Motif Extraction and Protein Classification

Vered Kunik
School of Computer Science
Tel Aviv University
Tel Aviv 69978, Israel
kunikver@tau.ac.il

Zach Solan
School of Physics and Astronomy
Tel Aviv University
Tel Aviv 69978, Israel
zsolan@tau.ac.il

Shimon Edelman
Department of Psychology
Cornell University
Ithaca, NY 14853, USA
se37@cornell.edu

Eytan Ruppin
School of Computer Science
Tel Aviv University
Tel Aviv 69978, Israel
ruppin@tau.ac.il

David Horn
School of Physics and Astronomy
Tel Aviv University
Tel Aviv 69978, Israel
horn@tau.ac.il

Abstract

We present a novel 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 data. Applying MEX to the oxidoreductases class of enzymes a relatively small set of motifs is obtained. This set spans a motif-space that is used for functional classification of the enzymes by an SVM classifier. The classification based on MEX motifs surpasses that of two other SVM based methods: SVMProt, a method based on the analysis of physical-chemical properties of a protein generated from its sequence of amino acids, and SVM applied to a Smith-Waterman distances matrix. Our findings demonstrate that the MEX algorithm extracts relevant motifs, supporting a successful sequence-to-function classification.*

keywords motif extraction, enzyme classification

Introduction

It is commonly accepted that high sequence similarity guarantees functional similarity of proteins. A contemporary analysis of enzyme function conservation by Tian and Skolnick [14] suggests that 40% pairwise sequence identity can be used as a threshold to certify functional similarity, i.e. the first three digits of the Enzyme Commission (EC) number are identical [1]. Using pairwise sequence similarity, and combining it with the Support Vector Machine (SVM) classification method [15, 10], Liao and Noble [7] have argued that they obtain a significantly improved remote homology detection relative to existing state-of-the-art algorithms.

There are alternative sequence-based approaches to the task of protein classification. One is based on general characteristics of the sequence, such as the number of specific amino-acids within it, as suggested in [6]. A recent variation of this approach represents the amino-acid sequence as a sequence of physical-chemical features [3, 4], 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 these feature vectors and reported that the SVMProt technique reaches a high degree of accuracy, at a level of two digits of the EC number hierarchy, on various enzyme subclasses.

An alternative to the straightforward sequence similarity approach is the usage of motifs. Appropriately chosen sequence motifs may be expected to reduce noise in the data and indicate active regions of the protein, hence 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 work by Ben-Hur and Brutlag [2], based on the eMOTIF approach [9, 8], led to very good results. Starting out with 5911 enzyme sequences of the oxidoreductases class, which consisted 129

[1] The function of an enzyme is specified by a name and a number given to it by the Enzyme Commission (EC). The EC number consists of four numbers, n1:n2:n3:n4, corresponding to four levels of classification. The oxidoreductases class discussed in this paper corresponds to n1=1, one of the six main divisions. For this class, n2 (subclass) specifies electron donors, n3 (sub-subclass) specifies electron acceptors and n4 indicates the exact enzymatic activity.

EC subclasses, they based their analysis on 59783 regular-expression eMOTIFs. By using an appropriate feature selection method they obtained success rates well over 90% for a variety of classifiers.

The approach presented in this work is motif based. Its novelty is the employed motif extraction algorithm (MEX). Conventional approaches [5] construct motifs in terms of position specific weight matrices, or else use hidden Markov models and Bayesian networks, hence are supervised to some extent. MEX extracts motifs from proteins sequential data in an **unsupervised** manner, without requiring over-representation of its amino-acid motifs in the data set. MEX motifs are explicit strings in contradistinction to position-specific weight matrices or regular expressions. In the application described below, 3165 MEX motifs are extracted. This is a low number of motifs in comparison with the 59783 regular-expression eMOTIFs used by Ben-Hur and Brutlag [2].

In what follows, we demonstrate that an SVM analysis of oxidoreductases enzymes based on MEX motifs leads to results that are comparable to those obtained by an SVM based on pairwise sequence similarity on a level 2 classification tasks and to better results on a level 3 classification tasks. Furthermore, it outperforms the results obtained by the SVMProt method, even though the latter is based on physical and chemical properties of the amino-acid sequence. Moreover, our algorithm is highly predictive of function, down to the third level (sub-subclass) of the EC hierarchy.

