Journal of Bioinformatics and Computational Biology
Vol. 6, No. 4 (2008) 709–725
© Imperial College Press

Imperial College Press
www.icpress.co.uk

RECOGNITION OF PROTEIN FUNCTION USING THE LOCAL SIMILARITY

KIRILL ALEXANDROV*, BORIS SOBOLEV†,
DMITRY FILIMONOV‡ and VLADIMIR POROIKOV§

Laboratory for Structure-Function Based Drug Design
Institute of Biomedical Chemistry, Russian Academy of Medical Sciences
Pogodinskaya Str. 10, Moscow 119121, Russia
**dzimmu@yandex.ru*
†boris.sobolev@ibmc.msk.ru
‡dmitry.filimonov@ibmc.msk.ru
§vladimir.poroikov@ibmc.msk.ru

Received 15 October 2007
Revised 14 December 2007
Accepted 19 January 2008

The functional annotation of amino acid sequences is one of the most important problems in bioinformatics. Different programs have been successfully applied for recognition of some functional classes; nevertheless, many functional groups still cannot be predicted with the required accuracy. We developed a new method for protein function recognition using the original approach of sequence description. Each sequence of the training set is compared with the query sequence, and the local similarity scores are calculated for the query sequence positions and used as input data for the original classifier. The method was tested using leave-one-out cross-validation for three data sets covering 58 enzyme classes. Two tested sets including noncrossing functional classes were recognized with high accuracy at various levels of classification hierarchy. The majority of these classes were predicted with 100% accuracy, showing a prediction ability comparable with the HMMer method and an accuracy superior to the SVM-Prot program. When the tested set was composed of intersected classes of ligand specificity, the prediction accuracy was less; however, the accuracy increased as the size of the predicted class expanded. The proposed method can be used for both predicting protein functional class and selecting the functionally significant sites in a sequence.

Keywords: Functional annotation of proteins; sequence similarity; machine learning; recognition of functional classes.

1. Introduction

Functional annotation of newly sequenced genes presents one of the most important challenges in bioinformatics. Since only a small part of encoded proteins is characterized experimentally, the methods of computational functional classification of new amino acid sequences are being intensively developed. Homology-derived annotation based on pairwise sequence alignment was a general way to predict protein function for a long time. This approach is shown to have certain limitations.[1,2]

Phylogenetic methods combined with the established experimental annotations (so-called "phylogenomics") reveal significant advantages in the characterization of large, functionally diverged protein families, but these methods usually require the intensive interference of the expert in annotation procedures.[3] In contrast, the machine learning approach underlies automated functional annotation based on the training set of experimentally annotated proteins. The following methods have been used for protein function prediction: naive Bayes classifier, artificial neural network, k-nearest neighbor, decision tree, and support vector machine.[4] These methods enable to avoid the sequence alignment procedure by using different values representing the protein sequences, such as amino acid composition; dipeptide, tripeptide, and tetrapeptide compositions; and descriptors showing the distribution of amino acid residues within the sequence.[4–6]

Machine learning methods show high accuracy of the functional class recognition — exceeding 95% for certain functional classes. However, many functional groups cannot be predicted with reasonable accuracy.[7] Thus, the problem of protein function prediction is far from the final solution.

Earlier, to predict the functional classes of proteins based on their amino acid sequences, we used the original classification algorithm PASS. The sequences were represented by structural Multilevel Neighborhoods of Atom (MNA) descriptors.[8] This approach provides high accuracy of prediction, but the use of MNA descriptors requires significant computational resources.

In this study, we propose a new method of sequence representation that enables to represent a query sequence in terms of local similarity with the proteins of the training set. Similarity scores calculated for all amino acid positions are the input data for the classifier program. In the new approach, one can adapt the program to define the protein features associated with single or multiple sequence regions of different lengths. The simplicity and high computational speed allow an automated search for optimal parameters. The functional annotation also includes the detection of functionally significant amino acid residues. This task can be solved using the sequences, alignments, and phylogenetic trees as input data.[9,10] As the suggested procedure directly assigns similarity scores to the amino acid positions in the query sequence, it is easy to select functionally significant positions and obtain functional maps of proteins without construction of the alignment. The new method is named Projections of Amino Acid Sequences (PAAS).

2. Method

2.1. *Local sequence similarity*

We have suggested the description of the amino acid sequence A by its local similarity to a sequence B. At first, raw similarity scores are detected by shifting the sequences A and B each to the others (Fig. 1). Each region of sequence A is compared with each superposed region of sequence B. The raw score is calculated as the sum of scores determined for each pair of superposed residues.

