Available at www.ComputerScienceWeb.com
POWERED BY SCIENCE DIRECT®

Information Sciences 155 (2003) 1–18

INFORMATION SCIENCES
AN INTERNATIONAL JOURNAL

www.elsevier.com/locate/ins

Automated data-driven discovery of motif-based protein function classifiers ☆

Xiangyun Wang, Diane Schroeder, Drena Dobbs, Vasant Honavar *

Artificial Intelligence Laboratory, Department of Computer Science and Graduate, Program in Bioinformatics and Computational Biology, Iowa State University, Ames, IA 50011-1041, USA

Received 4 July 2001; received in revised form 10 March 2002; accepted 26 April 2002

Abstract

This paper describes an approach to data-driven discovery of decision trees or rules for assigning protein sequences to functional families using sequence motifs. This method is able to capture regularities that can be described in terms of presence or absence of arbitrary combinations of motifs. A training set of peptidase sequences labeled with the corresponding MEROPS functional families or clans is used to automatically construct decision trees that capture regularities sufficient to assign the sequences to their respective functional families. The performance of the resulting decision tree classifiers is then evaluated on an independent test set. We compared the rules constructed using motifs generated by a multiple sequence alignment based motif discovery tool (MEME) with rules constructed using expert annotated PROSITE motifs (patterns and profiles). Our results indicate that the former provide a potentially powerful high throughput technique for constructing protein function classifiers when adequate training data are available. Examination of the generated rules in relation to

☆ This research was supported in part by grants from the National Science Foundation (0219699, 9972653), the Carver Foundation, and Pioneer Hi-Bred, Inc. This research has benefited from interactions with Dr. Dake Wang, Zhong Gao, Changhui Yan, and Carson Andorf of the Iowa State University Artificial Intelligence Research Laboratory. A preliminary version of this paper appeared in the Proceedings of the Conference on Computational Biology and Genome Informatics, part of the Joint Conference on Information Sciences (JCIS), held at Durham, North Carolina, in 2002.

* Corresponding author. Tel.: +1-515-294-1098.
E-mail address: honavar@cs.iastate.edu (V. Honavar).
URL: http://www.cs.iastate.edu/~honavar/aigroup.html.

doi:10.1016/S0020-0255(03)00067-7

known three-dimensional structures of members in the case of two families (MEROPS families C14 and M12) suggests that the proposed technique may be able to identify combinations of sequence motifs that characterize functionally significant three-dimensional structural features of proteins.

1. Background and introduction

Proteins are the main catalysts, structural elements, signaling messengers, and molecular machines in tissues. Hence, assigning putative *functions* to protein sequences is one of the most important problems in functional genomics. Until recently, the primary source of information about protein function has come from biochemical, structural, or genetic experiments on individual proteins. The post-genomic era offers new opportunities and challenges in characterization of protein function from multiple perspectives, using diverse sources of information [17,37].

Of the various sources of data that can be used for assigning proteins to functional families, protein sequence information is perhaps the least expensive and the most readily available. Consequently, sequence-based approaches to protein function prediction are among the best developed. One such approach to assignment of function to protein sequences is a nearest neighbor approach using sequence similarity. Nearest neighbors, i.e., sequences that are most similar to query sequences are detected using programs such as Blast [1] or Fasta [30]. Such tools typically assist users in picking the highest scoring hit(s) with informative annotation to generate a plausible function of the query sequence. Sequence search often returns multiple results, so significant human expertise is needed in interpreting the results. The reliability of homologues detected by multiple sequence alignment falls rapidly once the pairwise sequence identity drops below 30% [34]. Furthermore, at shorter alignment lengths (9 out of 16 aligned residues), it becomes impossible to infer structural similarity although results can be improved by careful exploration of related sequences to accumulate further evidence. While there is substantial evidence that structure is preserved among *homologous* proteins (i.e., those encoded by genes that have evolved from a common ancestor), sequence similarity is strongly correlated with the structure [13,14], the evidence is less clear with respect to preservation of function [12].

A second class of sequence-based function classification approaches have evolved from early work on protein pattern recognition which suggested that short sequences of amino acids (*motifs*) may be conserved in a protein family [15]. Currently, motif composition is often used to assign putative functions to novel protein sequences based on the known functions of other proteins that share one or more motifs with the novel protein. Several motif databases have

been developed, including those that contain relatively short *motifs*, e.g., PROSITE [18]; or groups of motifs referred to as *fingerprints*, e.g., PRINTS [3], or BLOCKS [20]; or sequence patterns, often based on position-specific scoring matrices or hidden Markov models generated from multiple sequence alignments e.g., called *profiles*, PROSITE [18] or *domains*, e.g., Pfam [9]. Such motif databases or resources that integrate several databases, e.g., InterPro [2], MetaFam [36], can be queried using a protein sequence to obtain a list of motifs that are found in the sequence as well as the functions or structures associated with these motifs.

