Motif Extraction and Protein Classification

Vered Kunik[1] Zach Solan[2] Shimon Edelman[3]

Eytan Ruppin[1] David Horn[2]

[1]School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel
{kunikver,ruppin}@tau.ac.il
[2]School of Physics and Astronomy, Tel Aviv University, Tel Aviv 69978, Israel
{zsolan,horn}@tau.ac.il
[3]Department of Psychology, Cornell University, Ithaca, NY 14853, USA
se37@cornell.edu

Abstract

We introduce an unsupervised method for extracting meaningful motifs from biological sequence data. This *de novo* motif extraction (MEX) algorithm is data driven, finding motifs that are not necessarily over-represented in the entire corpus, yet display over-representation within local contexts. We apply our method to the problem of deriving functional classification of proteins from their sequence information. Applying MEX to amino-acid sequences of a group of enzymes, we obtain a set of motifs that serves as the basis for description of these proteins. This motif-space, derived from sequence data only, is then used as a basis for functional classification by an SVM classifier. Using the set of the oxidoreductase super-family, with about 7000 enzymes, we show that classification based on MEX motifs surpasses that of two other SVM based methods: SVMProt that relies on physical and chemical properties of the protein sequence of amino-acids, and SVM applied to a Smith-Waterman distance matrix. This demonstrates the effectiveness of our MEX algorithm, and the feasibility of sequence-to-function classification.

keywords motif extraction, enzyme classifi cation

Introduction

It is well-known that high sequence similarity guarantees functional similarity of proteins. A recent analysis of enzymes by Tian and Skolnick [14] suggests that 40% pairwise sequence identity can be used as a threshold for safe transferability of the fi rst three digits of the Enzyme Commission (EC) number [1]. Using pairwise sequence similarity, and combining it with the Support Vector Machine (SVM) classifi cation approach [15, 10], Liao and Noble [7] have argued that they obtain a signifi cantly improved remote homology detection relative to existing state-of-the-art algorithms.

[1]The EC number, which is of the form: n1:n2:n3:n4 specifi es the location of the enzyme on a tree of functionalities.

There are two alternative approaches to the task of protein classification that are sequence-based. One relies on general characteristics of the sequence, such as numbers of specific amino-acids within it, as suggested in [6]. A recent variation of this approach replaces the amino-acid sequence with a feature sequence[3, 4], using physical and chemical features such as hydrophobicity, normalized Van der Waals volume, polarity, polarizability, charge, surface tension, secondary structure and solvent accessibility. Cai $et\ al.$ [3, 4] have applied SVM to the resulting vectors, and reported that this SVMProt technique reaches high degrees of accuracy on many enzyme classes, defined in terms of two digits of the EC number.

The other alternative to straightforward sequence similarity is the use of motifs. Appropriately chosen sequence motifs may be expected to reduce noise in the data and to reflect active regions of the protein, thus improving predictability of its function. A protein can then be represented as a 'bag of motifs' [1] (i.e. neglecting their particular order on the linear sequence), or a vector in a space spanned by these motifs. A recent investigation by Ben-Hur and Brutlag [2], based on the eMotif approach [9, 8] has lead to very good results. Starting out with 5911 oxidoreductases, which fit into 129 classes, they base their analysis on 59783 regular-expression emotifs. With appropriate feature selection methods they obtain success rates well over 90% for a variety of classifiers.

Our approach is also motif based. Its novelty is the motif extraction algorithm (MEX) that we employ. Conventional approaches [5] construct motifs in terms of position specific weight matrices, or else use hidden Markov models and Bayesian networks. They all rely on some sort of supervised approach, such as specifying a particular class of proteins for which some statistical significant over-representation is uncovered. MEX extracts motifs from protein sequences without demanding over-representation of such strings of amino-acids in the data set. Instead, our **unsupervised** method relies on over-representation in small contexts, as explained in the next section. Moreover our motifs are specific strings and not position-specific weight matrices or regular expressions. In the application described below, MEX extracts 3165 specific motifs. This low, yet (in our case) effective number may be compared with the 59783 regular-expression emotifs of Ben-Hur and Brutlag [2].