The Motif Extraction Algorithm (MEX)

MEX is a motif extraction algorithm that serves as the basic unit of ADIOS [12, 13], an unsupervised method for extraction of syntax from linguistic corpora. We apply it to the task of finding sequence-motifs within biological data. Consider a data set of 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 are placed as ordered paths. Each sequence defines such a path over 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 search-path, we find a particular instantiation of a variable order Markov model up to order k, where k is the length of the search-path. For each such search-path $(e_1;e_k) = e_1 e_2 \cdots e_k$ we define a right-moving probability function, whose value at $i, j \leq k$ is

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

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

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

Fig. 1 demonstrates the type of structures that we expect to find in our graph - an assimilation of paths over a subsequence of the search-path. Such a subsequence is a candidate motif. The criteria for motif selection are defined by local maxima of P_L and P_R signifying, respectively, the beginning and ending of a motif.

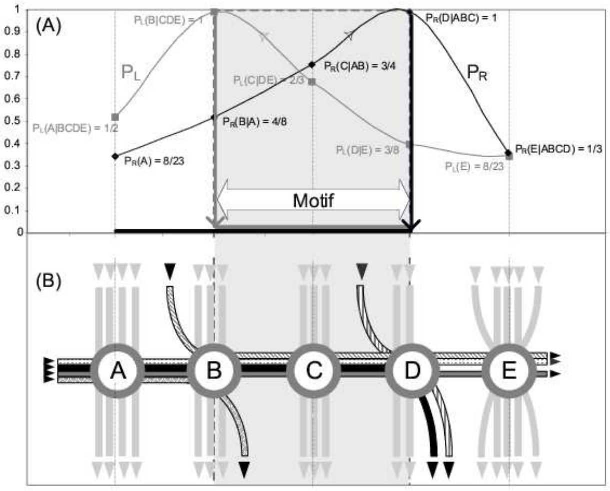

Figure 1. The definition of a motif within the MEX algorithm. Note that the maxima of P_L and P_R defines the beginning and ending of the motif, respectively. Descents in P_L and P_R following the maxima signify divergence of paths.

A drop in the probability functions is defined as:

$$D_R(e_i;e_j) = P_R(e_i;e_j)/P_R(e_i;e_{j-1}) \quad (3)$$

$$D_L(e_j;e_i) = P_L(e_j;e_i)/P_L(e_j;e_{i+1}) \quad (4)$$

The threshold parameter , η, is defined as follows: the location e_{j-1} is declared as the ending of the motif if $D_R(e_i;e_j) < \eta$. Analogously, e_{i+1} is declared as the beginning of the motif if $D_L(e_j;e_i) < \eta$. Since the experimental probabilities, $P_R(e_i;e_j)$ and $P_L(e_j;e_i)$, are determined by finite numbers of paths, a statistical measure is introduced in order to avoid 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 maximum be smaller than a parameter $\alpha < 1$. In the following application we have set $\eta = 0.9$ and $\alpha = 0.01$. Once the algorithm reaches the stop criteria (e.g. ceases to locate new patterns) they are sorted in a length-significance descending order, by which their loci are identified on the original data.

SVM functional classification based on MEX motifs

We have concentrated our analysis on the oxidoreductases class of enzymes. 6602 protein sequences and their EC number annotations were extracted from the SwissProt database Release 40.0. These proteins served as the dataset to which MEX was applied. The enzymes are classified into 16 distinct subclasses of level 2 and 32 distinct sub-subclasses of level 3.

The algorithm identified 3165 motifs of various lengths. These motifs are found on 3739 of the enzyme sequences to which MEX was applied. Classification was tested on levels 2 (subclass) and level 3 (sub-subclass) of the EC number. Subclasses were required to have a sufficient number of elements to ensure reasonable statistics. Protein sequences 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 subclass separately, taking the protein sequences of the subclass as positive examples and the protein sequences of other subclasses as negative examples. 75% of the examples were used for training and the remaining examples for testing. The train-test procedure was repeated on six different random choices of train-test sets in order to accumulate statistics. We have tested various subsets of MEX motifs and discovered that the subset of motifs longer than five amino-acids leads to optimal results in the classification task. There are 1222 such motifs, spanning the space of 3739 enzymes. Enzymes are classified into 16 distinct subclasses of level 2 and 32 distinct sub-subclasses of level 3.

The obtained results are compared to those of two other approaches. The first, SVMProt [3, 4], uses a performance measurement parameter defined as

$$Q = \frac{TP + TN}{TP + TN + FP + FN}, \tag{5}$$

where TP, TN, FP and FN denote the number of true positive, true negative, false positive, and false negative outcomes respectively. The SVMProt results presented below are obtained from their published results. However, since the large negative set used in each classification task quickly yields a high Q value, we have chosen to use the Jaccard score

$$J = \frac{TP}{TP + FP + FN} \tag{6}$$

instead. Not taking into account TN, this performance measurement is more discriminative than Q.