Several automated tools for generating a set of motifs that capture conserved sequence regularities among a given set of sequences are available (see [22] for a review). They fall into two broad classes. The first class of methods relies on (typically local) multiple sequence alignment to extract conserved patterns among set of (functionally) related sequences, such as MEME (Multiple Expectation Maximization for Motif Elicitation) [4]. A second class of methods uses a combinatorial approach to build a dictionary of motifs from a given set of sequences without making any assumptions about the functional family memberships of the sequences in question [33]. The latter are especially useful for extracting sequence regularities among divergent families. Motifs or sequence patterns distill information from groups of related sequences to facilitate detection of weaker sequence similarities. Therefore, pattern based searches are often more sensitive and selective than sequence database searches. For example, Jaakkola et al. [23] have shown that HMM profiles generated from local alignment of sequence fragments can be used to build classifiers that can help identify distantly related sequences (where sequence similarity is less than 30%).

Motif-based techniques for protein function prediction focus similarity searches on parts of the protein that are likely to be functionally or structurally significant, and hence more likely to be conserved. Current motif-based approaches to protein function prediction are not without drawbacks. Many proteins contain several motifs and the same motif may be found in proteins belonging to several different functional families. More generally, it may be necessary to identify combinations of motifs that must present, or perhaps even absent in a sequence, in order to reliably assign it to a functional family. Indeed, in the PRINTS database [3], the fingerprints used to assign proteins to functional families can be simple motifs or a combination of motifs. However, the process of identifying a fingerprint for each protein family of interest can be labor intensive and requires considerable domain knowledge. Thus, there is a need for sophisticated tools that *automate* the discovery of sequence regularities predictive of protein function and allow efficient updating of databases.

Proteins with similar three-dimensional structural features very often, but not always, have similar functions because the shape of the protein both constrains and facilitates the ways in which the protein can interact with

substrates, ligands, or other proteins. Hence, if the structure of a protein is known, one might assign a putative function to it on the basis of its structural similarity to a known structure [28]. Several algorithms have been developed for recognizing structurally related proteins e.g., [21], accompanied by the establishment of a number of structure databases and structural class databases such as PDB [10], SCOP [25], CATH [29], and DALI [21]. However, experimental determination of protein structures using NMR or X-ray crystallography techniques is time consuming and expensive. While there are 254,293 protein records in PIR-NREF database [7] (Release 1.05, 9 September 2002) contains 1,011,453 entries, there are only 14,339 experimentally determined three-dimensional protein structures in the Protein Data Bank (PDB) [10] contains 18,691 structures, corresponding to approximately 3000 different proteins (as of 10 September 2002). Hence, protein function prediction often relies on protein structure prediction using computational approaches. Ab initio methods that predict the conformation of a protein from its amino acid sequence are computationally very demanding and are currently limited to relatively short proteins or peptides [35]. A number of structure-based approaches to function determination are therefore focused on identification of functionally significant structural elements (e.g., active sites, binding sites) of proteins [5]. A recent study by Fetrow et al. [19] has shown that a sequence-to-structure-to-function paradigm that exploits knowledge of functionally relevant three-dimensional structural elements together with sequence information significantly improves the accuracy of function annotation of disulphide oxidoreductases in *S. cervisiae*. However, experimental determination of functionally relevant structural features is time consuming and expensive. With the accumulation of known structures, there is both the potential as well as a need for data-driven computational methods for identification of functionally meaningful structural features (and their sequence correlates) that can serve as reliable predictors of function.

In this paper, we test the feasibility of a fully automated approach for protein function classification. We present a data-driven approach to discovery of rules for assigning protein sequences to functional families on the basis of the presence or absence of specific motifs or combinations of motifs. (For simplicity, we will use the term *motif* to include short conserved sequence patterns as well as profiles.) Machine learning algorithms [26] offer one the most cost effective approaches to automated discovery of a priori unknown predictive relationships from large data sets in computational biology [6]. Decision tree induction algorithms are relatively fast, and produce rules that are easy to interpret. Machine learning approaches have been previously used for protein function classification. For example, King et al. [24] investigated an inductive logic programming approach to the construction of protein function classifiers using alternative representations of protein sequences (amino acid residue frequencies, phylogeny, and predicted structure). In a previous study,

we used the C4.5 family of decision tree induction algorithms [31] to discover rules for protein classification on the basis of presence or absence of combinations of PROSITE motifs with encouraging results [38]. The study demonstrated, for several protein families, that decision tree classifiers generated using PROSITE patterns and motifs can provide more accurate protein family classification than the use of a single characteristic motif. PROSITE patterns are usually fairly short (less than 20 amino acids) and typically correspond to biologically significant sites experimentally identified in PROSITE functional families. PROSITE profiles, on the other hand, correspond to Hidden Markov models that usually match longer sequence fragments (often over 100 amino acids). These longer profiles are useful as "signatures" for protein families, but make it difficult to identify underlying sequence regularities that are predictive of protein function, or may correspond to biologically significant structural features.

Here we explore whether it is possible to use relatively short, automatically generated motifs to discover rules for protein classification. In this study, we used MEME (Multiple Expectation Maximization for Motif Elicitation) a motif discovery program that can be used to automate the construction of motif databases from any given set of sequences [4]. We also explore the use of the resulting classifiers as a source of information about the sequence correlates of functionally significant structural features of proteins. For our data set, we chose a well-characterized subset of protein families from the MEROPS protease database [Release 5.4 23 March 2000] [32]. We compared rules discovered based on motifs automatically generated using MEME with those generated based on PROSITE patterns and profiles [18]. Further, we investigated the ability of decision trees to identify functionally significant structural features of proteins using the caspase protease family as a test case.

