BIOINFORMATICS ORIGINAL PAPER

Vol. 21 no. 19 2005, pages 3711–3718
doi:10.1093/bioinformatics/bti608

Structural bioinformatics

Motif-based protein ranking by network propagation

Rui Kuang[1], Jason Weston[4], William Stafford Noble[5] and Christina Leslie[2,3,*]

[1]Department of Computer Science, [2]Center for Computational Biology and Bioinformatics and [3]Center for Computational Learning Systems, Columbia University, New York, NY 10027, USA, [4]NEC Labs, New Jersey, CA 95014, USA and [5]Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA

Received on May 30, 2005; revised on July 23, 2005; accepted on July 29, 2005
Advance Access publication August 2, 2005

ABSTRACT

Motivation: Sequence similarity often suggests evolutionary relationships between protein sequences that can be important for inferring similarity of structure or function. The most widely-used pairwise sequence comparison algorithms for homology detection, such as BLAST and PSI-BLAST, often fail to detect less conserved remotely-related targets.

Results: In this paper, we propose a new general graph-based propagation algorithm called MotifProp to detect more subtle similarity relationships than pairwise comparison methods. MotifProp is based on a protein-motif network, in which edges connect proteins and the k-mer based motif features that they contain. We show that our new motif-based propagation algorithm can improve the ranking results over a base algorithm, such as PSI-BLAST, that is used to initialize the ranking. Despite the complex structure of the protein-motif network, MotifProp can be easily interpreted using the top-ranked motifs and motif-rich regions induced by the propagation, both of which are helpful for discovering conserved structural components in remote homologies.

Availability: http://www.cs.columbia.edu/compbio/motifprop

Contact: cleslie@cs.columbia.edu

1 INTRODUCTION

The most widely-used algorithms for detecting homology in protein sequences have focused on estimating pairwise sequence similarity with sequence–sequence or profile–sequence alignments (Altschul *et al*., 1990, 1997; Waterman *et al*., 1991). These algorithms use dynamic programming or faster heuristic strategies to produce optimal or nearly optimal pairwise local alignments. Moreover, these algorithms exploit a well-established methodology for estimating an E-value to assess the statistical significance of the alignment score between the query and target sequences. However, for remotely related protein sequences, even profile-based ranking algorithms like PSI-BLAST will fail to detect sequence similarity due to the weak statistical significance of the computed alignments. Interestingly, probably because of the lack of a comprehensive enough protein motif database, very few of these previous works make use of motifs to rank proteins. If we simply consider the common motif hits between proteins as the similarity measure, existing motif databases, such as PROSITE database (Hulo *et al*., 2004) and eMOTIF database (Nevill-Manning *et al*., 1998), are more useful for characterizing homologous proteins than for inferring remote homologies.

In this paper, we introduce a new graph-based rank propagation algorithm called MotifProp, where we represent relationships between proteins and motif-based features in a protein-motif network. Instead of relying on a direct measure of pairwise sequence similarity, we assume that most distantly-related proteins share a conserved structure or at least some common conserved structural components. A natural way of trying to capture these subtle similarities is to measure whether the proteins share sequence motifs that might correspond to conserved structural elements. To incorporate this notion into a network structure, we define the protein-motif network—where similarity between protein nodes is indirectly represented by the set of edges connecting protein nodes and motif-based feature nodes. This network is a bipartite graph, where edges connect protein nodes and motif nodes, and a set of common motifs in the network can serve as ‘bridges’ connecting similar proteins. To measure the similarity to a given query, each protein node and motif node is first assigned an activation score with an initial ranking algorithm, which is called the base algorithm. The MotifProp algorithm propagates activation scores through the protein-motif network to learn rankings relative to the query sequence. After convergence of the algorithm, the final protein node activation scores are used to rank the sequences and retrieve homologs of the query. An illustration of a protein-motif network is shown in Figure 1.

For MotifProp to perform well in the ranking task, we need to choose an appropriate set of feature motifs that can capture the remote homology information we are interested in. Recent supervized remote homology detection approaches (Karplus *et al*., 2001; Jaakkola *et al*., 1999; Liao and Noble, 2002; Leslie *et al*., 2004; Ben-Hur and Brutlag, 2003; Kuang *et al*., 2005) have used many different representations of protein sequences, including string kernels based on short inexact-matching k-mer features. Among these recent methods, SVM classification using the profile kernel (Kuang *et al*., 2005)—a k-mer based string kernel defined on profiles—is one of the most successful methods for remote homolog detection. In the profile kernel, each sequence is associated with a sequence profile (derived, for example, from PSI-BLAST) and each k-length window of the profile is implicitly mapped to a vector in the feature space indexed by all k-length subsequences from the alphabet of amino acids; the non-zero features in this vector correspond to k-mers similar to the profile. In this paper, we use k-mers as motifs in the protein-motif network, and we define a set of k-mer features that are similar to windows of a protein sequence either by using a profile (similar to the profile kernel approach) or by using fixed substitution probabilities or exact k-mer matches. In addition, we

[*]To whom correspondence should be addressed.

© The Author 2005. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

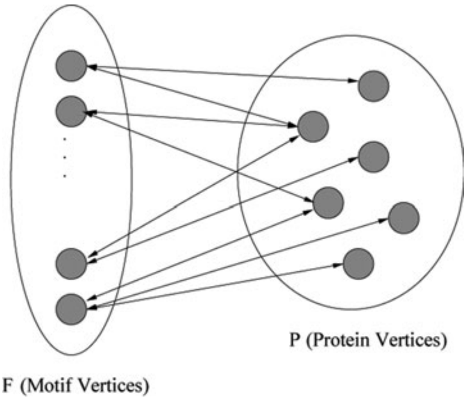

Fig. 1. Protein-motif bipartite network. Edges in the network represent occurrences of a motif within a protein sequence. The edges are weighted according to the significance of the motif occurrence.

experiment with using eMOTIFs (Ben-Hur and Brutlag, 2003) and PROSITE motifs (Hulo *et al*., 2004) in the network. We also explore how to integrate complementary motif sets, such as PROSITE motifs and k-mers, with a sequential MotifProp algorithm to achieve more refined ranking performance.

Our earlier work introducing the RankProp algorithm (Weston *et al*., 2004) showed that one can exploit the global structure of the protein similarity network to achieve significant improvement over PSI-BLAST for the protein ranking problem. The protein similarity network is a weighted graph defined on a large protein database, where protein sequences are the nodes and pairwise PSI-BLAST E-values are used to compute edge weights. RankProp is able to propagate through densely connected regions of the protein similarity network to detect subtle sequence relationships that are missed by methods like PSI-BLAST, which only uses local similarity near the query. Compared with MotifProp, the RankProp algorithm has two major drawbacks. First, it does not provide an explanation of why the target is related to the query by detecting conserved sequence information between them. Second, considerable pre-processing time is needed to compute the similarity network for a large database.