In what follows, we demonstrate that an SVM analysis of oxidoreductase enzymes based on MEX motifs leads to results that are better than SVM based on pairwise sequence similarity. It also outperforms SVMProt on these enzymes, even though the latter is based on physical and chemical properties of the amino-acid sequence. Moreover, it is highly predictive of function, down to the third level of hierarchy of the EC classification.

The Motif Extraction Algorithm (MEX)

ADIOS [12, 13] is an unsupervised method for extraction of syntax from linguistic corpora. Its basic building block extracts significant patterns from a given text. Here we apply this basic motif extraction algorithm (MEX), to the problem of finding sequence-motifs within biological texts. Consider a corpus constructed of many sequences of variable length, each such sequence expressed in terms of an alphabet of finite size N (e.g. N=20 amino-acids in proteins). The N letters form vertices of a graph on which the sequences will be placed as ordered paths. Each sequence defines

such a path over the graph. The number of paths connecting nodes is used to define probabilities:

$$p(e_i) = \frac{\text{number of paths leaving } e_i}{\text{total number of paths}} \tag{1}$$

$$p(e_j|e_i) = \frac{\text{number of paths moving from } e_i \text{ to } e_j}{\text{total number of paths leaving } e_i} \tag{2}$$

and so on, where e_i is a vertex on the graph. In terms of all $p(e_j|e_i)$ the graph defines a Markov model. Moreover, using any path on the graph, to be called henceforth a trial-path, we find a particular instantiation of a variable order Markov model up to order k, where k is the length of the trial-path. For each such trial-path $e_1 e_2 \cdots e_k = (e_1; e_k)$ we define a right-moving probability function, whose value at site $i, j \leq k$ is

$$P_R(e_i; e_j) = p(e_j|e_i e_{i+1} e_{i+2} ... e_{j-1}) = \frac{l(e_i; e_j)}{l(e_i; e_{j-1})} \tag{3}$$

where $l(e_i; e_j)$ is the number of occurrences of sub-paths $(e_i; e_j)$ in the graph. Starting from the other end of the path we define a left-moving probability function

$$P_L(e_j; e_i) = p(e_i|e_{i+1} e_{i+2} ... e_{j-1} e_j) = \frac{l(e_j; e_i)}{l(e_j; e_{i+1})} \tag{4}$$

Examples of such path-dependent probability functions are shown in Fig. 1.

In Fig. 1 we show the type of structure that we expect to find in our graph - the appearance of coherence of paths over some section, or subsequence, of the path. We select this subsequence pattern as a motif because the coherence of paths implies that the same motif appears in several contexts, i.e. in several promoters. The criteria for motif selection are derived from the observation that such motifs lie between local maxima of P_L and P_R signifying the beginning and the end of the motif. To be specific we define the drop in probability functions as

$$D_R(e_i; e_j) = P_R(e_i; e_j)/P_R(e_i; e_{j-1}) \tag{5}$$

$$D_L(e_j; e_i) = P_L(e_j; e_i)/P_L(e_j; e_{i+1}) \tag{6}$$

We introduce a threshold parameter η such that if $D_R(e_i; e_j) < \eta$, the location e_{j-1} will be declared as the end-point of the motif. Similarly, if $D_L(e_j; e_i) < \eta$, e_{i+1} will be declared as the beginning of the motif. Since the relevant probabilities ($P_R(e_i; e_j)$ and $P_L(e_j; e_i)$) were determined by some finite numbers of paths we have to face the problem of low statistics which may lead to erroneous results. Hence we calculate the significance values of both $D_R(e_i; e_j) < \eta$ and $D_L(e_j; e_i) < \eta$ and require that their average be smaller than a parameter $\alpha < 1$. In the following application we will set $\eta = 0.9$ and $\alpha = 0.01$.