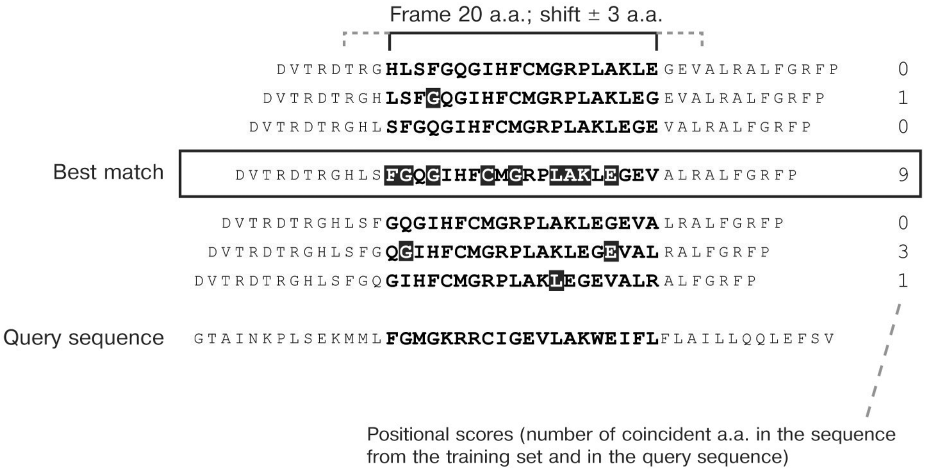

Frame 20 a.a.; shift ± 3 a.a.
DVTRDTRG HLSFGQGIHFCMGRPLAKLE GEVALRALFGRFP 0
DVTRDTRGH LSFGQGIHFCMGRPLAKLEG EVALRALFGRFP 1
DVTRDTRGHL SFGQGIHFCMGRPLAKLEGE VALRALFGRFP 0
Best match
DVTRDTRGHLS FGQGIHFCMGRPLAKLEGEV ALRALFGRFP 9
DVTRDTRGHLSF GQGIHFCMGRPLAKLEGEVA LRALFGRFP 0
DVTRDTRGHLSFG QGIHFCMGRPLAKLEGEVAL RALFGRFP 3
DVTRDTRGHLSFGQ GIHFCMGRPLAKLEGEVALR ALFGRFP 1
Query sequence
GTAINKPLSEKMML FGMGKRRCIGEVLAKWEIFL FLAILLQQLEFSV
Positional scores (number of coincident a.a. in the sequence from the training set and in the query sequence)

Fig. 1. Calculation of raw similarity scores. The training set sequence (upper strings) is compared with the query sequence (lower string) at different shifts (from −3 to 3 amino acid residues), and the maximal scores calculated at the given frame are detected for the query sequence positions. The black boxes denote the amino acid matches.

Thus, the raw similarity scores are calculated as

$$R_k = \max_j (I_{k+F-1,j} - I_{kj}), \quad I_{kj} = \sum_{i=1}^{k} s(a_i, b_{i+j}), \quad M_\mathrm{l} \le j \le M_\mathrm{r}, \tag{1}$$

where R_k is the raw similarity score in position k of sequence A; F is the length of the comparable sequence fragments or a "frame"; $s(a_i,\, b_{i+j})$ is the similarity of amino acids a_i and b_{i+j}; j is the current shift; and M_l and M_r are the maximal allowable shifts at the left and right sequence edges, respectively. In Fig. 1, $F = 20$, $M_\mathrm{l} = -3$, $M_\mathrm{r} = 3$, and $R_k = 9$ for the left edge position in the frame "FGMGK..." of the query sequence.

To run the local similarity calculation, the following parameters should be defined: the frame and the maximal band ($\max B$). The second value specifies M_l and M_r, depending on the relation between the lengths of query and training sequences (L_q and L_t, respectively).

$$\begin{aligned} M_\mathrm{l} &= \begin{cases} \max B/2 | L_\mathrm{q} \le L_\mathrm{t} \\ -\max B/2 - L_\mathrm{q} + L_\mathrm{t} | L_\mathrm{q} > L_\mathrm{t} \end{cases} \\ M_\mathrm{r} &= \begin{cases} -\max B/2 | L_\mathrm{q} \ge L_\mathrm{t} \\ -\max B/2 + L_\mathrm{t} - L_\mathrm{q} | L_\mathrm{q} < L_\mathrm{t} \end{cases} \end{aligned} \tag{2}$$

In this study, we used the simplest measure of similarity between the amino acid residues: 1 for identical residues and 0 for different ones. It was shown that the use of substitution matrices does not increase the accuracy of prediction. The local score $S_k(A, B)$ calculated for position k of sequence A is estimated by the maximum of the values R_k calculated for all frames in which this position put in:

$$S_k(A, B) = \max_m R_{k+m}, -(F-1) \le m \le F-1. \quad (3)$$

The query sequence A is compared to each sequence B of the training set. Thus, we obtain the local similarity scores of the query sequence with all training set sequences.

The suggested procedure is similar to the well-known dot-matrix method.[11] Sequence alignment can be considered as the joining of diagonal fragments, which fit into the narrow band providing the best alignment score.[12]

In PAAS, local similarity scores are used as the input data for the classifier program. Note that many positions, which could be ignored in alignment, are accounted for in this procedure. The number of scores representing each training set sequence equals the length of the query sequence. If the scores calculated for proteins belonging to a certain class are averaged for each query sequence position, one obtains the class projection on the query sequence (Fig. 2). The averaged local similarity scores obviously represent the motifs that determine the functional similarity or difference between the studied proteins.

2.2. *Classification algorithm*

We adopted an algorithm that was originally proposed for the prediction of biological activity spectra for chemical substances (PASS) based on a well-known naive