The second approach, the Smith-Waterman algorithm [11], is based on a one-versus-all sequence similarity approach. This algorithm has been applied to the same set of 3739 oxidoreductases sequences represented by MEX motifs. The ariadne tool has been used (written by R. Mott, available online at http://www.well.ox.ac.uk/ariadne) in order to obtain the p-values distances matrix, M_{SW}, defining the feature space of the SVM classifier. A minimal p-value threshold of 10^{-6} was imposed in order to allow usage of p-values logarithm, defining a normalized distances matrix D_{SW}. This procedure is similar to the approach described in [7], however, the entire vector of D_{SW} has been used in our analysis for specifying an enzyme. The classification task has been performed with the same SVM classifier (linear kernel) employed to the data driven by MEX. The dataset has been preprocessed in order to produce an appropriate input file for the learning task. A random $75\% : 25\%$ partition of the data into a training set and a testing set, respectively, has been used for each learning task. The train-test procedure was repeated on six different random choices of data sets in order to accumulate statistics.

Fig. 2 shows a comparison of the Jaccard score obtained by MEX, Smith-Waterman analysis and SVMProt (error deviations are not presented for the latter as they were not included in their published results). The scores obtained by MEX are clearly higher than those obtained by SVMProt and are comparable to those obtained by Smith-Waterman. The average J-scores are 0.89 ± 0.05 for MEX, 0.74 ± 0.13 for SVMProt and 0.89 ± 0.06 for the Smith-Waterman method. Noticeably, there is no correlation between the size of the subclass and the J-scores obtained by both MEX and the Smith-Waterman methods. Clearly, if the size of the subclass is too small, i.e. the number of the positive examples is small, a large variance in the train/test different divisions may exist, resulting in large error deviations. However, in most cases, the average J-scores are high, independent of the tested subclass.

Third level classification results were not compared to SVMProt as none were included in their published results. Table 1 presents a comparison of the Jaccard scores obtained by MEX and Smith-Waterman analysis. The scores obtained by MEX are clearly higher. The average J-scores are 0.89 ± 0.08 for MEX and 0.86 ± 0.15 for Smith-Waterman. These findings attest to the meaningful information embodied in MEX selected motifs, facilitating a fine tuned classification of these proteins.

Motif selection

Motifs of various lengths were extracted by applying the MEX algorithm. The enzyme function classification ca-