Other related work on network-based propagation algorithms has focused on information retrieval on the web (Kleinberg, 1999) and classification tasks in natural language processing (Radev, 2004). Kleinberg models the hyperlinked web network as a bipartite graph comprising two sets of nodes—hubs (web pages with many pointers to related pages) and authorities (web pages pointed to by hubs)—and propagates activation scores between these node sets. Radev proposes a tripartite updating scheme for a number classification task in text processing by defining a bipartite graph where feature vertices are connected with labeled examples and unlabeled examples, and he cyclically propagates activation between these three kinds of nodes.

In contrast, we formulate a propagation algorithm for learning a ranking in a bipartite graph of proteins and motifs. MotifProp takes the ranking scores of any base ranking algorithm, such as PSI-BLAST, as the initial activation values for protein nodes and initializes activation scores of the motif nodes estimated by matching motifs against the query. It then performs a simple two-direction propagation algorithm that can efficiently propagate activation values between protein nodes and motif nodes to improve over the initial ranking.

In addition to obtaining a ranking of proteins relative to the query, MotifProp induces a ranking of motif features. Those top-ranked motif features are considered important to the ranking and suggest common conserved components among distantly-related proteins. It is biologically interesting to analyze these selected motifs for a better understanding of the relation between motifs and protein superfamilies. Another important property of MotifProp is that high motif activation values suggest consensus patterns with the relatives of the query sequence. If we map the motif activation values back to their original matching positions on the query sequence, then those regions that are important for ranking will be associated with high activation scores. These high activation or 'motif-rich' regions are particularly useful for motif discovery and structural study, and they can provide a biological explanation of the performance improvement over the base ranking algorithm, especially for queries where alignment-based methods cannot successfully retrieve target sequences.

Our paper is organized as follows. In Section 2, we describe all the networks and their corresponding propagation algorithms. In Section 3, we report our experiments on a benchmark dataset of query sequences that have known structural annotations, based on the SCOP database (Murzin *et al*., 1995). We also analyze structural and functional properties of top-ranked motifs and motif-rich regions induced by MotifProp. Finally, in Section 4, we discuss several possible future directions for our network propagation work.

2 METHODS

In this section, we introduce our new method, the MotifProp algorithm. Several natural variants of MotifProp based on different protein-motif networks with motif nodes from either one or multiple motif databases are also discussed. Finally, we describe how to extract motif-rich regions for protein superfamilies using the MotifProp-induced feature activation scores.

2.1 MotifProp: protein-motif bipartite network for rank propagation

A protein-motif bipartite network has two sets of nodes, one set **P** representing protein sequences and another set **F** representing motifs. Edges in the graph only connect nodes in **P** with nodes in **F**, giving a bipartite graph. Let q be the query sequence, H the connectivity matrix between protein vertices and motif vertices, and P and F the vectors of activation values associated with these two sets of nodes with respect to q. MotifProp uses an alternating update rule. In one direction, we propagate the activation values of the protein nodes to the connected feature nodes through weighted edges; in the other direction, we propagate the activation values of the motif nodes to the connected protein nodes through those edges. In both directions, we use a parameter $\alpha \in (0, 1)$ to control the amount of propagation across the bipartite graph relative to reinforcement of the initial activation values. The update rules for the tth round of propagation are given by

$$P^{t+1} = \alpha \widetilde{H} F^t + (1-\alpha) P^0$$
$$F^{t+1} = \alpha \widetilde{H}' P^t + (1-\alpha) F^0,$$

where $\widetilde{H}$ is obtained from H by normalizing so that entries in each row sum to 1 and $\widetilde{H}'$ is a similarly row-normalized version of the transpose of H. F^0 is the vector of initial motif activation values and P^0 is the vector of initial activation values from the base ranking algorithm, each normalized so that entries sum to 1.

The proof of the convergence for the MotifProp algorithm is related to the proof in Zhou *et al*. (2004). We reformulate the MotifProp algorithm as a propagation algorithm with protein nodes and motif nodes. In this formulation, the update rule is given by

$$\begin{pmatrix} P \\ F \end{pmatrix}^{t+1} = \alpha \begin{pmatrix} 0 & \widetilde{H} \\ \widetilde{H}' & 0 \end{pmatrix} \begin{pmatrix} P \\ F \end{pmatrix}^{t} + (1-\alpha) \begin{pmatrix} P \\ F \end{pmatrix}^{0}.$$

The connectivity matrix (of size $|P| + |F|$ by $|P| + |F|$) can be divided into four block submatrices, one between protein nodes ($|P|$ by $|P|$), one between motif nodes ($|F|$ by $|F|$), one between protein nodes and motif nodes ($|P|$ by $|F|$) and a final one between motif nodes and protein nodes. For our algorithm, the diagonal blocks are all 0s, so the row-normalization of H and its transpose is the same as row-normalizing the full $|P| + |F|$ by $|P| + |F|$ connectivity matrix to a stochastic matrix, and the bi-directional update rule is identical to the single update rule given above. The original proof of convergence can therefore be directly applied to our algorithm.

In various propagation algorithms in the machine learning literature, edge weights between nodes are often based on a notion of distance between node data examples. Typically, given a distance $d(i, j)$ between nodes i and j, the corresponding edge weight is first set to $\exp -d(i, j)/\sigma$, where $\sigma > 0$ is a parameter, and then normalized so that the sum of weights of the (directed) edges terminating at any vertex is 1. In RankProp, PSI-BLAST E-values were used as the distance function between protein nodes to compute these Gaussian edge weights, and the Gaussian of the E-value distance from the query to each protein node was used to initialize the activation scores. Following the same motivation, for MotifProp the initial activation values P^0, F^0 and (unnormalized) edge weight H_{ij} are given by

$$P_j^0 = \exp(-S_q(j)/\sigma)$$

$$F_j^0 = \exp(-E_q(j)/\sigma)$$

$$H_{ij} = \exp(-E_i(j)/\sigma),$$

where $\sigma > 0$, $S_q(j)$ is PSI-BLAST E-value assigned to protein j given query q and $E_i(j)$ is the E-value associated with a match of the j-th motif in the i-th protein. Sequences that are left unranked by PSI-BLAST are assigned an initial activation score of 0, assuming a very large E-value against the query. Various types of motifs can be used in the protein-motif network, such as the eMOTIF database, the PROSITE database, or the set of all possible k-length sequences from the amino acid alphabet. To estimate E-values for motif matches, we follow the methodology of ungapped alignments (Altschul *et al.*, 1990), in which the E-value of an optimal match with score S is approximated by