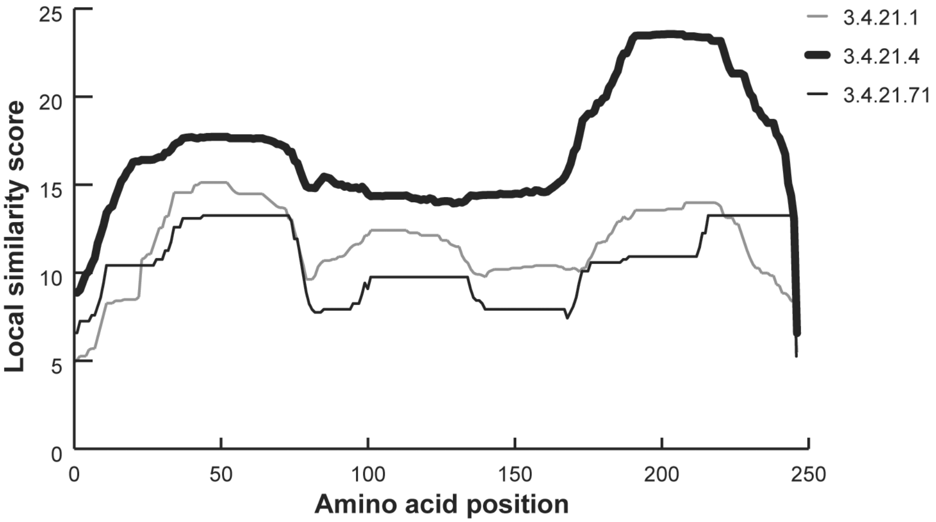


Fig. 2. Projecting the training set sequences on the query sequence. The local similarity scores are averaged over the classes, whose EC numbers are shown in the top right corner. The native class of the query protein is shown in bold.

Bayes classifier.[13] It is assumed that amino acid sequence is described by a set of descriptors $\{D_1, \ldots, D_n\}$, and the probability of its belonging to a given class A is estimated by the conditional probability $P(A|D_1, \ldots, D_n)$. As follows from Bayes theorem,

$$P(A|D_1, \ldots, D_n) = \frac{P(D_1, \ldots, D_n|A)P(A)}{P(D_1, \ldots, D_n)}, \tag{4}$$

where $P(D_1, \ldots, D_n|A)$ is the conditional probability of the descriptor set $\{D_1, \ldots, D_n\}$ occurrence in a sequence from class A, $P(A)$ is the class A prior probability, and $P(D_1, \ldots, D_n)$ is the descriptor set $\{D_1, \ldots, D_n\}$ prior probability. According to the naive Bayes approach, classification features are independent. Therefore,

$$P(D_1, \ldots, D_n|A) \cong P(D_1|A)P(D_2|A)P(D_3|A) \cdots P(D_n|A) = \Pi_i P(D_i|A). \tag{5}$$

As a result, the log-likelihood ratio of the conditional probability $P(A|D_1, \ldots, D_n)$ of class A to $P(\neg A|D_1, \ldots, D_n)$ of its complement $\neg A$ can be expressed as

$$\ln\left[\frac{P(A|D_1, \ldots, D_n)}{P(\neg A|D_1, \ldots, D_n)}\right] = \ln\left[\frac{P(A)}{P(\neg A)}\right] + \sum_i \ln\left[\frac{P(D_i|A)}{P(D_i|\neg A)}\right]. \tag{6}$$

Taking into account that $P(\neg A|D_1, \ldots, D_n) = 1 - P(\neg A|D_1, \ldots, D_n)$ and using Bayes' theorem for ratios $P(D_i|A)/P(D_i|\neg A)$, we find

$$\begin{aligned}\ln\left[\frac{P(A|D_1, \ldots, D_n)}{1 - P(A|D_1, \ldots, D_n)}\right] &= \ln\left[\frac{P(A)}{1 - P(A)}\right] \\ &\quad + \sum_i \left\{\ln\left[\frac{P(A|D_i)}{1 - P(A|D_i)}\right] - \ln\left[\frac{P(A)}{1 - P(A)}\right]\right\}.\end{aligned} \tag{7}$$

The use of the naive Bayes approach faces several problems. As it is known, the logarithm of the probability ratio tends to $\pm\infty$. We substituted the logarithms of probability ratios $\ln[P(A|D_i)/(1 - P(A|D_i))]$ for $\text{ArcSin}(2P(A|D_i) - 1)$. The $\text{ArcSin}(2P(A|D_i) - 1)$ shape coincides with the shape of $\ln[P(A|D_i)/(1 - P(A|D_i))]$ for almost all values of $P(A|D_i)$, but the $\text{ArcSin}(2P(A|D_i) - 1)$ value is bounded by $\pm\pi/2$. To take into account the interdependencies of similarity scores, we use the averaging of arcsine values instead of the sum of the logarithmic values.

In this study, the query protein belonging to class A is estimated by the B-statistic, calculated as follows:

$$t_0 = \frac{\displaystyle\sum_{k=1}^{N} [W_k(A) - W_k(\neg A)]}{\displaystyle\sum_{k=1}^{N} [W_k(A) + W_k(\neg A)]} \tag{8}$$

$$t_i = \frac{\sum_{k=1}^{N} S_{ik}[W_k(A) - W_k(\neg A)]}{\sum_{k=1}^{N} S_{ik}[W_k(A) + W_k(\neg A)]} \tag{9}$$

$$t = \mathrm{Sin}\left[\frac{1}{n}\sum_{i=1}^{n} \mathrm{ArcSin}(t_i)\right] \tag{10}$$

$$B = \frac{t - t_0}{1 - tt_0}, \tag{11}$$

where N is the number of sequences in the training set, $W_k(A)$ and $W_k(\neg A)$ are the weights of sequence k in class A and its complement (0 or 1 in this paper), S_{ik} is a similarity score in a position i of the query sequence with the training sequence k, and n is the number of amino acid residues in the query sequence.

The classifier program that we developed allows the assignment of multiple classes to a single sequence. The classes can be intersected simulating, for example, the situation with overlapped substrate specificity of cytochromes P450.