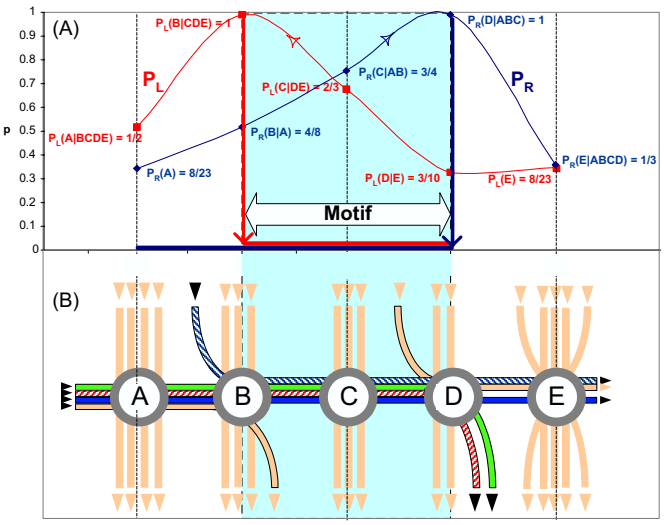

Figure 1: The definition of a motif within the MEX algorithm. Note that the maxima of P_L and P_R define the beginning and the end of the motif. Drops following the maxima signify divergence of paths at these points. The nodes in this figure are labeled by different letters. Note, however, that different letters may also label the same node appearing at different locations on the trial-path. In this example, E and A may represent the same node.

SVM functional classification based on MEX motifs

We evaluated the ability of motifs extracted by MEX to support functional classification of proteins (enzymes). In this experiment, we concentrated on the oxidoreductases superfamily (EC 1.x.x.x). Protein sequences and their EC number annotations were extracted from the SwissProt database Release 40.0. First, MEX was loaded with the 7095 proteins of the oxidoreductases superfamily. Each path in the initial graph thus corresponded to a sequence of amino acids (20 symbols). MEX was run on the data using the parameters $\eta = 0.9$ and $\alpha = 0.01$. The algorithm identified 3165 motifs of different lengths that exist in the enzyme superfamily.

Classification was tested on levels 2 and 3 of the EC on classes that have sufficient numbers of elements to ensure reasonable statistics. Proteins were represented as 'bags of MEX-motifs'. A linear SVM classifier (SVM-Light package, available online at http://svmlight.joachims.org/) was trained on each class separately, taking the proteins of the class as positive examples, and others as negative examples. 75% of the examples were used for training and the remainder for testing. This was repeated on six different random choices of train and test sets to accumulate statistics. We have tested various subsets of MEX motifs and discovered that the subset of motifs longer than five leads to optimal results in the classification analysis. There are 1222 such motifs, spanning the space in which we represent all enzymes and classify them into 16 classes of level 2 and 39 classes of level 3.

We compare our results to two other approaches. One is SVMProt [3, 4], whose SVM procedure we have adopted. Here we use their published results. They have measured performance with

the parameter

$$Q = \frac{TP + TN}{TP + TN + FP + FN}, \tag{7}$$

where TP, TN, FP and FN are the numbers of true positive, true negative, false positive, and false negative outcomes respectively. Since they (and us too) use a relatively large negative set for each class, it is quite easy to get high Q values. We find the Jaccard score

$$J = \frac{TP}{TP + FP + FN} \tag{8}$$

to be more discriminative than Q because J does not include TN in both its numerator and denominator. Hence we adopt this score in the comparisons described below.

The second method we compare ourselves to is based on a one versus all sequence similarity approach. We use the Smith-Waterman [11] algorithm on all pairs of oxidoreductases that were analysed by MEX. We have used the ariadne tool (written by R. Mott, available online at http://www.well.ox.ac.uk/ariadne). The resulting matrix M_{SW} of p-values serves as a distance-matrix to be used as input for an SVM classification. We impose a minimal p-value threshold at 10^{-6}, and then use the logarithms of the p-values to define a normalized distance matrix D_{SW}. This procedure is similar to the approach of [7]. We differ somewhat in the rest of the analysis, by using the full line of D_{SW} as the vector specifying an enzyme, and the same SVM procedure (linear kernel) as employed on the 'bag of MEX motifs'. Learning experiments were conducted by preprocessing the dataset to produce an appropriate input file for the learning task. Each experiment was conducted using a random $75\% : 25\%$ partition of the data into a training set and a testing set, respectively. We have run three experiments for each input file and calculated the average score and standard deviation.

