.1 Design of new prediction tool

The design of new prediction tool was inspired by MAPP algorithm, originally developed in 2005 by Stone and Sidow [2]. The prediction core of this tool is based solely on the phylogenetic analysis and differences in physico-chemical properties between wild-type and mutant amino acid. In MAPP algorithm, the following properties were employed: hydropathy [3], polarity [4], charge [4], side-chain volume [5], free energy in α-helical conformation [6] and free energy in β-sheet conformation [6].

The phylogenetic analysis requires the multiple-sequence alignment and phylogenetic tree with branch lengths. To construct these inputs, the need of using third-party software occurred. CLUSTALΩ [7] and FastTree [8] were employed to construct these inputs. BLASTp [9] was applied to choose 200 homologs with e-value 10^{-12} from nr90 database [10]. This database contains amino acid sequences from all species with less than 90% sequence identity to reduce highly similar cases which could bias the predictions. In case of proteins with more than 200 homologs, selection is performed uniformly across whole set of hits meeting the criteria and sorted by e-value. The pipeline and associated threshold for calculation of alignment and tree was inspired by workflow of HotSpot Wizard tool [11] for calculation of conservation of individual residues.

The output of FastTree method is an unrooted tree. Since the algorithm requires the rooted form of phylogenetic tree as an input, the transformation needs to be done. To accomplish such a task, mid-point rooting method was used.

With rooted phylogenetic tree constructed, the impact score of each possible amino acid mutation at each position in a multiple-sequence alignment is given by the following steps:

- Based on the topology and branch lengths of the tree, the weights are calculated for each sequence to obtain the phylogenetic correlation among all sequences. This is performed by Felsenstein's algorithm [12] that calculates the weighted average of the "best weights" obtained by rooting the tree at the midpoint of each branch.
- The sequence weights are multiplied by the relative frequency of each amino acid occurring at analyzed position to obtain "alignment summary" matrix.
- This summary is interpreted using a matrix of physico-chemical properties. Such an interpretation is expressed by multiplication of alignment summary matrix with matrices of physico-chemical properties.
- Constraint violation is measured for each position of the sequence. Dissimilarity scale between original and substituting amino acid in combination with conservation rate (obtained by previous steps) yields the probability that substituting amino acid is neutral.

2.2 Testing datasets

The testing of our tool was performed on four datasets. The MMP dataset consisted of 13 massively mutated proteins, 11 from Yampolsky and Stoltzfus study [13] and two from patent applications issued by Danisco Inc. [14] [15]. The PMD dataset was obtained as the subset of PMD database [16] (last update in 2007) and comprises 1,406 sequences. The BSIFT dataset was taken from the study of Lee at al. [17] and it consists of a diverse set of experimentally described mutagenesis experiments extracted from Swiss-Prot database [18]. PredictSNP dataset [1] was compiled from five different sources containing 10,081 sequences in total. All four datasets together with the number of neutral and deleterious mutations are summarized in Table 1.

2.3 Design of experiments

Performance evaluation was measured on all MMP, PMD, BSIFT and PredictSNP datasets. The multiple sequence alignments and phylogenetic trees were calculated by CLUSTALΩ and FastTree at first. In order to achieve the best possible success rate, the following options were tested:

[1] **p-value:** stands for significance level of the statistical t-test where null hypothesis is "mutation is neutral"

Table 1. Summary of testing datasets.

Datasets	Mutations			Sequences
	Neutral	Deleterious	All	
MMP	7,538	4,456	11,994	13
PMD	1,248	2,249	3,497	1,406
BSIFT	3,081	11,738	14,819	4,036
PredictSNP	24,082	19,800	43,882	10,801
Summary	34,921	43,729	74,192	16,256

- Two different third-party software for construction of multiple-sequence alignment: CLUSTALΩ and MUSCLE.
- Two thresholds of e-value $\{10^{-6}, 10^{-12}\}$ and four thresholds of maximum number of homologs $\{50, 100, 150, 200\}$ to select the most suitable set of homolog sequences by BLASTp.
- Ten decision thresholds of native prediction algorithm responsible for assigning mutations to deleterious / neutral subset; these thresholds were chosen from the interval ⟨0.01, 0.1⟩ with the step of 0.01.
- Two feature selection methods (information and experiment based).