$$E(S) = K|x||f|\mathrm{e}^{-\lambda S},$$

where K and λ are parameters depending on the substitution matrix and background frequency of amino acids, and $|x|$ and $|f|$ are the length of protein x and motif f. We note that the PSI-BLAST E-values, used to compute initial activation scores for protein nodes, and the motif E-values, used to compute edge weights and initial activation scores for motif nodes, are not directly comparable, since they give E-values for different statistical contexts. Moreover, in the latter case, the E-value formula is motivated by asymptotic theory, so the dependence on the length of the motif is only a rough approximation. Nevertheless, the difference between these E-values is not a concern here, since protein nodes and motif nodes play different roles in the MotifProp algorithm: protein and motif activation scores are normalized separately, and the algorithm induces separate rankings of proteins and motifs relative to the query. We emphasize that the E-values are used only to obtain a notion of distance between motifs and protein sequences; other reasonable choices of distance should also work.

If we use k-mers as feature nodes, all k-mers that match exactly or inexactly to k-length subsequences of sequence x are represented by edges between node x and the corresponding motif nodes in the bipartite graph. A good way of computing the matches between k-mers and sequences is motivated by the profile-based k-mer mapping (Kuang *et al.*, 2005). In this case, the substitution score between a k-mer f and sequence x is $\max_p S_f(x_{p..p+k-1})$, where $p = 1...|x| - k + 1$ and $S_f(x_{p..p+k-1}) = \sum_l^k s_l(f_l, x_{p+l-1})$ is a sum of log odds substitution scores $s_l(\cdot,\cdot)$ depending on the profile at position $p + l - 1$, and all the k-mers with a substitution score less than a user-defined threshold are considered to be motif hits against sequence x. If we do not have a profile for our protein sequences, then we can restrict the network to include only edges that correspond to exact occurrences of k-mers f in sequences x (similar to the k-spectrum mapping (Leslie *et al.*, 2002)) and use the diagonal entries of a fixed substitution matrix for scoring. (In principle, the parameters K and λ depend on the profile, but for simplicity, we use the same parameter values for all sequences.)

One can also consider using motifs derived from the PROSITE database. However, PROSITE contains different types of motifs, such as regular expressions and profiles, making it hard to establish a consistent substitution scoring system. For simplicity, we consider only exact matches between motifs and sequences, and we estimate the substitution score using diagonal entries of a fixed substitution matrix as described above. For the case of the eMOTIF database, the E-value based on the substitution probability between amino acids and amino acid groups is provided by the eBAS package (Huang and Brutlag, 2001) and can be used directly for estimating edge weights.

The protein-motif network is much less expensive to compute than the protein similarity network used with the RankProp algorithm (Weston *et al.*, 2004). Optimal sequence alignment requires time that is quadratic in sequence length, rendering it expensive to compute in practice. In that case, the time complexity for building a full protein similarity network is $O(\sum_{i=1}^{|P|}\sum_{j=1}^{|P|}(|p_i||p_j|))$, where $|p_i|$ denotes the length of protein sequence p_i. With heuristic algorithms, such as PSI-BLAST, we can obtain a reduced complexity of approximately $O(n \log n)$ in sequence length. However, it is still often prohibitively expensive to compute the similarity network if you have a large database of proteins. On a SUN cluster with 20 host machines, it takes longer than 1 month to compute a protein similarity network with 101 602 sequences from Swiss-Prot. On the other hand, the time complexity for building a protein-motif network is $O(\sum_{i=1}^{|P|}\sum_{j=1}^{|F|} C(p_i, f_j))$, where f_j denotes a motif and C is the time complexity for matching a motif against a sequence, which is usually bounded by $O(|p_i|)$. Empirically, on a single machine in the SUN cluster, it takes <2 days to compute a protein-motif network with 101 602 sequences from Swiss-Prot and either 223 390 motifs from the eMOTIF database, 1639 PROSITE motifs or all 160 000 4mers (using exact matches).

2.2 Sequential MotifProp: integration of multiple motif databases

Existing protein motif databases are built using various methods. They are represented as position-specific scoring matrices (Hulo *et al.*, 2004), regular expressions (Nevill-Manning *et al.*, 1998) or k-mers (Kuang *et al.*, 2005). Ideally we expect that the superset of all motifs will produce the most comprehensive protein-motif network. However, given different properties of these motif sets, the expected number of motif hits and corresponding E-values for a given query protein sequence can vary significantly across these motif sets. This specificity difference makes combination of motif hits into a unified scoring schema for MotifProp a hard problem. Our empirical experiments suggest that a simple weighted linear combination of PROSITE motifs and k-mer motifs does not improve over using each individual set of motifs (results not shown). It is also computationally intensive to cross-validate for an optimal weighting between the motif sets, considering the possible number of motif databases we could use. We propose a simple sequential MotifProp based on protein-motif networks containing multiple motif sets to resolve this specificity problem.

In a protein-motif network with multiple motif sets, motif nodes F can be divided into an n-set partition $\{F_{(1)}, F_{(2)}, ..., F_{(n)}\}$, in which $F_{(i)}$ is a set of motifs from the ith motif set. Edges only connect each subset $F_{(i)} (i \in [1, n])$ and protein vertices P. The edge weights are estimated using only on the motif model for each individual $F_{(i)}$. The connectivity matrices $H_{(i),n}$ and $H'_{(i),n}$ are normalized individually in the same manner as a protein-motif network comprising vertices P and $F_{(i)}$ only. In this larger network, we run MotifProp to convergence with each set of motifs for n sequential runs, using the final activation values from the ith motif set to initialize the activation

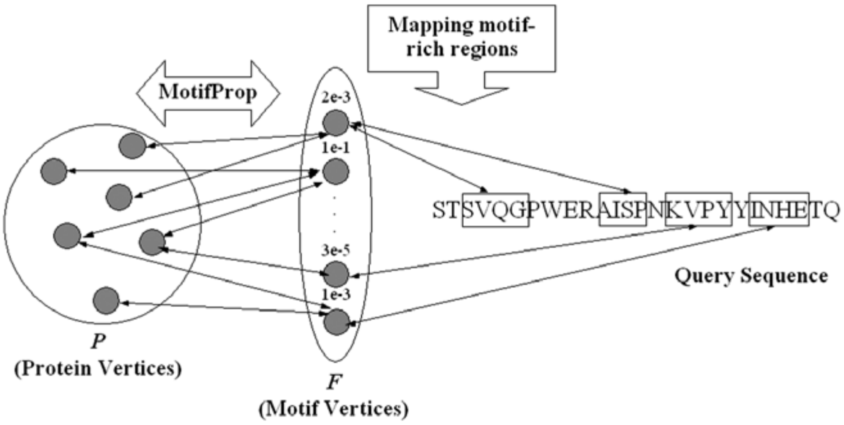

Fig. 2. Mapping motif-rich regions. After motif propagation, each motif node is associated with an activation score. We map each motif back to the query sequence and accumulate the activation scores for each position. Those positions with high scores are classified as motif-rich regions.

values for the $(i+1)$st motif set. This simple sequential MotifProp algorithm avoids the normalization issue involved in combining different motif sets, while allowing a cumulative and consistent updating to the activation values of protein nodes P.