2.3. *Validation of prediction accuracy*

To estimate the accuracy of prediction, we used the leave-one-out cross-validation (LOOCV) procedure. At each step of the LOOCV procedure, one sequence was removed from the training set and used as a query sequence. The obtained B-statistic values were used to calculate the Independent Accuracy of Prediction (IAP) for each class A[13]:

$$IAP = \frac{\sum_{i,j} \theta(B_{i\in A} - B_{j\in\neg A})}{N_A \cdot N_{\neg A}}, \tag{12}$$

where B_i is the estimation of the sequence i belonging to class A if i actually belongs to class A; B_j is the estimation of the sequence j belonging to class A if j actually belongs to its complement $\neg A$; $\theta(x) = 1$ if $x > 0$, $\theta(x) = 1/2$ if $x = 0$, $\theta(x) = 0$ if $x < 0$; N_A is the number of sequences in class A; and $N_{\neg A}$ is the number of sequences in its complement $\neg A$. This criterion is defined as “independent” because it does not depend on any additional assumptions concerning the parent population and risk function.

If all B-statistic values calculated for the sequences belonging to class A exceed the values calculated for the sequences not belonging to class A, than IAP equals one. We consider the class recognition accuracy as 100% if the IAP value is exactly equal to one.

3. Testing the Prediction

The accuracy of the method was evaluated on three protein collections representing the serine protease set, the "gold standard", and the cytochrome P450 superfamily.

3.1. *Serine proteases*

At first, the method was tested on the serine proteases (Enzyme Classification Number 3.4.21.X). Proteins of this class are known to be successfully recognized by the phylogenomic approach.[14] Due to this reason, 623 amino acid sequences composing 28 classes were selected to test the efficiency of our method. These classes were defined by the fourth value of the EC Number and therefore did not intersect each to the others. As can be seen from Fig. 3, the prediction accuracy increased with the extension of the frame and a maximal band parameter. The average accuracy reached the maximum (close to one) at the maximal band of 50 and frame of 50. Twenty-four of 28 classes were recognized at these parameter values with 100% accuracy.

3.2. *Gold standard*

At the next stage, we tested the method vs. the so-called "gold standard". The gold standard is the training set especially designed for testing of the functional annotation methods. It represents experimentally characterized proteins as well as their very close homologs.[15] The gold standard sequences belong to five enzyme superfamilies — amidohydrolases, crotonases, enolases, haloacid dehalogenases, and proteins forming vicinal oxygen chelates. These superfamilies are divided into 98 families. Forty-two families represented by the single sequences were excluded from

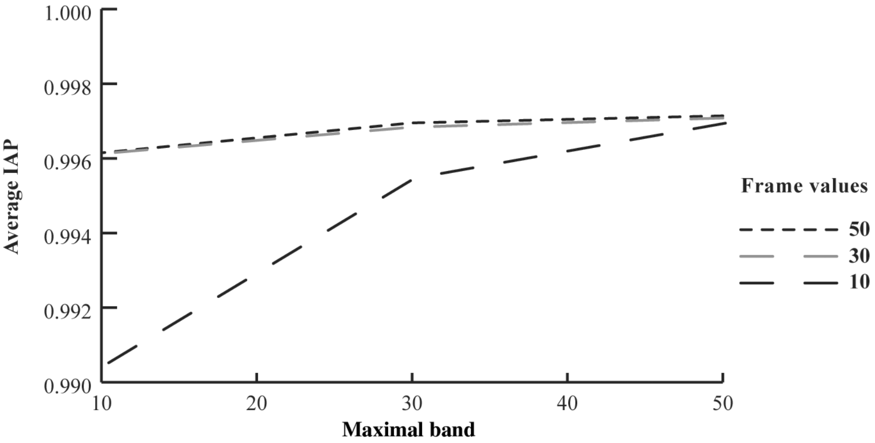


Fig. 3. Testing of the method vs. the serine proteases. Dependence of the recognition accuracy on parameter values.

the original set, so the used evaluation set contained 832 sequences. Thus, the training set consisted of nonintersecting protein classes. Each protein belongs to a single family and a single superfamily. LOOCV testing of our program included two tasks: classification by families and superfamilies.

High accuracy of recognition was obtained at both superfamily and family levels. The average IAP values calculated for the majority of parameter pairs (maximal band/frame) exceeded 0.99. Given the most favorable parameter values, 4 superfamilies were recognized with 100% accuracy and 1 family was recognized with IAP = 0.9996, while 45 families were recognized with IAP = 1 and 11 families were recognized with IAP > 0.96 (Table 1).

Table 1. Accuracy of recognition of gold standard families obtained using the LOOCV test (adopted from Brown *et al.*[15]).

Family	EC number[a]	IAP
1,2-dihydroxynaphthalene dioxygenase	n/a	0.9996
1,4-dihydroxy-2-napthoyl-CoA synthase	n/a	1
2,2',3-trihydroxybiphenyl dioxygenase	n/a	0.9994
2,3-dihydroxybiphenyl dioxygenase	1.13.11.39	0.9929
2,3-dihydroxy-p-cumate-3,4-dioxygenase	n/a	1
2,4,5-trihydroxytoluene oxygenase	n/a	1
2,6-dichlorohydroquinone dioxygenase	n/a	1
2-haloacid dehalogenase	3.8.1.2	1
3,4-dihydroxy-phenylacetate 2,3-dioxygenase	1.13.11.15	1
3-hydroxyisobutyryl-CoA hydrolase	3.1.2.4	1
3-isopropylcatechol-2,3-dioxygenase	n/a	0.9988
3-methylcatechol 2,3-dioxygenase	n/a	0.9587
4-hydroxyphenylpyruvate dioxygenase	1.13.11.27	1
adenosine deaminase	3.5.4.4	1
allantoinase	3.5.2.5	1
ammelide aminohydrolase	n/a	1
AMP deaminase	3.5.4.6	1
aryldialkylphosphatase	3.1.8.1	1
catechol 2,3-dioxygenase	1.13.11.2	0.9979
chloromuconate cycloisomerase	5.5.1.7	0.9969
crotonobetainyl-CoA hydratase	n/a	1
cytosine deaminase	3.5.4.1	1
delta(3,5)-delta(2,4)-dienoyl-CoA isomerase	n/a	1
deoxy-d-mannose-octulosonate 8-phosphate phosphatase	3.1.3.45	1
d-hydantoinase	3.5.2.2	1
dihydroorotase1	3.5.2.3	1
dihydroorotase2	3.5.2.3	1
dihydroorotase3	3.5.2.3	1
dipeptide epimerase	n/a	0.9740
dodecenoyl-CoA delta-isomerase (mit.)	5.3.3.8	1
enolase	4.2.1.11	1
enoyl-CoA hydratase	4.2.1.17	0.9999
epoxide hydrolase *n*-terminal phosphatase	n/a	1
feruloyl-CoA hydratase/lyase	n/a	1
fosfomycin resistance protein FosA	2.5.1.18	1