3. Results

3.1 Performance of the new tool

Table 2 shows the merged results measured on PredictSNP, MMP, PMD and BSIFT datasets. These results were obtained with using CLUSTALΩ in configuration for 50, 100, 150 and 200 homologs from BLASTp with e-value 10^{-6} and 10^{-12}. According to the achieved results, all sets of experiments with e-value 10^{-12} showed better normalized accuracy than those with e-value 10^{-6}, approximately about 0.01. The best results were observed for variants with 100 and 200 homologs (0.667 in both cases). The average coverage[2] ranged from 0.704 to 0.788. The reason of low coverage is twofold: (a) for some protein & homology search settings, BLASTp was not able to find any homologs and therefore the alignment & tree could not be calculated and (b) mutations on positions with more than 50% gaps were omitted. The significant influence of exclusion of gap-rich positions is clearly visible in the results as the sets of experiments with e-value threshold 10^{-12} achieved systematically higher coverage, about 0.03 on average, against the experiments with e-value threshold 10^{-6}. It is caused by the fact that less strict e-value threshold brings more distant homologs to alignment and more gaps occur. Similarly, the number of gaps increases together with the alignment size – e.g. for PredictSNP dataset, there were 72% and 80% of gaps in the alignment matrix of 50 and 200 homologues, respectively.

Table 2. Prediction accuracy of developed prediction tool implementation with p-value threshold set on 0.01 and default set of six decision features. Results are merged for PredictSNP, MMP, PMD and BSIFT datasets, using CLUSTALΩ for multiple-sequence alignment.

e-val.		50	100	150	200
1e-6	TN[3]	22,164	21,458	20,621	20,277
	FP[4]	3,670	3,331	3,206	3,152
	TP[5]	13,773	13,274	12,926	12,882
	FN[6]	16,900	16,462	16,108	15,921
	Coverage	0.762	0.735	0.712	0.704
	Accuracy	0.636	0.637	0.635	0.635
	Acc. norm.[7]	0.653	0.656	0.655	0.656
1e-12	TN	22,730	22,038	21,319	20,901
	FP	4,228	4,165	4,057	3,909
	TP	15,218	15,082	14,463	14,332
	FN	16,303	15,560	15,363	14,863
	Coverage	0.788	0.766	0.744	0.728
	Accuracy	0.649	0.653	0.648	0.652
	Acc. norm.	0.663	0.667	0.663	0.667

3.2 Influence of different multiple-sequence alignment tools

Developed tool was extensively tested to find the most suitable third-party sequence alignment software. The comparison was performed between MUSCLE and CLUSTALΩ. An experiment with utilization of MUSCLE for multiple sequence alignment construction was performed, again with the different configurations of BLASTp. Table 3 and Table 4 show a comparison between these two tools on the MMP dataset. The results yields that the MUSCLE achieved higher accuracy, approximately about 0.001 on average. The increase is even higher for the desired configuration (200 homologs, e-value 10^{-12}) where it attained 0.008. However, this negligible increase is at the expense of unacceptably high increase of the time requirements.

[2] **Coverage:** ratio of successfully evaluated mutations

[3] **TN:** number of neutral mutations predicted as neutral

[4] **FP:** number of neutral mutations predicted as deleterious

[5] **TP:** number of deleterious mutations predicted as deleterious

[6] **FN:** number of deleterious mutations predicted as neutral

[7] **Acc. norm.:** is insensitive to the problem of inbalance of the evaluated datasets. Calculated as $[TP/(TP+FN)+TN/(TN+FP)]/2$

Table 3. Performance of developed prediction tool with different integrated multiple-sequence alignment software. New prediction tool was used with default threshold 0.01 and default properties.