2.3 Identifying motif-rich regions

In the MotifProp algorithm, we learn a vector of activation values for both protein nodes and motif nodes. Motifs with high activation values are those connecting the query sequence and its homologs by strong edges. Thus, their corresponding matches on the query sequence correspond to motifs that are conserved during the evolution of the protein superfamily or fold. To analyze these motif regions, we compute an accumulated activation value for each position on the query sequence and mark those positions with high score density as regions of interest using

$$\sum_j F_j^* \delta(p, j),$$

where F^* is the vector of final activation values for all motif features and $\delta(p, j)$ is a binary function denoting whether the p-th position falls inside a match of the j-th feature (Fig. 2). With this mapping, a distribution of activation scores can be obtained over the positions. We sort the positions by their activation values and take the top-scoring positions that contribute 90% of the cumulative total score. We define these top-scoring positions to be our motif-rich regions. We expect these regions to correspond to conserved regions among homologs of the query, so they are potentially useful for structural or functional motif analysis. In Section 3.4, we compare motif-rich regions extracted in our experiment with PDB annotations, and we show the corresponding functional and structural properties of motif-rich regions for several superfamily examples.

3 EXPERIMENTS

We test MotifProp algorithm on 7329 protein domains with known 3D structures in the human-annotated SCOP database version 1.59 (Murzin *et al*., 1995). After purging with http://astral.berkeley.edu, no pair of proteins have >95% sequence identity. Following the setup in Weston *et al*. (2004), these 7329 sequences are divided into a training set (4246 sequences) and a test set (3083 sequences). The training set is used for estimating optimal parameters for the propagation algorithm. In order to propagate through a large network, we use 101 602 proteins from Swiss-Prot (version 40) as additional nodes. For all rank propagation algorithms we perform 20 iterations, which is generally sufficient for convergence, and we compare against PSI-BLAST and RankProp for the protein ranking task of retrieving target sequences from the same superfamily as the query. Although the propagations are on sequences from both the SCOP database and the Swiss-Prot database, we only consider the ranking among SCOP domains, i.e. the performance is reported on the ranking of SCOP domains obtained when Swiss-Prot sequences are removed from the ranked list.

We evaluate the ranking performance by using three variants of the receiver operating characteristic (ROC) score (Hanley and McNeil, 1982). For a given query, this score measures the area under a curve that plots true positives as a function of false positives for varying classification thresholds. The ROC$_n$ score computes this score up to the n-th false positive (Gribskov and Robinson, 1996). The score is normalized to fall between 0 and 1, and for this problem, the expected score from a random query is close to 0. For each method, we report the mean ROC$_{50}$, ROC$_{10}$ and ROC$_1$ scores across all queries in the test set. These three scores place increasing emphasis on producing high-quality predictions at the top of the ranked list. ROC$_1$ is probably of the most interest for this application.

3.1 Protein-motif rank propagation

We test the propagation algorithm on three different protein-motif networks, one produced from the eMOTIF database (version 3.6) (Ben-Hur and Brutlag, 2003), one from the PROSITE motif database (version 18.34) (Hulo *et al*., 2004) and the last produced from a k-mer feature mapping (Leslie *et al*., 2002; Kuang *et al*., 2005). We use the eBAS package (Ben-Hur and Brutlag, 2003) to scan proteins against the eMOTIF database and the PROSCAN package (Gattiker *et al*., 2002) for scanning against the PROSITE database. For the k-mer mapping, we set $k=4$, and we use a profile feature mapping (Kuang *et al*., 2005) for SCOP sequences with a threshold of six, but for improved efficiency, we use an exact k-mer matching (Leslie *et al*., 2002) for Swiss-Prot sequences. The PSI-BLAST profiles of SCOP sequences are produced by searching against the SCOP dataset plus Swiss-Prot sequences.

By measuring the error rate on the training set queries, the parameter $\alpha = 0.85$ is chosen for the k-mer MotifProp algorithm, $\alpha = 0.4$

Table 1. Comparison of overall ranking quality produced by various algorithms

Algorithm	ROC_1	ROC_{10}	ROC_{50}
Sequential MotifProp	0.640	0.663	0.688
k-mer MotifProp	0.621	0.648	0.679
RankProp	0.592	0.667	0.725
PROSITE MotifProp	0.600	0.643	0.664
eMOTIF MotifProp	0.527	0.612	0.666
PSI-BLAST	0.594	0.616	0.641

The table lists, for each algorithm, the mean ROC_1, ROC_{10} and ROC_{50} scores over 3083 test query sequences.

Table 2. Pairwise comparisons of algorithms

ROC type	Algrotihm 1	Algorithm 2	Win (%)	Lose (%)	Tie (%)
ROC_1	k-mer MotifProp	PSI-BLAST	25.0	2.6	72.4
ROC_{10}	k-mer MotifProp	PSI-BLAST	28.0	3.4	68.6
ROC_{50}	k-mer MotifProp	PSI-BLAST	32.4	9.9	57.8
ROC_1	Sequential MotifProp	RankProp	27.3	12.6	60.1
ROC_{10}	Sequential MotifProp	RankProp	26.1	20.5	53.4
ROC_{50}	Sequential MotifProp	RankProp	12.7	29.3	58.1

Column ‘Win’ is the percentage of queries with better results in the first method. Column ‘Tie’ is the percentage of queries with tied results. Column ‘Lose’ is the percentage of queries with worse results in the first method.