Table 1. (*Continued*)

Family	EC number[a]	IAP
galactonate dehydratase	4.2.1.6	1
glucarate dehydratase	4.2.1.40	1
glyoxalase I	4.4.1.5	1
histone acetyltransferase	2.3.1.48	1
guanine deaminase	3.5.4.3	1
isoaspartyl dipeptidase	n/a	1
l-hydantoinase	3.5.2.2	1
mandelate racemase	5.1.2.2	1
methylaspartate ammonia-lyase	4.3.1.2	1
methylglutaconyl-CoA hydratase	4.2.1.18	1
methylmalonyl-CoA epimerase	5.1.99.1	1
muconate cycloisomerase	5.5.1.1	0.9993
n-acetylgalactosamine-6phosphate deacetylase	n/a	1
n-acyl-d-amino-acid deacylase	3.5.1.81	1
o-succinylbenzoate synthase	n/a	0.9969
phosphonoacetaldehyde hydrolase	3.11.1.1	1
phosphoserine phosphatase	3.1.3.3	1

[a] n/a designates that an EC number is not assigned to this class.

Table 2. Testing the method vs. the gold standard set grouped by families: averaged IAP values.

		Frame				
		10	20	30	40	50
Maximal band	10	0.54434	0.99395	0.99673	0.99708	0.99774
	20	0.99865	0.99868	0.99870	0.99863	0.99860
	30	0.99862	0.99864	0.99660	0.99848	0.99848
	40	0.99842	0.99849	0.99847	0.99838	0.99833
	50	0.99852	0.99854	0.99852	0.99847	0.99847

Table 3. Testing the method vs. the gold standard set grouped by superfamilies: averaged IAP values.

		Frame				
		10	20	30	40	50
Maximal band	10	0.97193	0.98471	0.99193	0.99508	0.99652
	20	0.99598	0.99849	0.99945	0.99979	0.99987
	30	0.99887	0.99959	0.99988	0.99995	0.99997
	40	0.99952	0.99985	0.99996	0.99998	0.99999
	50	0.99971	0.99987	0.99995	0.99997	0.99997

The accuracy of superfamily prediction reached the highest values at the maximal frame, while the accuracy of family prediction reached the maximal values at the less frames (Tables 2 and 3; Figs. 4 and 5). Superfamilies are better recognized at the frame of 50, while accuracy of the family recognition reached the maximum at the frame of 30. The superfamilies seem to be clearly recognized

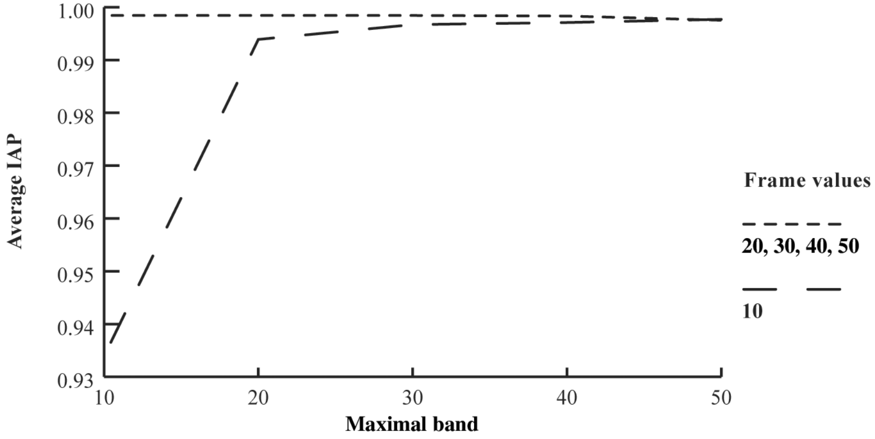


Fig. 4. Testing of the method vs. the gold standard set grouped by families. Dependence of recognition accuracy on parameter values. Curves for some frame values look very similar and are given as a single curve.