Software	Size/eval.	50	100	150	200
CLUSTALΩ	**1e-6**	0.680	0.680	0.678	0.676
	1e-12	0.679	0.702	0.687	0.688
MUSCLE	**1e-6**	0.681	0.680	0.676	0.676
	1e-12	0.695	0.700	0.676	0.696

Table 4. Comparison of alignment software in terms of time requirements. New prediction tool was used with default threshold 0.01.

Software	Size/eval.	50	100	150	200
CLUSTALΩ	**1e-6**	15 s	39 s	41 s	57 s
	1e-12	11 s	30 s	36 s	48 s
MUSCLE	**1e-6**	70 s	253 s	435 s	715 s
	1e-12	60 s	211 s	367 s	606 s

3.3 Influence of different threshold configurations

All previously mentioned results were acquired with the default p-value decision threshold 0.01. However, the analysis of these results in Table 2 revealed large number of false positives. It raised the question whether it is possible to improve the accuracy by increasing of p-value threshold. Figure 1 shows the change in accuracy for PredictSNP dataset with the decision threshold set on 10 different values from the interval ⟨0.01, 0.1⟩.

The obtained results suggest that the higher threshold provides better accuracy than the default threshold of 0.01. More specifically, there is up to 1.6% accuracy growth in MMP dataset with the two maxima around the threshold of 0.05 and 0.08 and up to 2.6% accuracy growth in PredictSNP dataset with the maximum around 0.08. We can conclude that the higher threshold can significantly improve the normalized accuracy of prediction. However, the decision about the most suitable p-value threshold should be adjusted to the intended purpose. In some applications, the value of sensitivity metrics (ratio of correctly recognized deleterious mutations) is more important than the normalized accuracy. The differences between sensitivity and specificity can be observed in Figure 1 (detail informations are available in supplements as Table **5** and **6**).

3.4 Influence of exclusion of gap-rich positions

In the process of performance evaluation, only positions with lower than 50% gaps in the column of alignment were taken into account. This raises a question whether such limitation affects the prediction accuracy. In the MMP dataset, there is just a slight difference in the final results. In PredictSNP dataset, the drop of accuracy is approximately 1% if condition of at least 50% of non-gaps is ignored. In conclusion, the condition of at least 50% non-gaps in the column of multiple-sequence alignment does not have a negative influence on the final accuracy.

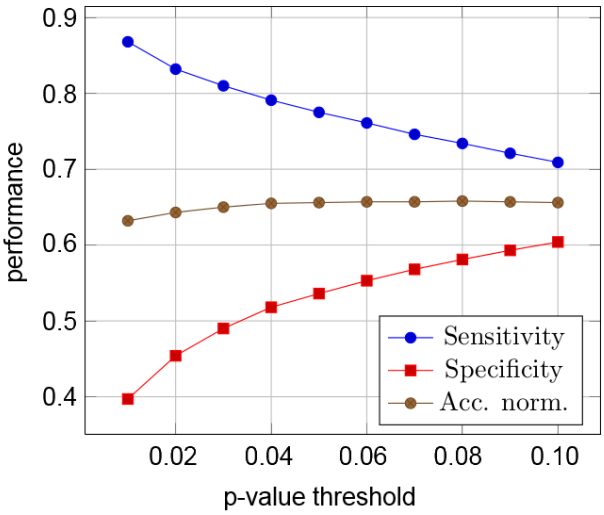


Figure 1. The influence of different p-value threshold in prediction accuracy (measured on PredictSNP dataset).

3.5 Influence of different sets of properties

Up to this point, all experiments were processed with the properties mentioned in 2.1. In this section we are focusing on the question, if the accuracy can be further increased by choosing a different set of physicochemical properties from AAindex database [19]. This is a multidimensional problem as there are 544 highly redundant properties in AAindex database. To cope with this problem, the advanced techniques of feature selection were applied.