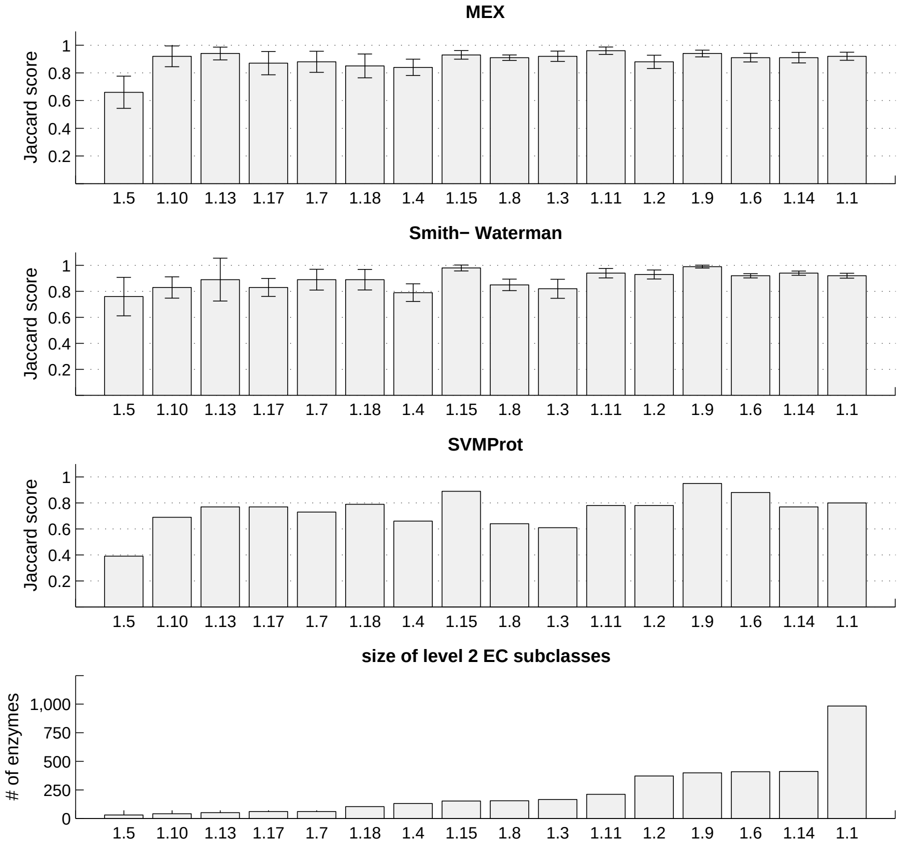

Figure 2. Jaccard scores for second-level EC subclasses obtained by MEX (upper panel), Smith-Waterman (second panel) and SVMProt (third panel). The bottom panel depicts the size of each subclass MEX and the Smith-Waterman method were applied to. The subclasses are labeled according to their EC number and are ordered according to size.

pabilities of the motifs were tested using various length-dependent subsets of motifs. It is important to note that the coverage (i.e., number of enzyme sequences represented by the motifs) varies according to the motifs subset selection. The results demonstrate that the classification task performed by the subset of 1222 motifs of length 6 and longer obtains high J-scores. In order to comprehend the predictive capabilities of MEX motifs, we have analyzed which of these motifs are unique (i.e., belong to a single EC subclass at the second level). Statistics are presented in Fig. 3.

Evidently, motifs of length 6 are both abundant and, concomitantly, comprise a large fraction of motifs unique of a single subclass. Out of the 601 motifs of length 6, 493 are unique. A level 3 classification task performed solely with motifs of length 6 yielded an average J-score of 0.88 ± 0.1, which is essentially as high as the average J-score obtained by classifying the same set of sequences using the 1222 motifs of length 6 and longer. Apparently, the 601 motifs of length 6 serve as an adequate basis for this refined classification task.

An additional interesting insight, clarifying the relatively lower J-scores obtained by an SVM analysis based on the set of MEX motifs longer than 4 (average J-scores are 0.85 ± 0.05 for a level 2 classification task and 0.83 ± 0.09 for a level 3 classification task) is the large fraction of non-unique motifs of length 5 (see Fig. 3) that clearly impairs the predictive power of the unique motifs.

class	# of elements	MEX J	SW J
1.1.1	959	0.91 ± 0.03	0.85 ± 0.04
1.9.3	399	0.92 ± 0.2	0.80 ± 0.11
1.2.1	333	0.94 ± 0.14	0.52 ± 0.00
1.6.5	331	0.78 ± 0.17	0.77 ± 0.11
1.11.1	211	0.98 ± 0.02	0.89 ± 0.01
1.14.14	203	0.92 ± 0.09	0.83 ± 0.00
1.15.1	153	0.90 ± 0.06	0.62 ± 0.08
1.3.3	91	0.87 ± 0.14	0.69 ± 0.10
1.8.4	89	0.81 ± 0.16	0.67 ± 0.12
1.18.6	87	0.82 ± 0.12	0.71 ± 0.12
1.14.13	65	0.93 ± 0.02	0.91 ± 0.07
1.8.1	62	0.91 ± 0.12	0.85 ± 0.13
1.17.4	60	0.93 ± 0.1	0.80 ± 0.08
1.4.1	26	0.89 ± 0.14	0.94 ± 0.10
1.6.99	72	0.89 ± 0.07	0.85 ± 0.09
1.13.11	51	0.92 ± 0.06	0.96 ± 0.00
1.7.1	50	1	0.60 ± 0.20
1.4.3	39	0.86 ± 0.04	0.90 ± 0.02
1.14.99	38	0.77 ± 0.31	0.69 ± 0.14
1.3.1	34	0.88 ± 0.08	0.93 ± 0.03
1.2.4	31	0.88 ± 0.03	0.89 ± 0.03
1.14.15	30	0.83 ± 0.06	0.91 ± 0.03
1.3.99	28	0.84 ± 0.1	0.68 ± 0.03
1.10.3	24	0.96 ± 0.04	0.88 ± 0.05
1.14.19	24	1	1
1.5.1	23	0.76 ± 0.09	0.61 ± 0.09
1.14.11	22	0.86 ± 0.07	0.82 ± 0.03
1.14.16	22	0.93 ± 0.11	0.68 ± 0.07
1.6.2	18	0.92 ± 0.13	0.80 ± 0.08
1.18.11	17	0.67 ± 0.19	0.68 ± 0.10
1.4.99	17	1	0.87 ± 0.12
1.1.99	16	0.81 ± 0.16	0.67 ± 0.12

Table 1. J-values derived from MEX and Smith-Waterman analysis, corresponding to level 3 classification tasks.