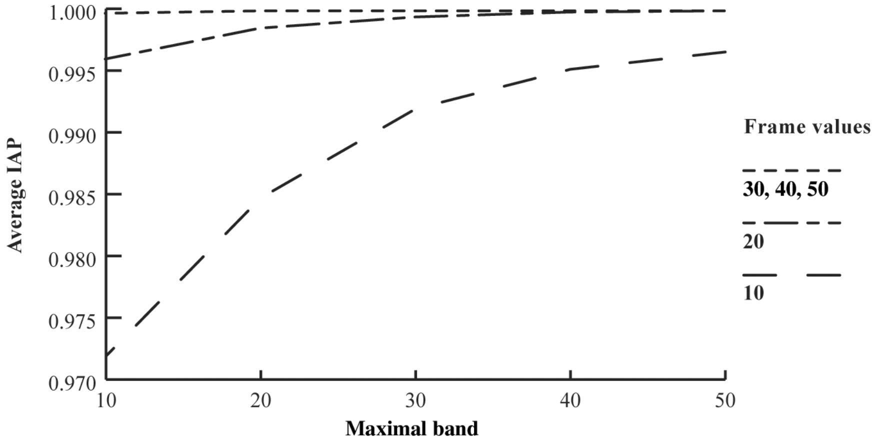


Fig. 5. Testing of the method vs. the gold standard set grouped by superfamilies. Dependence of recognition accuracy on parameter values. Curves for some frame values look very similar and are given as a single curve.

by alignment-based methods; however, families of the same superfamily are worse recognized by the analysis of aligned sequences with phylogenomics methods.[4] Our approach enables to successfully recognize different functional classes belonging to the same superfamily.

3.3. *Cytochrome P450 superfamily*

The cytochrome P450 superfamily represents quite a challenging task. Many of the P450 members are characterized by the wide ligand spectrum, and the subclasses of ligand specificity are intersected. As it was shown in our previous study, strong correspondence between the homology of proteins and the similarity of their ligands was not found for several P450 families.[16] In any case, clustering the aligned sequences is not an absolutely reliable way to annotate P450 sequences. We tested our program to understand whether the local similarity features, which allow the recognition of ligand specificity, exist. The data sets of experimentally annotated P450 proteins were retrieved from the Cytochrome Protein Database (http://cpd.ibmh.msk.su/).[17] We collected two training sets representing specificity subclasses composed of two or more proteins. The first set represented 211 proteins, metabolizing 578 substrates. The second set represented 139 proteins, induced by 272 compounds. The proteins specifically interacting with the same ligand were considered as belonging to the same subclass of substrate or inducer specificity.

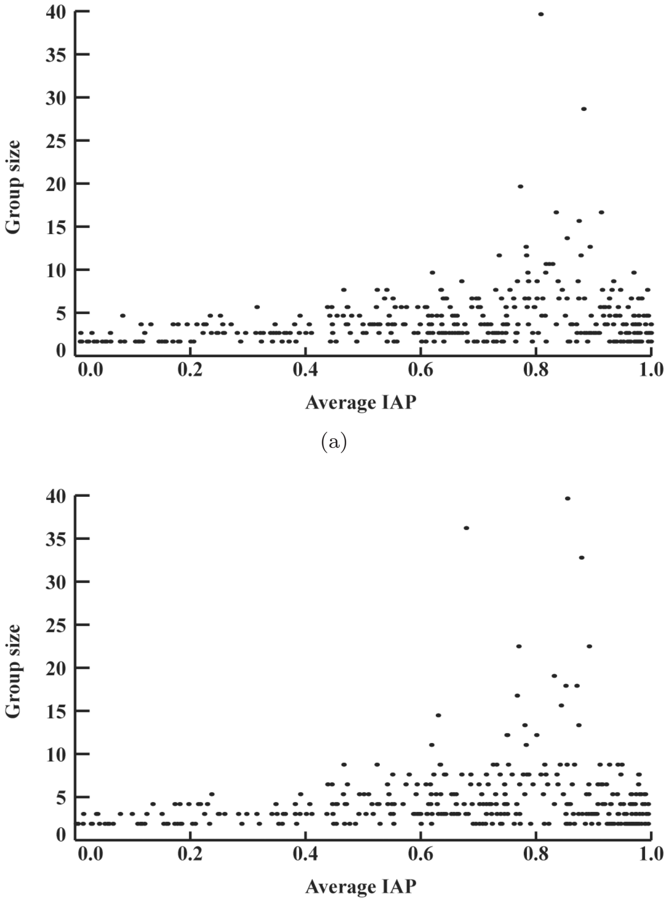
The LOOCV procedure showed that prediction accuracy is significantly less than in the case of other studied sets (Fig. 6). However, we can see a clear trend of increasing accuracy with an increasing size of functional group for both the substrate and inducer specificities.

Though the ligand specificity of cytochrome P450 does not reveal clear correlation with homology,[16] P450 families are perfectly separated by alignment. Therefore, the different methods predict P450 ligand specificity with relatively low accuracy. PAAS recognizes the P450 substrate and inducer specificities with low efficiency too. On the other hand, the larger subclass sets provide a relatively higher accuracy obtained with our method due to the larger representation of different families in the P450 set. The possible contribution of certain local motifs on specificity recognition is the subject of further study.

3.4. *Comparison with other methods*

In order to estimate the comparative power of the suggested method, we performed functional class prediction with two existing methods: the HMMer program, which uses hidden Markov models (HMMs) based on the sequence alignments[18]; and the SVM-Prot program, which implements the machine learning algorithm using the unaligned sequences.[19] The comparative results are shown in Table 4.

We extracted the evaluation set from the gold standard. It represented all superfamilies (designated by "sf" in Table 4) and 10 families (designated as "f" in Table 4). The families were selected so that they were rather large (from 7 to 215 proteins) and had EC numbers (to evaluate the SVM-Prot prediction). Eight families predicted with the highest accuracy (IAP = 1), and two families predicted with the minimal IAP value. The HMMer program was tested by building the models for all evaluated families and scanning each model vs. the evaluation set. The SVM-Prot program was used with its own training set and tested by three sequences belonging to each evaluated family, which used input data for the SVM-Prot server (http://jing.cz3.nus.edu.sg/cgi-bin/svmprot.cgi/).