In Fig. 2 we compare the Jaccard scores of MEX with the Smith-Waterman analysis and with SVMProt results (the latter have no errors on them because none were included in the published results). Clearly MEX scores are the best. The average J-scores are 0.89 ± 0.06 for MEX, 0.74 ± 0.13 for SVMProt and 0.79 ± 0.12 for the Smith-Waterman analysis.

An interesting observation from Fig. 2 is that there is no correlation between the size of the class and the success of the tools that we apply. Clearly, if the size is too small, i.e. there are too few enzymes available in the class, there may exist large variance in different trials of train/test divisions, hence large error-bars. However, in general, the average J-values are high, independent of the class that is being tested.

On level 3 of the classification there are no published data of SVMProt. In Fig. 3 we present the comparison between MEX and the Smith-Waterman analysis. MEX has a clear advantage, both in the large class 1.1.1 and overall. The average J-scores are 0.89 ± 0.08 for MEX and 0.78 ± 0.15 for Smith-Waterman. We can thus conclude that the MEX selected motifs carry meaningful information for fine tuned classification of these proteins.

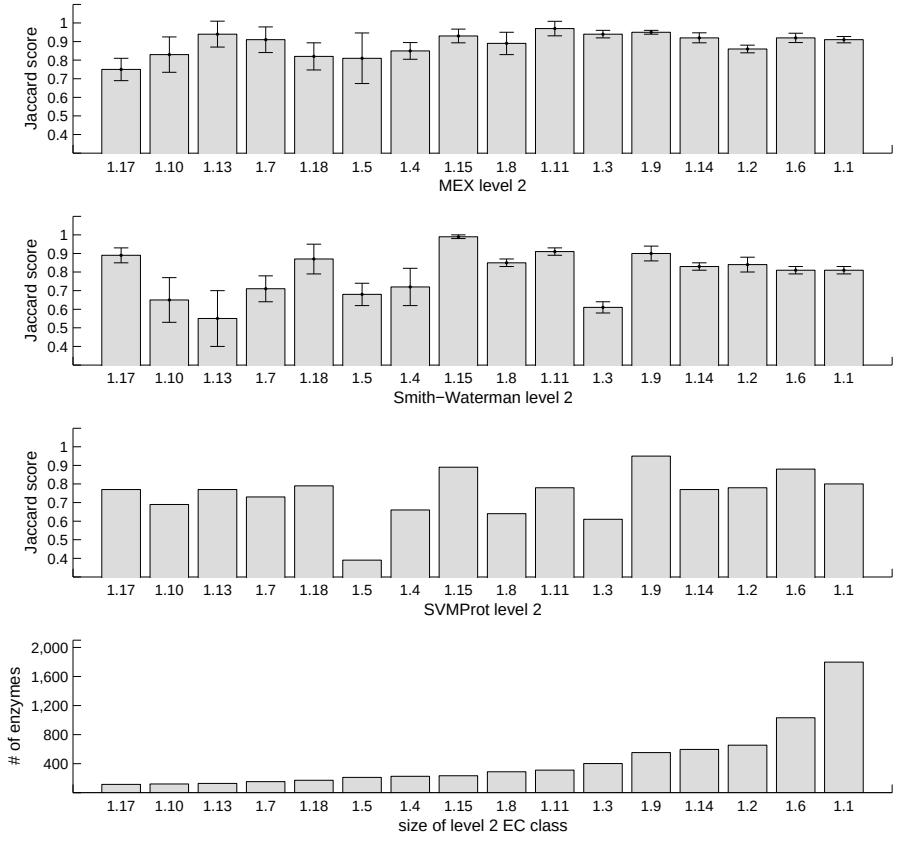

Figure 2: Jaccard scores (see definition in text) for second-level EC classes according to MEX (upper frame) Smith-Waterman (second frame) and SVMProt (third frame). The lowest frame displays the size of each class. The classes are labeled by their EC number and are rank ordered according to size.