First approach was based on the idea of selecting a subset of properties with high orthogonality. To prove this idea, 544 properties from AAIndex were clustered by K-means algorithm [20] into $\{5,6,7,8,9,10\}$ clusters and 20,000 of experiments were proceeded on MMP dataset for each cluster count by random selection of one property from each cluster at one time. From Table 5 it can be seen that the best results were obtained with five clusters generated by K-means algorithm. However the comparison with the implementation utilizing the original set of six expertly chosen properties revealed that this method is inefficient since the obtained accuracy is worse about 0.034 (0.663 versus 0.697).

Table 5. Summary of the best results obtained with $\{5,6,7,8,9,10\}$ clusters (measured on MMP dataset) New prediction tool was used with default threshold 0.01.

Clust. count	5	6	7	8	9	10
Accuracy	0.663	0.651	0.652	0.660	0.658	0.656

Second approach utilized experiment based methods, more specifically the combination of forward selection (FS) and backward elimination (BE). The main disadvantage of FS method lies in its non-recovery factor. In the experiment performed on MMP dataset purely by FS, the algorithm tended to select the feature that had raised the best results independently but have been distortive in the combination with the others. To cope with this problem the following method was designed. At first, the FS process selects 25 properties. This upper bound was set as an acceptable trade-off between the time demands and the size of a feature space for following BE. With the utilization of BE, this subspace of features is analyzed by successive reduction to the number of 5 features. FS is then processed again starting from the subset of properties, where BE obtained the best accuracy. The lower bound of 5 features was determined experimentally as the minimum number providing quality results and it is just slightly lower than usual number of features used in the most of the existing predictors based on machine learning approaches. FS and BE is processed continuously in several iterations.

Described algorithm is computationally costly and calls for massive parallelization, but only after four iterations the selected subset of 13 features has led to improvement approximately 2.4% with decision threshold 0.1 (69.7% up to 72.13% on MMP dataset). Comparison of the RAPHYD algorithm (in both variations - 6 / 13 properties) with existing tools can be observed in Table 6.

Table 6. Comparison of RAPHYD algorithm with existing tools on MMP dataset evaluated in [1]. RAPHYD is presented in two variations - 6 / 13 physico-chemical properties.

Tool	nsSNPAnalyzer	PANTHER	PhD-SNP	PPH-1	PPH-2
TN	4,264	4,336	3,739	4,390	3,518
FP	2,687	834	3,798	3,053	3,925
TP	2,510	329	3,399	3,330	3,769
FN	1,518	1,428	1,058	944	505
Coverage	0.915	0.619	1.000	0.977	0.977
Accuracy	0.617	0.695	0.595	0.659	0.622
Acc. norm.	0.618	0.603	0.629	0.684	0.677

Tool	SIFT	SNAP	PredictSNP	RAPHYD - 6	RAPHYD -13
TN	2,887	5,338	4,291	4,658	5,719
FP	4,463	2,200	3,247	2,541	1,480
TP	3,675	3,163	3,773	3,244	2,820
FN	416	1,293	683	1,104	1,528
Coverage	0.954	1.000	1.000	0.963	0.963
Accuracy	0.574	0.709	0.672	0.684	0.739
Acc. norm.	0.646	0.709	0.708	0.697	0.721

4. Conclusions & Outlook

- The new predictor of the effect of amino acid substitutions on protein function was developed. The decision core is based on the complex phylogenetic analysis and the differences in the physico-chemical properties.
- The extensive evaluation on the four datasets revealed that the best trade-off between normalized accuracy and time consumption was provided by CLUSTALΩ (third-party software for multiple-sequence alignment) launched on maximum 200 homolog sequences from nr90 database found by BLASTp with e-value threshold 10^{-12}.
- The change of default p-value threshold, responsible for assigning the mutant amino acids to potentially neutral and deleterious subsets, led to significantly improved prediction accuracy. The highest improvement of prediction is about 2% for p-value threshold set on 0.08 (tested on both PredictSNP and MMP datasets).
- The exclusion of gap-rich positions having more than 50% gaps in the column of alignment did not exhibit any negative effect on the prediction accuracy.
- The combination of experiment-based methods (forward selection & backward elimination) resulted in the set of 13 attributes with the increase of accuracy about 2.4%.