for eMOTIF MotifProp and $\alpha = 0.7$ for PROSITE MotifProp. Parameter σ is set to 100 in all experiments for consistency with RankProp. In Table 1 we show the average ROC scores across the test queries. All three networks produce improved ROC_{50} scores over PSI-BLAST, while the PROSITE and k-mer networks also improve ROC_{10} and ROC_1, and the eMOTIF network produces much weaker ROC_1 scores. The comparison of ROC scores between the MotifProp methods and PSI-BLAST or RankProp is shown in Figure 3 and Table 2. We can conclude from our results that MotifProp with various types of motifs gives significant improvement over the PSI-BLAST ranking, but in terms of ROC_{50} performance, it does not perform quite as strongly as RankProp with our network setup and parameter choices. We note that there also exists an adaptive version of RankProp, where the weighting parameter σ on the edges is allowed to change per query. By learning a rule for adapting sigma using the training set as explained in the supplementary material for Weston *et al.* (2004), adaptive RankProp achieved an ROC_1 performance of 0.6502, significantly outperforming PSI-BLAST. However, as MotifProp can also be used with the same adaptive σ technique, for simplicity we did not do this final tuning step and report results with non-adaptive versions of both RankProp and MotifProp. We observe in all experiments a steep improvement of ROC scores at around $(0.7, 0.9)$. We found most of the improved queries from below 0.7 to this region are members of the immunoglobulin superfamily, which is the largest superfamily in the SCOP database. This suggests that MotifProp can exploit the cluster of protein sequences with the protein-motif network just as RankProp does (Weston *et al.*, 2004).

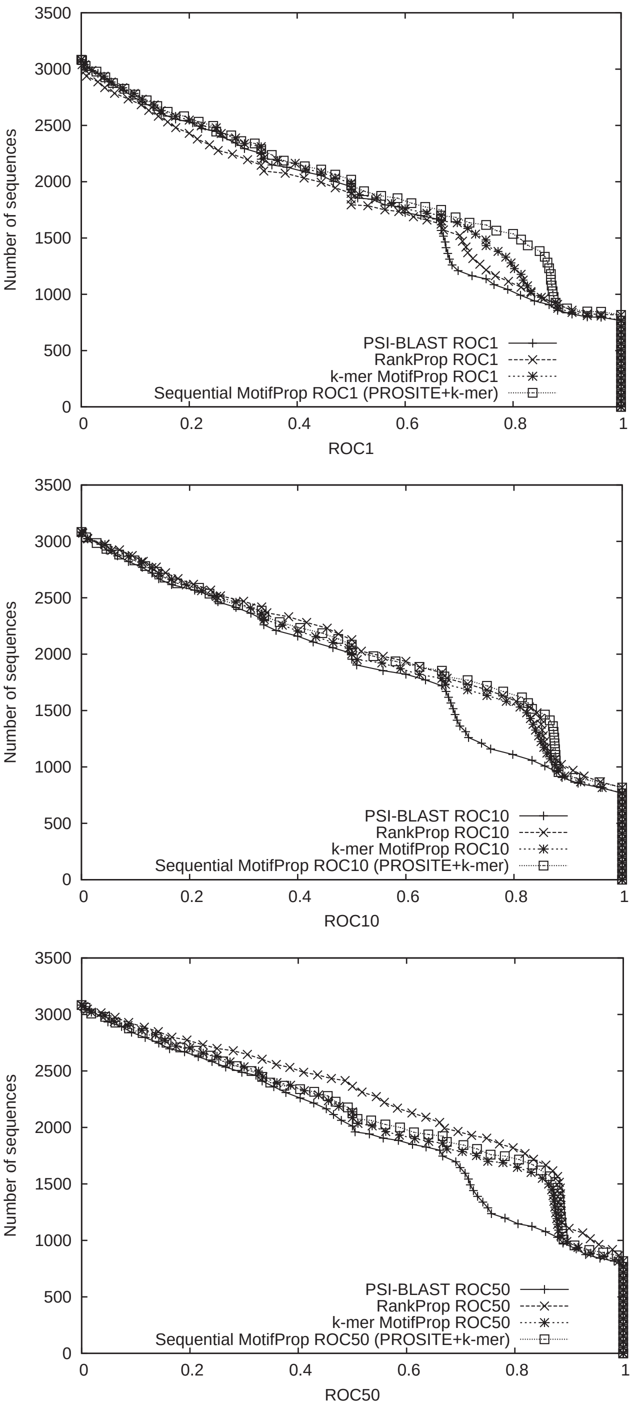

Fig. 3. Comparison of the ROC scores of PSI-BLAST, RankProp and MotifProp for test queries. The graph plots the total number of queries for which a given method exceed an ROC_n score threshold ($n = 1$, 10 and 50).

3.2 Sequential MotifProp for protein-motif network with multiple motif sets

We experiment with a protein-motif network using two sets of motifs, PROSITE motifs and k-mers. In the first step of sequential MotifProp, we run the MotifProp algorithm between protein nodes and

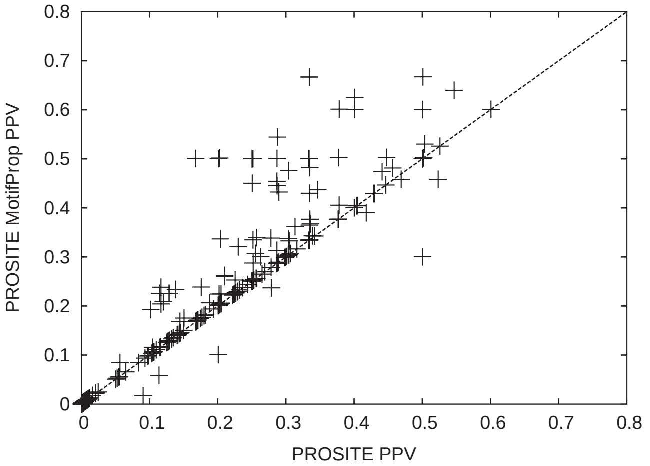

Fig. 4. Comparison of average positive predictive values of PROSITE motifs for 330 superfamilies before and after propagation.

Table 3. Comparison of PROSITE motif hits before and after propagation for domains from Chaperone J-domain superfamily and domains from Nitrilase/N-carbamoyl-D-aminoacid amidohydrolase superfamily (N/N-D amidohydrolase)

ID	Chaperone J-domain					N/N-D amidohydrolase		
	1fpoa1	1hdja1	1xbl	1fafa	1gh6a	1emsa2	1f89a	1er2a
PS00003							X	X
PS00004							X	
PS00005	X		X			X	XO	X
PS00006	X	X	XO	XO		XO	XO	X
PS00007					X	X	XO	
PS00008		XO	XO	XO	X	XO	XO	XO
PS00636	O	XO	XO		XO			
PS00637	O	O	O		O			
PS00920						O	O	O
PS00921						O	O	O
PS01227						XO	XO	O
PS50076	XO	XO	XO		XO			
PS50263						XO	XO	XO
PS50910					O			

A cross denotes a motif hit before propagation and a circle denotes a hit after.