(a)

(b)

Fig. 6. Testing the prediction of (a) substrate and (b) inducer specificity of cytochromes P450. The results are shown for groups including two or more members. The frame and maximal band values are 20 and 100, respectively.

The superfamilies were predicted by only HMMer and our method because the EC numbers cannot be assigned to these classes.

The output data were formalized for our method and HMMer as follows:

- "+ + +" designates 100% accuracy of prediction; and
- "+ +" designates that our method predicts the corresponding class with $0.9 < IAP < 1$, and that HMMer predicts the same class with the number of true-positive (TP) results higher than the sum of false-positive (FP) and false-negative (FN) ones.

The threshold for HMMer results was determined from the program output data. It was equal to the maximal E-value of the correctly predicted proteins belonging to

Table 4. Accuracy of protein functional class prediction with different methods.

Protein class	PAAS	HMMer	SVM-Prot
Amidohydrolase (sf)	+ + +	+ + +	
Crotonase (sf)	+ + +	+ + +	
Enolase (sf)	+ + +	+ + +	?
VOC (sf)	+ + +	+ + +	
Haloacid dehalogenase (sf)	++	+ + +	
Histone acetiltransferase (f)	+ + +	+ + +	++
Enolase (f)	+ + +	+ + +	+ + +
AMP deaminse (f)	+ + +	+ + +	+ + +
d-hydantoinase (f)	+ + +	+ + +	+
dihydroorotase2 (f)	+ + +	+ + +	—
Guanine deaminase (f)	+ + +	+ + +	++
p-type atpase (f)	+ + +	+ + +	—
Urease (f)	+ + +	+ + +	+ + +
2,3-dihydroxybiphenyl dioxygenase (f)	++	++	+
chloromuconate cycloisomerase (f)	++	+ + +	++

a given family. For four superfamilies, the maximal E-value significantly exceeded the bounds of the threshold proposed by the authors of HMMer.

In the case of SVM-Prot, the number of "+" signs corresponds to the number of correctly annotated sequences of three evaluated ones with the highest P-values. It should be noted that SVM-Prot predicts only two first positions of EC number (the superfamilies are not assigned to any EC numbers).

The sign "—" denotes that the method does not recognize this protein family.

Hidden Markov models (HMM) were built without removal of the query sequences. SVM-Prot uses its own training set, which may also contain the query sequences. So, the accuracy of the recognition performed by these two programs could be somewhat overestimated due to the self-recognition. Keeping in mind this fact, we estimate the prediction results obtained by HMMer and PAAS as comparable. The results obtained with SVM-Prot are less accurate than the prediction performed by PAAS and HMMer.

4. Conclusions

The proposed approach revealed high efficiency in protein function prediction. A high accuracy of prediction was obtained for different levels of protein functional classifications. We showed that our method enables to predict effectively the functional class of proteins when these classes do not intersect with each other. The prediction accuracy is high — up to 100% recognition for the majority of these classes. These results are comparable with data obtained from alignment-based methods. However, the PAAS method has the following advantages:

(1) The PAAS method provides fine tuning (by changing of the band and frame values) of the program for searching both the global and local

sequence similarities. This feature enables to classify and functionally map new sequences.

(2) The PAAS method takes into account more information about sequence similarity than alignment methods. The local similarity scores ignored by alignment can make a contribution to protein class recognition.

(3) The PAAS method runs without a preliminary alignment procedure, which often requires expert interference.

(4) HMM building is a time-consuming procedure, while our relatively simple algorithm provides the sequence recognition at a very high speed. It is especially important for detection of the parameter values optimizing a certain class recognition, as the solution of this task can require multiple recalculations.

The suggested approach provides a more accurate prediction compared to the machine learning method (SVM-Prot). Our method provides accurate distinguishing of the large protein superfamilies as well as functional subclasses related to the same superfamily. So, the method allows perfect recognition at the different levels of structural and functional specificity.

We suggest that our method can be adapted for different types of sequence similarity. The classes associated with global sequence similarity are perfectly predicted by our program. We suggest that related features with separate sequential motifs should also be recognized by this approach. The superfamilies seem to be clearly recognized by alignment-based methods; however, the families of the same superfamily are worse recognized by the analysis of aligned sequences. Our approach enables to recognize different functional classes belonging to the same superfamily. The families are predicted with maximal accuracy at shorter frame values compared to the superfamily level. Thus, the relatively short sequential motifs are more important for recognition of the classified groups, which are closer to each other.

Testing of the method vs. the P450 superfamily reveals a less accurate recognition of broadly intersected functional subclasses within the large group of homological proteins. Sophisticated sequence–function relationships result in the difficulties of function recognition. However, the larger groups were predicted with significantly higher accuracy. It is possible that remote homologs can have three-dimensional structural features that provide affinity to the same ligands, which are not recognized in a sequence.

Our approach can be applied for both functional specificity prediction and sequence mapping, i.e. to reveal local determinants of the functional specificity.

Acknowledgments

This work was supported by the Russian Federation of Basic Research (grant N 04-04-49390-a). We are grateful to A. V. Lisitsa for providing the data on cytochrome P450 substrate and inducer specificity.