class	# of elements	MEX J	SW J
1.1.1	1699	0.91 ± 0.03	0.85 ± 0.04
1.1.99	59	0.92 ± 0.2	0.80 ± 0.11
1.10.2	69	0.94 ± 0.14	0.52 ± 0.00
1.10.3	38	0.78 ± 0.17	0.77 ± 0.11
1.11.1	310	0.98 ± 0.02	0.89 ± 0.01
1.12.99	26	0.92 ± 0.09	0.83 ± 0.00
1.13.11	112	0.90 ± 0.06	0.62 ± 0.08
1.14.11	47	0.87 ± 0.1	0.69 ± 0.10
1.14.13	101	0.82 ± 0.12	0.71 ± 0.12
1.14.14	233	0.93 ± 0.02	0.91 ± 0.07
1.14.15	38	0.91 ± 0.12	0.85 ± 0.13
1.14.16	28	0.93 ± 0.1	0.80 ± 0.08
1.14.19	26	0.89 ± 0.14	0.94 ± 0.10
1.14.99	72	0.89 ± 0.07	0.85 ± 0.09
1.15.1	233	0.92 ± 0.06	0.96 ± 0.00
1.16.1	21	1	0.60 ± 0.20
1.17.4	113	0.86 ± 0.04	0.90 ± 0.02
1.18.1	47	0.77 ± 0.31	0.69 ± 0.14
1.18.6	123	0.88 ± 0.08	0.93 ± 0.03
1.2.1	512	0.88 ± 0.03	0.89 ± 0.03
1.2.4	66	0.83 ± 0.06	0.91 ± 0.03
1.3.1	156	0.84 ± 0.1	0.68 ± 0.03
1.3.3	139	0.96 ± 0.04	0.88 ± 0.05
1.3.5	18	1	1.00 ± 0.00
1.3.99	73	0.76 ± 0.09	0.61 ± 0.09
1.4.1	83	0.86 ± 0.07	0.82 ± 0.03
1.4.3	89	0.93 ± 0.11	0.68 ± 0.07
1.4.99	31	0.92 ± 0.13	0.80 ± 0.08
1.5.1	167	0.67 ± 0.19	0.68 ± 0.10
1.6.1	21	1	0.87 ± 0.12
1.6.2	20	0.81 ± 0.16	0.67 ± 0.12
1.6.5	814	0.87 ± 0.02	0.84 ± 0.01
1.6.99	177	0.70 ± 0.09	0.63 ± 0.04
1.7.1	58	0.91 ± 0.08	0.76 ± 0.15
1.7.2	26	1	0.72 ± 0.10
1.7.99	43	1	0.40 ± 0.20
1.8.1	138	0.91± 0.03	0.86 ± 0.04
1.8.4	137	0.93 ± 0.05	0.88 ± 0.13
1.9.3	552	0.94 ± 0.02	0.90 ± 0.03

Figure 3: Table of J-values derived from MEX and compared with the Smith-Waterman analysis, corresponding to classes at level 3 EC classifi cation.

Motif selection

MEX has extracted motifs of various lengths. We have tested the predictive power of SVM on enzyme classification, along the lines described above, by selecting particular subsets of motifs according to the length of the motif. We discovered that the subset of motifs of length 6 gave excellent results on its own, quite close to those quoted in the previous section, that were derived for all motifs longer than 5.