Acknowledgement: The author thanks colleagues from Loschmidt laboratories for valuable and inspiring consultations, advices and recommendations. Computational resources were provided by the MetaCentrum under the program LM2010005 and the CERIT-SC under the program Centre CERIT Scientific Cloud, part of the Operational Program Research and Development for Innovations, Reg. no. CZ.1.05/3.2.00/08.0144.

References

[1] Bendl J., Stourac J., Salanda O., Pavelka A., Wieben ED., Zendulka J., et al. Predictsnp: Ro-

bust and accurate consensus classifier for prediction of disease-related mutations. *PLoS Comput Biol.*, 10: e1003440., 2014.

[2] Stone EA., Sidow A. Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. *Genome Res.*, (15):978–986, 2005.

[3] Kyte J., Doolittle RF. A simple method for displaying the hydropathic character of a protein. *J Mol Evol.*, (157):105–132, 1982.

[4] Berg JM., Tymoczko JL., Stryer L., Berg JM., Tymoczko JL., Stryer L. *Biochemistry*. W H Freeman, 2002.

[5] Zamyatnin AA. Protein volume in solution. *Prog Biophys Mol Biol.*, (24):107–123, 1972.

[6] Muñoz V., Serrano L. Intrinsic secondary structure propensities of the amino acids, using statistical – matrices: Comparison with experimental scales. *Proteins Struct Funct Bioinforma.*, (20):301–311, 1994.

[7] Sievers F., Wilm A., Dineen D., et al. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Mol Syst Biol.*, (7):539–545, 2011.

[8] Price MN., Dehal PS., Arkin AP. Fasttree 2 – approximately maximum-likelihood trees for large alignments. *Mol Syst Biol.*, 5(e9490), 2010.

[9] Altschul SF., Madden TL., Schäffer AA., Zhang J., Zhang Z., Miller W., et al. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res.*, (25):3389–3402, 1997.

[10] NCBI Resource Coordinators. Database resources of the national center for biotechnology information. *Nucleic Acids Res.*, (41):D8–D20, 2013.

[11] A. Pavelka, E. Chovancova, J. Damborsky. Hotspot wizard: a web server for identification of hot spots in protein engineering. *Nucleic Acids Research*, 43:1–8, 2009.

[12] Felsenstein J. Evolutionary trees from dna sequences: A maximum likelihood approach. *J Mol Evol.*, (17):368–376, 1981.

[13] Yampolsky LY., Stoltzfus A. The exchangeability of amino acids in proteins. *Genetics*, (170):1459–1472, 2005.

[14] Aehle W., Cascao-pereira LG., Estell DA., Goedegebuur F., Kellis Jr. JT., Poulose AJ., et al. Compositions and methods comprising serine protease variants. US Patent 20150031589.

[15] Cuevas WA., Estell DE., Hadi SH., Lee S-K., Ramer SW., et al. Geobacillus stearothermophilus -amylase (amys) variants with improved properties. US Patent US8084240.

[16] Kawabata T., Ota M., Nishikawa K. The protein mutant database. *Nucleic Acids Res.*, (27):355–357, 1999.

[17] Lee W., Zhang Y., Mukhyala K., Lazarus RA., Zhang Z. Bi-directional sift predicts a subset of activating mutations. *PLoS ONE*, 4(e8311), 2009.

[18] Boeckmann B., Bairoch A., Apweiler R., Blatter M-C., Estreicher A., Gasteiger E., et al. The swiss-prot protein knowledgebase and its supplement trembl in 2003. *Nucleic Acids Res.*, (31):365–370, 2003.

[19] Kawashima S., Ogata H., Kanehisa M. Aaindex: Amino acid index database. *Nucleic Acids Res.*, (27):368–369, 1999.

[20] J. MacQueen. *Some methods for classification and analysis of multivariate observations*. University of California, 1967.