References

1. Devos D, Valencia A, Practical limits of function prediction, *Proteins* **41**:98–107, 2000.
2. Devos D, Valencia A, Intrinsic errors in genome annotation, *Trends Genet* **17**(8):429–431, 2001.
3. Sjölander K, Phylogenomic inference of protein molecular function: Advances and challenges, *Bioinformatics* **20**:170–179, 2004.
4. Han L, Cui J, Lin H, Ji Z, Cao Z, Li Y, Chen Y, Recent progresses in the application of machine learning approach for predicting protein functional class independent of sequence similarity, *Proteomics* **6**(14):4023–4037, 2006.
5. Andorf C, Silvescu A, Dobbs D, Honavar V, Learning classifiers for assigning protein sequences to gene ontology (GO) functional families, *Proceedings of the Fifth International Conference on Knowledge Based Computer Systems* (*KBCS*), pp. 256–265, 2004.
6. Saha S, Raghava GP, VICMpred: An SVM-based method for the prediction of functional proteins of Gram-negative bacteria using amino acid patterns and composition, *Genomics Proteomics Bioinformatics* **4**:42–47, 2006.
7. Jensen J, Gupta R, Stærfeldt H-H, Brunak S, Prediction of human protein function according to Gene Ontology categories, *Bioinformatics* **19**:635–642, 2003.
8. Fomenko A, Filimonov D, Sobolev B, Poroikov V, Prediction of protein functional specificity without an alignment, *OMICS* **10**:56–65, 2006.
9. Kalinina OV, Novichkov PS, Mironov AA, Gelfand MS, Rakhmaninova AB, SDPpred: A tool for prediction of amino acid residues that determine differences in functional specificity of homologous proteins, *Nucleic Acids Res* **32**(Web Server issue):W424–W428, 2004.
10. Kalinina OV, Rassel RB, Rakhmaninova AB, Gelfand MS, Computational method for prediction of protein functional sites using specificity determinants, *Mol Biol (Mosk)* **41**(1):151–162, 2007.
11. McLachlan AD, Test for comparing related amino acid sequences: Cytochrome C and cytochrome C551, *J Mol Biol* **61**:409–424, 1971.
12. Barton GJ, Protein sequence alignment and database scanning, in Sternberg MJE (ed.), *Protein Structure Prediction — A Practical Approach*, IRL Press at Oxford University Press, Oxford, 31–64, 1996.
13. Lagunin A, Stepanchikova A, Filimonov D, Poroikov V, PASS: Prediction of activity spectra for biologically active substances, *Bioinformatics* **16**(8):747–748, 2000.
14. Rose T, Di Cera E, Substrate recognition drives the evolution of serine proteases, *J Biol Chem* **277**(22):19243–19246, 2002.
15. Brown SD, Gerlt JA, Seffernick JL, Babbitt PC, A gold standard set of mechanistically diverse enzyme superfamilies, *Genome Biol* **7**(1):R8, 2006.
16. Borodina Y, Lisitsa A, Poroikov V, Filimonov D, Sobolev B, Archakov A, If there exists correspondence between similarity of substrates and protein sequences in cytochrome P450 superfamily?, *Nova Acta Leopold* **87**(329):47–55, 2003.
17. Lisitsa AV, Ponomarenko EA, Gusev SA, Kuznetsova GP, Karuzina II, Lewi P, Archakov AI, Cytochrome P450 knowledgebase: Structure and functionality, *Proceedings of the 14th International Conference on Cytochromes P450: Biophysics and Bioinformatics*, pp. 29–34, 2005.
18. Durbin R, Eddy S, Krogh A, Mitchison G, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, Cambridge, 1998.
19. Cai CZ, Han LY, Ji ZL, Chen X, Chen YZ, SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence, *Nucleic Acids Res* **31**:3692–3697, 2003.

Kirill Alexandrov received his Diploma from the Russian State Medical University of the Russian Academy of Medical Sciences, Moscow, Russia, in 2006. Since then, he has been a Ph.D. student at the Laboratory for Structure-Function Based Drug Design of the Institute of Biomedical Chemistry, Russian Academy of Medical Sciences. Alexandrov's field of interests include protein function recognition and search for functional determinants in amino acid sequences.

Boris Sobolev received his Ph.D. degree at the Ivanovsky Institute of Virology of the Russian Academy of Medical Sciences in 1993. From 1994 to 1995, he was in Bakh Institute of Biochemistry of the Russian Academy of Sciences as a senior scientist. He is currently in the Orechovich Institute of Biomedical Chemistry of the Russian Academy of Medical Sciences, Laboratory for Structure-Function Based Drug Design, as a leading scientist. His research focuses on protein function prediction and search of potential drug targets.

Dmitry Filimonov graduated in Mathematics, Physics, and Biophysics from the Moscow Institute of Physics and Technology. He is a leading scientist at the Laboratory for Structure-Function Based Drug Design of the Institute of Biomedical Chemistry, the Russian Academy of Medical Sciences. He is a member of the Cheminformatics and QSAR Society. Dr. Filimonov's field of research is bioinformatics and computer-aided drug design.

Vladimir Poroikov graduated from the Department of Physics of Moscow State University in 1974, with an M.Sc. in Physics. He has since then received a Professor in Biochemistry (Institute of Biomedical Chemistry, Russian Academy of Medical Sciences, Moscow, 2000), D.Sc. in Pharmacology (National Research Center for Biologically Active Compounds, Staraya Kupavna, Moscow, 1995), and Ph.D. in Biophysics (Department of Biology, Moscow State University, 1981). Deputy Director (Research) in the Institute of Biomedical Chemistry of the Russian Academy of Medical Sciences (IBMC, Moscow) since 1998, he has also been Head of the Laboratory for Structure-Function Based Drug Design in the IBMC since 1995 and Professor of the Medical and Biological Faculty of the Russian State Medical University since 1996. He has been a member of the organizing committee and/or invited speaker of more than 10 international conferences during the past 4 years, and is a co-author of more than 300 published works and 12 non-open published R&D reports of new pharmaceuticals. Poroikov's field of interest is bioinformatics, chemoinformatics, and computer-aided drug design.