To understand why this is the case let us ask which of our motifs is unique to only one of the enzyme classes. The statistics of the results are displayed in Fig. 4. As evident, motifs of length 6 are both abundant and, concomitantly, have a very large fraction of motifs unique to only one class. In fact, out of the 601 motifs of length 6, 493 are unique to only one single EC class at the third level. Yet these 493 motifs are not sufficient for the classification task, since their coverage of all proteins within one EC class is limited. In the large class 1.1.1, containing 1699 proteins, the relevant 125 motifs of length 6 that are unique to this class appear on only 63% of the protein sequences. Nonetheless, a classification task based on all length 6 motifs (adding just 108 non-unique motifs) allows us to obtain a Jaccard score of 0.91 ± 0.03 for this class and overall J-values of 0.89 ± 0.06 for level 2 and 0.86 ± 0.13 for level 3. It is only on level 3 that limiting ourselves to motifs of length 6 only led to a lower result than the one quoted in the previous section, which was 0.89 ± 0.08. This is the price we pay for moving from the basis of 1222 motifs longer than 5 to the basis of 601 motifs of length 6 only.

Fig. 4 leads to an interesting insight why, if we base our SVM analysis on all motifs longer than 4, the quality of the results deteriorates: average J-values are lowered to 0.83 ± 0.09 for level 2 and 0.83 ± 0.14 for level 3. The reason must be the large fraction of non-unique motifs of length 5, that harm the quality of the analysis.

Discussion

We have run our MEX algorithm on a group of 7095 enzymes, and the motifs that it has constructed turned out to form an excellent basis for classifying these enzymes into small classes known to have different functional roles. In fact, they serve to define the best classification method from structure to function on the tested enzyme superfamily.

We have compared our results with two approaches:

1. Classification based on pairwise sequence similarity, of the type employed by [7]. We run the system using the same SVM procedure that we employ for MEX. The results demonstrate that MEX derived motifs form a better basis for classification, thus indicating that their selection improves the signal/noise inherent in the original sequences.

2. The SVMProt method of [3, 4]. We have used their published results to draw the comparison on level 2 data. Our results show that we do better, although their method is based on semantic information, i.e. physical and chemical properties of the sequence of amino-acids. This is another indication that the MEX selected motifs carry relevant information.

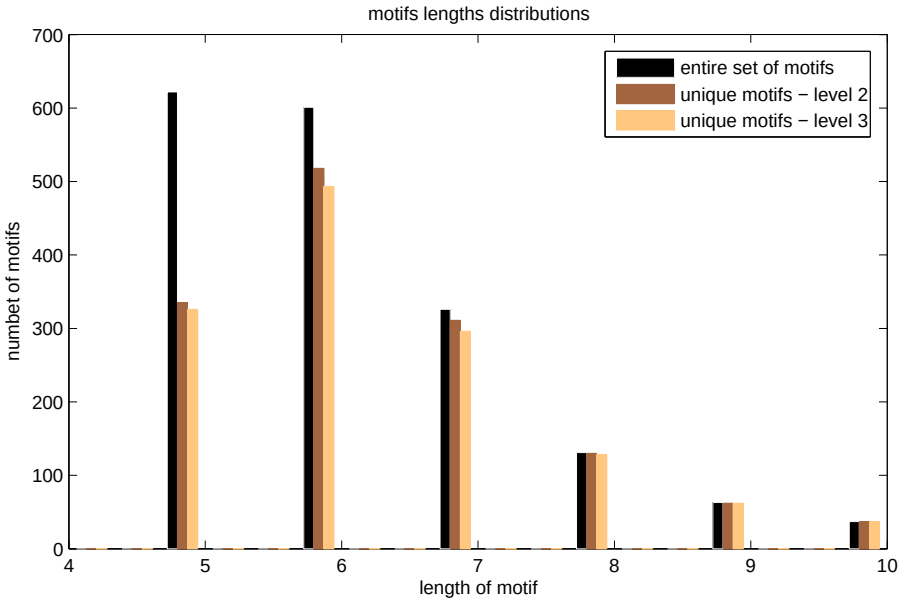

Figure 4: The number of MEX motifs of lengths 5-10. The three sets correspond to (left) all motifs, (middle) unique motifs to single classes at level 2, and (right) unique motifs to single classes at level 3.