PROSITE motifs. Afterwards, keeping the activation values from the first round for the protein nodes, a second round of MotifProp propagates between the protein nodes and k-mers. We directly use the optimal $\alpha = 0.7$ (selected using the training set) for the PROSITE MotifProp in the first step, and parameter $\alpha = 0.55$ is picked by again using the training set for the k-mer MotifProp in the second step. The comparisons of ROC scores for PSI-BLAST, RankProp, k-mer MotifProp and sequential MotifProp on test queries are shown in Table 1 and Figure 3. A pairwise comparison between RankProp and sequential MotifProp is reported in Table 2. On average, the sequential MotifProp based on PROSITE motifs and k-mers achieves stronger performance on ROC$_1$, comparable ROC$_{10}$ but slightly weaker ROC$_{50}$ compared with RankProp.

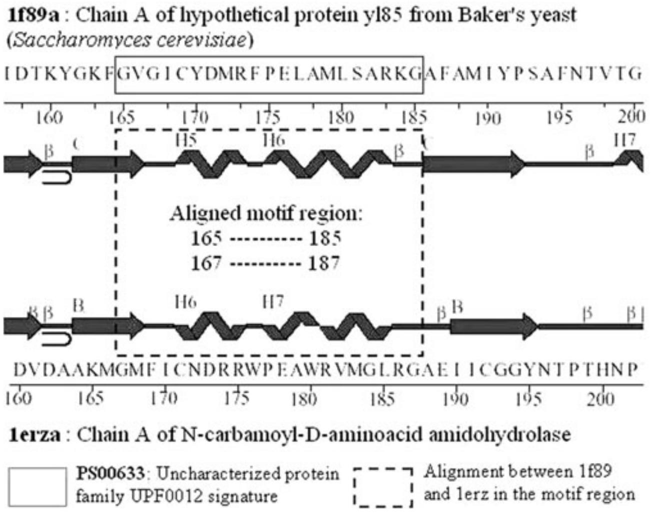
1f89a : Chain A of hypothetical protein yl85 from Baker's yeast (Saccharomyces cerevisiae)
IDTKYGKFGVGICYDMRFPELAMLSARKGAFAMIYPSAFNTVTG
160 165 170 175 180 185 190 195 200
H5 H6 H7
Aligned motif region:
165 ---------- 185
167 ---------- 187
H6 H7
DVDAAKMGMFICNDRRWPEAWRVMGLRGAEIICGGYNTPTHNP
160 165 170 175 180 185 190 195 200
1erza : Chain A of N-carbamoyl-D-aminoacid amidohydrolase
PS00633: Uncharacterized protein family UPF0012 signature
Alignment between 1f89 and 1erz in the motif region

Fig. 5. Identification of potential motif hit through MotifProp. PROSCAN cannot find a hit of motif PS01227 on chain A of SCOP domain 1er2, while MotifProp suggests that it is a potentially related motif. The aligned region on 1er2a shares similar structure pattern with the PS01227 motif region on chain A of SCOP domain 1f89.

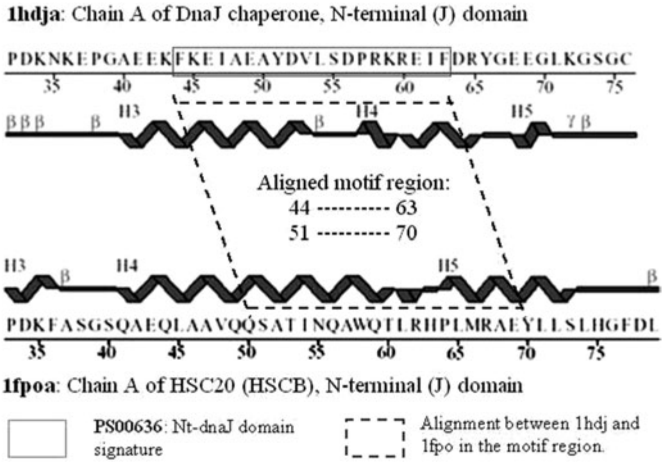
1hdja: Chain A of DnaJ chaperone, N-terminal (J) domain
PDKNKEPGAEEKFKEIAEAYDVLSDPRKREIFDRYGEEGLKGSGC
35 40 45 50 55 60 65 70 75
H3 H4 H5
Aligned motif region:
44 ---------- 63
51 ---------- 70
H3 H4 H5
PDKFASGSQAEQLAAVQQSATINQAWQTLRHPLMRAEYLLSLHGFDL
35 40 45 50 55 60 65 70 75
1fpoa: Chain A of HSC20 (HSCB), N-terminal (J) domain
PS00636: Nt-dnaJ domain signature
Alignment between 1hdj and 1fpo in the motif region.

Fig. 6. Identification of potential motif hit through MotifProp. PROSCAN cannot find a hit of motif PS00636 on chain A of SCOP domain 1fpo, while MotifProp suggests that it is a potentially related motif. The aligned region on 1fpo shares similar structure pattern with the PS00636 motif region on chain A of SCOP domain 1hdj.

3.3 PROSITE motif selection with MotifProp

To show that MotifProp is a natural way for motif feature selection, we compare the PROSITE motif hits detected by the PROSCAN program and top-ranked PROSITE motifs given by activation scores from MotifProp. Here, we retrieve the same number of motifs from MotifProp-ranked list as the number of PROSCAN-detected motif hits to make them easily comparable. We report the average positive predictive value (PPV) (Ben-Hur and Brutlag, 2005) of motif hits by SCOP superfamilies. Here, PPV is the ratio of the occurrence of a PROSITE motif m in a SCOP superfamily S, i.e. $\mathrm{ppv}(m) = \frac{\mathrm{count}(m|S)}{\mathrm{count}(m)}$, where count$(m)$ is the occurrence of m and