Discussion

Applying the MEX algorithm on a group of 7095 enzymes, it has been shown that the extracted motifs form an excellent basis for classifying these enzymes into small classes known to have different functional roles. In particular, the classification from sequence to function based on these motifs of this enzymes class was demonstrated to outperform any of the alternative methods.

Applying the MEX algorithm on a group of 6602 enzymes, it has been shown that the extracted motifs form an excellent basis for classifying the enzymes represented by these motifs into small classes known to have different functional roles. In particular, the classification from sequence to function based on these motifs of this enzymes class was demonstrated to outperform the SVMProt method on the second level classification task and the Smith-Waterman method on the third level classification task.

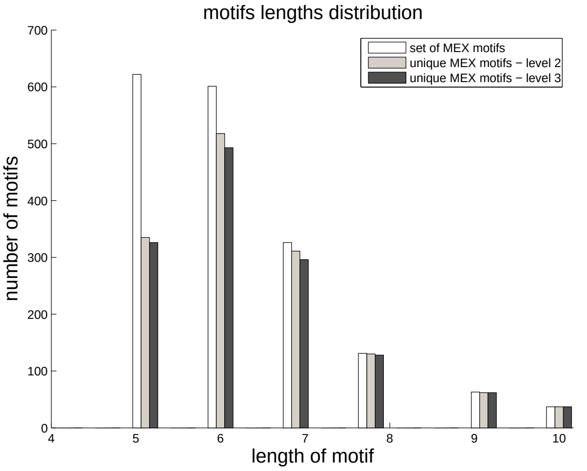

Figure 3. distribution of MEX motifs of lengths 5-10 according to their length. The three sets correspond to (left) entire set of MEX motifs, (middle) set of MEX motifs unique to a single level 2 subclasses and (right) set of MEX motifs unique to a single level 3 sub-subclass.

Our results are compared with two approaches: (i) Classification based on pairwise sequence similarity, analogous to the one employed by [7], using the same SVM procedure that was employed for MEX. As demonstrated, MEX derived motifs form a better basis for classification at the third EC number level, indicating that MEX selected motifs improve the signal to noise ratio inherent in the original sequences. (ii) The SVMProt method introduced by [3, 4] on level 2 data (using their published results). Despite the fact their method is based on semantic information, i.e. physical and chemical properties of the sequence of amino-acids, the results obtained by MEX are better, again indicating that the MEX selected motifs carry relevant information.

It should be noted that the MEX based classification is accomplished by using only 1222 motifs of length 6 or longer. Considering the 48 classification tasks for approximately 4000 proteins, the number of features allowing a successful classification by the MEX algorithm is surprisingly small. Furthermore, as opposed to the regular-expression motifs used by other methods, MEX motifs are all deterministic consecutive amino-acid sequences.

Such regular-expression motifs approach was presented by [2]. They have used regular-expression motifs of average length of 21 amino-acids (termed eMOTIFs) derived in a supervised manner. Applying a feature-selection procedure to select approximately 1000 eMOTIFs out of their original very large set of eMOTIFs, they have achieved impressive classification results. However, while the small number of selected eMOTIFs is comparable to the 1222 motifs used by our approach, it should be noted that the deterministic, consecutive motif sequences extracted by MEX spans a much smaller sequence space than the one spanned by the eMOTIFs, yet, achieving successful classification. Unfortunately, a direct comparison with this work could not be made due to insufficient data.

The application of the MEX algorithm studied here applies only a single level of feature extraction. Higher level patterns may be extracted by iteratively applying MEX, where each MEX iteration uses the observed sequence-motifs as vertices in the MEX graph. Moreover, utilizing the full extent of the ADIOS approach [13] may further reveal higher syntactic structures in biological sequence data, enabling a more extensive coverage of enzyme sequences.

Acknowledgment

This research was partially supported by the US Israel Binational Science Foundation. ZS has a student fellowship of the Horowitz Foundation for Complexity Sciences. We thank Asa Ben-Hur and Doug Brutlag for helpful conversations. Additionally, we thank Elhanan Borenstein for his helpful remarks.

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 evolutionary 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.