It should be noted that the our classification is carried out by using just 1222 motifs of length 6 and over, and similar results can also be obtained from just the 601 motifs of length 6. These are small numbers of features given the fact that they serve to describe 55 classification tasks for about 7000 proteins. All motifs are composed of deterministic consecutive and specific amino-acids. This is different from the regular expressions of eMotifs, the basis of the analysis by [2], or the general approach of probabilistic representations (PSSM) so common in Bioinformatics [5]. This makes the small numbers of MEX motifs even more striking. The deterministic nature of our motifs is a result of the logic used by MEX. This does not contradict the possible applicability of a PSSM structure. The latter could be a natural representation of clusters of MEX motifs that appear to have the same role, but there was no need for it in the analysis of the oxidoreductase enzymes.

The application of the MEX algorithm studied here is a first level of feature extraction from biological sequences. Higher level patterns may be extracted by repeated application of MEX after the observed sequence-motifs are incorporated as vertices in the MEX graph. Moreover, the full extent of the ADIOS approach [13] may lead in the future to revealing higher syntactic structures in biological sequence data.

Acknowledgment

This research was partially supported by the US Israel Binational Science Foundation. ZS has a student fellowship of the Hurwitz Foundation for Complexity Sciences. We thank Asa Ben-Hur

and Doug Brutlag for helpful conversations.

References

[1] Ben-Hur,A., Brutlag, D. (2003) Remote homology detection: a motif based approach. *Bioinformatics*, **19, Suppl. 1**, i26-i33.

[2] Ben-Hur,A., Brutlag, D. (2004) Sequence motifs: highly predictive features of protein function. *Neural Information Processing Systems 2004*.

[3] Cai,C. Z., Han,L. Y., Ji,Z. L., Chen, Y. Z. (2003) SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence. *Nuclear Acids Research*, **31**, 3692-3697.

[4] Cai,C. Z., Han,L. Y., Ji,Z. L., Chen, Y. Z. (2003) Enzyme family classification by support vector machines. *PROTEINS: Structure, Function and Bioinformatics*, **55**, 66-76.

[5] Durbin, R., Eddy, S., Krogh, A., Mitchison, G. (1998). *Biological sequence analysis Probabilistic models of proteins and nucleic acids*, Cambridge University Press.

[6] des Jardin, M., Karp, P. D., Krummenacker, M., Lee, T. J. and Ouzounis, C. A. (1997). Prediction of enzyme classification from protein sequence without the use of sequence similarity. Proceedings of ISMB.

[7] Liao, L., Noble, W. S., (2003) Combining pairwise sequence analysis and support vector machines for detecting remote protein evoluitionary and structural relationships. *J. of Comp. Biology*, **10**, 857-868.

[8] Huang, J. Y., Brutlag, D. L., (2001) The eMOTIF database. *Nuclear Acids research*, **29**, 202-204.

[9] Neville-Manning, C. G., Wu, T. D., Brutlag, D. L., (1998) Highly specific protein sequence motifs for genome analysis. *Proc. Natl. Acad. Sci. USA* **95**, 5865-5871.

[10] Schölkopf, B., (1997) *Support Vector Learning*. R. Oldenburg Verlag, Munich.

[11] Smith, T., Waterman, M., (1981) Identification of common molecular subsequences. *J. of Mol. Biology* **147**, 195-197.

[12] Solan, Z., Ruppin, E., Horn, D., Edelman, S., (2003) Automatic acquisition and efficient representation of syntactic structures. In S. Becker, S. Thrun and K. Obermayer, editors, *Advance in Neural Information Processing Systems* **15**, 91-98, MIT Press, Cambridge, MA .

[13] Solan, Z., Horn, D., Ruppin, E., Edelman, S., (2004) Unsupervised context sensitive language acquisition from a large corpus. In Sebastian Thrun and Lawrence Saul and Bernhard Schölkopf, editors, *Advances in Neural Information Processing Systems* bf 16 MIT Press, Cambridge, MA.

[14] Tian, W., Skolnick, J., (2003) How well is enzyme function conserved as a function of pairwise sequence identity? *J. Mol. Biol.* **333**, 863 - 882.

[15] Vapnik, V. (1995) *The Nature of Statistical Learning Theory*. Springer Verlag, NY.