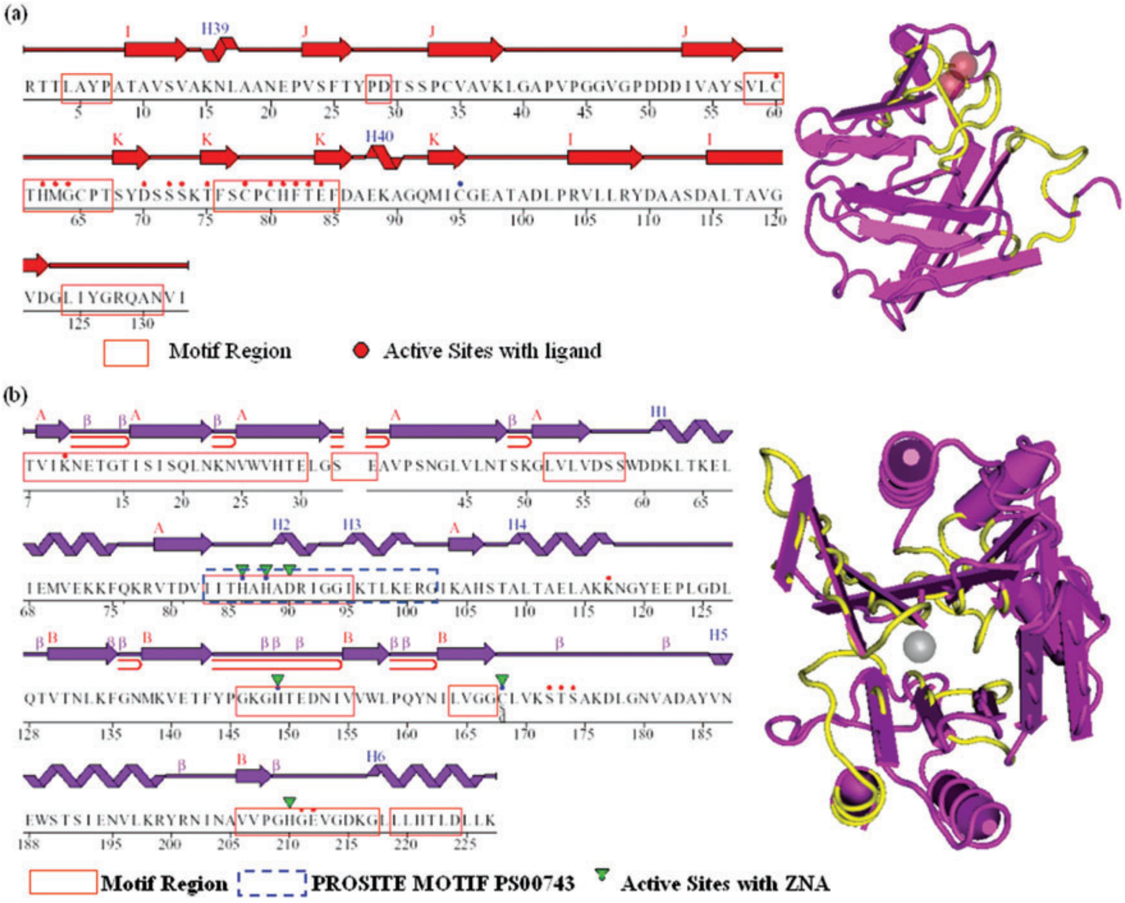
(a)
RTTLAYPATAVSVAKNLAANEPVSFTYPDTSSPCVAVKLGAPVPGGVGPDDDIVAYSVLC
THMGCPTSYDSSSKTFSCPCHFTEFDAEKAGQMICGEATADLPRVLLRYDAASDALTAVG
VDGLIYGRQANVI
Motif Region
Active Sites with ligand
(b)
TVIKNETGTISISQLNKNVWVHTELGS
EAVPSNGLVLNTSKGLVLVDSSWDDKLTKEL
IEMVEKKFQKRVTDVIITHAHADRIGGIKTLKERGIKAHSTALTAELAKKNGYEEPLGDL
QTVTNLKFGNMKVETFYPGKGHTEDNIVVWLPQYNILVGGCLVKSTSAKDLGNVADAYVN
EWSTSIENVLKRYRNINAVVPGHGEVGDKGLLLHTLDLLK
Motif Region
PROSITE MOTIF PS00743
Active Sites with ZNA

Fig. 7. Examples of motif-rich regions. (**a**) Motif-rich regions on chain B of arsenite oxidase protein from the ISP protein superfamily. (Left) The PDB sequence anotation (PDB id 1g8k) and motif-rich regions are given. There are no PROSITE motif hits for this sequence. (Right) The 3D protein structure is shown with motif-rich regions in yellow. The ligands are indicated as pink balls. The two motif-rich regions in the bottom-right of the picture are located at the interface of two different subunits. (**b**) Motif-rich regions on chain A of metallo beta-lactamase *ii* from bacillus cereus 569/h/9 from the metallo-hydrolase/oxidoreductase protein superfamily. (Left) The PDB sequence anotation (PDB id 2bc2), PROSITE motif regions and motif-rich regions are given. (Right) The 3D protein structure is shown with motif-rich regions in yellow. The zinc molecule is shown in gray.

$\text{count}(m|S)$ is the occurrence of m in S. A PPV reflects how specific a motif is with respect to a superfamily. In other words, motifs with higher PPVs can characterize the superfamily better since these motifs are rarely observed elsewhere. In Figure 4, we show that the average PPVs of motifs by superfamilies are higher in the MotifProp-induced feature representation than in the PROSCAN-detected one. The result suggests that MotifProp selects more specific motif features for the 330 SCOP superfamilies with >1 sequence in the test set.

We also look closer into several SCOP superfamilies with the most improvement of average PPV and find two examples of particular interest: Nitrilase/N-carbamoyl-D-aminoacid amidohydrolase (SCOP 1.59 superfamily d.160.1) and Chaperone J-domain (SCOP 1.59 superfamily a.2.3). In Table 3, we show the change of motif hits for the domains from these two superfamilies. MotifProp gets rid of frequently matched short PROSITE motifs (PS00003, PS00004, PS00005, PS00006, PS00007 and PS00008) and identifies more important functional motifs, i.e. PS01227 for Nitrilase/N-carbamoyl-D-aminoacid amidohydrolase and PS00636 for Chaperone J-domain. More interestingly, MotifProp also detects some motifs that do not occur in any test domain in the table, such as PS00920 and PS00921 (Nitrilases/cyanide hydratase signatures) for Nitrilase/N-carbamoyl-D-aminoacid amidohydrolase superfamily and PS00637 (*CXXCXGXG* DNAJ domain signature) for the Chaperone J-domain superfamily. These motifs are reported to be functionally related to the superfamilies in previous works (Kobayashi *et al*., 1993; Cyr *et al*., 1994). In Figure 5 and Figure 6, we show that we can identify the potential motif regions on chain A of 1er2 and on chain A of 1fpo through MotifProp, although the PROSCAN program cannot detect the motif hits on these domains.

3.4 Motif-rich region analysis

To study the motif-rich regions, we map the activation score of each k-mer back on the query sequence to get an activation score distribution over positions. We take those positions with high score density as our motif-rich regions. To show that activated motifs from MotifProp are potentially useful for analyzing structural features, we manually exam the motif-rich regions in two superfamilies with high MotifProp ROC_1 scores. By comparing them with PDB annotations, we identify common functional and structural characteristics captured by those regions in the same superfamily.

The first example is from the ISP (iron sulfur protein) superfamily (SCOP 1.59 superfamily b.33.1). Proteins in this superfamily consist of two conserved all-beta sub-domains. The small one has a rubredoxin-like fold, and the larger one comprises six beta-stands packed in either a sandwich of two three-stranded sheets or a closed barrel. We find that our motif-rich regions correspond very well with the ligand binding sites for all six sequences from this superfamily in our test set. In Figure 7a, we illustrate this phenomenon for one example from the superfamily, chain B of arsenite oxidase protein.

The second example is from the metallo-hydrolase/ oxidoreductase protein superfamily (SCOP 1.59 superfamily d.157.1). Proteins in this superfamily have a duplicate of a beta-alpha-beta-alpha motif, four layers of a/b/b/a and mixed beta-sheets. This protein uses zinc as the natural cofactor. We find that the functional binding sites with zinc are all captured by the motif-rich regions for all six members of the superfamily in our test set. One of the six sequences, chain A of metallo beta-lactamase *ii* protein, is illustrated in Figure 7b.

4 DISCUSSION

In this paper, we present the MotifProp algorithm, which performs rank propagation through a protein-motif bipartite network, a global network structure that represents similarity of protein sequences through common occurrences of motifs. Experiments show that MotifProp—used with eMOTIF, PROSITE, or *k*-mer motifs—can significantly improve over PSI-BLAST for protein ranking. MotifProp performs as well as the RankProp algorithm when measured using ROC$_{50}$ and ROC$_{10}$ scores and outperforms RankProp based on the more stringent ROC$_1$ score, and MotifProp also offers the ability to make motif selections and extract structurally and functionally significant motif-rich regions of the query sequence. Case studies on several protein superfamilies showed that MotifProp can select more conserved motifs for SCOP superfamilies, and motif-rich regions obtained from MotifProp coincide with PDB annotations of active or binding sites. Furthermore, MotifProp significantly reduces the expense of the all-versus-all network pre-computation (see section 2.1).

One possible direction for extending MotifProp is to combine the protein-motif bipartite network with the similarity network from RankProp to form a hybrid network. However, in our preliminary experiments we found that rank propagation in the hybrid network did not outperform MotifProp or RankProp. We hypothesize that the reason for the lack of improvement is that the PSI-BLAST E-values used to define the similarity network contain much of the same information as the motifs used as features in the bipartite network. However, if we could use features that complement PSI-BLAST alignment information, then we might be able to achieve improvement through a hybrid network.

ACKNOWLEDGEMENTS

The authors thank Dragomir R. Radev and Lan Xu for helpful discussions. The authors especially thank the computer support staff in the department of Genome Sciences at the University of Washington. This work is supported by NIH grant GM074257-01 and NSF grant ITR-0312706. WSN is an Alfred P. Sloan Research Fellow.

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al*. (1990) A basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Ben-Hur,A. and Brutlag,D. (2003) Remote homology detection: a motif based approach. In *Proceedings of the Eleventh International Conference on Intelligent Systems for Molecular Biology*, (ISMB 2003), Brisbane, Australia, June 29–July 3, 2003, *Bioinformatics*, **19**, i26–i33.

Ben-Hur,A. and Brutlag,D. (2005) Feature extraction, foundations and applications. *Sequence motifs: Highly Predictive Features of Protein Function*. Springer Verlag.

Cyr,D.M. *et al*. (1994) Dnaj-like proteins: molecular chaperones and specific regulators of hsp70. *Trends Biochem. Sci.*, **19**, 176–181.

Gattiker,A. *et al*. (2002) Scanprosite: a reference implementation of a prosite scanning tool. *Appl. Bioinformatics*, **1**, 107–108.

Gribskov,M. and Robinson,N.L. (1996) Use of receiver operating characteristic (roc) analysis to evaluate sequence matching. *Comput. Chem.*, **20**, 25–33.

Hanley,J.A. and McNeil,B.J. (1982) The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, **143**, 29–36.

Huang,J.Y. and Brutlag,D.L. (2001) The EMOTIF database. *Nucleic Acids Res.*, **29**, 202–204.

Hulo,N. *et al*. (2004) Recent improvements to the prosite database. *Nucleic Acids Res.*, **32**, 134–137.

Jaakkola,T., Diekhans,M. and Haussler,D. (1999) Using the Fisher kernel method to detect remote protein homologies. In *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*, AAAI Press, pp. 149–158.

Karplus,K. *et al*. (2001) What is the value added by human intervention in protein structure prediction? *Proteins*, **45**(S5), 86–91.

Kleinberg,J. (1999) Authoritative sources in a hyperlinked environment. *J. ACM*, **46**, 604–632.

Kobayashi,M. *et al*. (1993) Nitrilase in biosynthesis of the plant hormone indole-3-acetic acid from indole-3-acetonitrile: Cloning of the alcaligenes gene and site-direted mutagenesis of cysteine residues. *Proc. Natl Acad. Sci. USA*, **93**, 247–249.

Kuang,R., Ie,E., Wang,K., Wang,K., Siddiqi,M., Freund,Y. and Leslie,C. (2005) Profile-based string kernels for remote homology detection and motif extraction. *J. Bio. Comp. Biol.*, **3**(3), 527–550.

Leslie,C., Eskin,E., Cohen,A., Weston,J. and Noble,W.S. (2004) Mismatch string kernels for discriminative protein classification. *Bioinformatics*, **20**, 467–476.

Leslie,C. *et al*. (2002) The spectrum kernel: a string kernel for SVM protein classification. *Pac. Bio. Symp.*, **7**, 566–575.

Liao,L. and Noble,W.S. (2002) Combining pairwise sequence similarity and support vector machines for remote protein homology detection. In *Proceedings of the Sixth Annual International Conference on Research in Computational Molecular Biology*, Washington, DC, ACM Press, pp. 225–232.

Murzin,A.G. *et al*. (1995) SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.

Nevill-Manning,C.G. *et al*. (1998) Highly specific protein sequence motifs for genome analysis. *Proc. Natl Acad. Sci. USA*, **95**, 5865–5871.

Radev,D. (2004) Weakly supervised graph-based methods for classification. Technical Report CSE-TR-500-04, Department of Electrical Engineering and Computer Science, University of Michigan.

Waterman,M.S., Joyce,J. and Eggert,M. (1991) Computer alignment of sequences, *Phylogenetic Analysis of DNA Sequences*, Oxford University Press, pp. 59–72.

Weston,J. *et al*. (2004) Protein ranking: from local to global structure in the protein similarity network. *Proc. Natl Acad. Sci. USA*, **101**, 6559–6563.

Zhou,D. *et al*. (2004) Ranking on data manifolds. *Adv. Neural Inf. Process. Syst.*, **16**, 169–176.