2. Data-driven discovery of rules for protein function classification using sequence motifs

The basic computational problem is the following: Given a database or training set of amino acid sequences corresponding to proteins with known (i.e., experimentally determined) function, our goal is to induce a classifier that would be able to assign novel protein sequences to one of the protein families represented in the training set. The general approach is illustrated in Fig. 1.

2.1. Data representation

A majority of algorithms for data-driven induction of pattern classifiers represent instances to be classified using a fixed set of *attributes*. Hence, we first map each protein sequence into a corresponding *attribute-based representation*

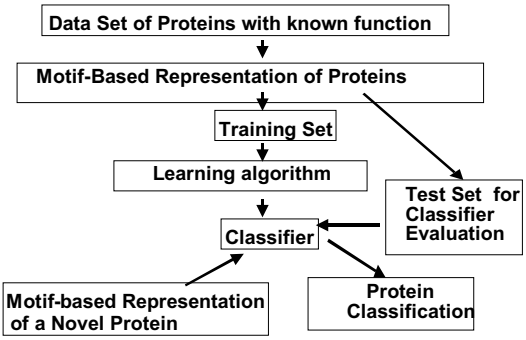

Fig. 1. Data mining approach to motif-based protein function classification.

[38]. The choice of attributes plays a critical role in the data mining process. Here, we represent protein sequences using a suitable *vocabulary* of *sequence motifs*. The set of motifs to be used can be chosen to correspond to one of the existing motif databases (e.g., PROSITE) or the set of motifs identified by running a suitable motif-finding program (e.g., MEME) on the set of protein sequences. Suppose the vocabulary contains N motifs. Any given sequence typically contains a few of these motifs. We encode each sequence as an N-bit binary pattern where the ith bit is 1 if the corresponding motif is present in the sequence; otherwise the corresponding bit is 0. Each N-bit sequence is associated with a *label* which identifies the functional family of the sequence (if known). A training set is simply a collection of N-bit binary patterns, each of which has associated with it a label that identifies the functional family of the corresponding protein. This training set is used to train a classifier which can then be used to assign novel sequences to one of the several functional families represented in the training set.

2.2. Data set

A subset of the peptidase (protease) families classified according to the MEROPS (Release 5.5 15 June 2000) two-level classification system [32] was used in this study. The choice of the peptidase families was motivated by the diversity of the proteins in the family and the fact that many of them are well-characterized and have known structures and functions [8]. The MEROPS database (http://www.merops.co.uk/) classifies proteases into functional families and clans. Clans are groupings of evolutionarily related functional families. This classification structure permits analysis of the performance of protein classifiers at two levels of sequence diversity.

For this study, all MEROPS-defined protease families that had more than two protein members and belonged to a clan were chosen. Clans with fewer

than two member proteins were excluded from the data set. Protein sequences that were only fragments were removed. The resulting dataset consisted of 84 families (out of a total of 161 in MEROPS) with between 3 and 313 members, and 19 clans with between 1 and 18 families. In order to avoid excessive bias in favor of large families (i.e., those consisting of a large number of members with high levels of sequence identity), the PURGE program [27] was used to select sequences from large families. This resulted in a data set of 1627 proteins. MEME motifs were extracted from each of the 84 families of proteins. The data set used in the study can be obtained by contacting xiangyun.wang@astrazeneca.com.

2.3. Motif-based representation of the protein sequences

Decision trees were constructed using motif-based representation of sequences generated using two different sources of motifs:

- A database of motifs generated by the MEME (Multiple Expectation Maximization for Motif Elicitation) program [4] for each peptidase family used in the study. MEME was chosen as a representative of automated motif identification programs because of its ability to identify motifs among highly divergent sequences [22]. The MAST (Motif Alignment and Search Tool) program was used to determine the motif composition of a sequence. Several perl scripts were used to transform the MAST output into the appropriate format for the C4.5 program.
- Motifs and profiles from PROSITE, which is one of the most carefully curated motif databases [18]. The PROSITE database associates with each functional family, a *characteristic* motif or HMM profile which can be used to identify members of the family. ProfileScan (available from PROSITE) was used to identify PROSITE motifs or profiles (with a length of at least five amino acids) in each peptidase sequence.

2.4. Decision tree algorithm

We used the C4.5 decision tree algorithm [31] for building protein sequence classifiers. C4.5 uses a greedy procedure that selects the attributes that yield the maximum information gain to recursively partition the training set. It also uses post-pruning to compensate for any over fitting that may have occurred. The decision trees generated were evaluated using 5-fold cross-validation (i.e., five independent runs using 80% of the data for training and the remaining 20% for testing). Decision trees produced were then converted into rules for further analysis. Each rule is of the form "if *condition* then *class*" where *condition* checks for a motif combination whose presence or absence is a reliable predictor for the corresponding *class* (e.g., protein family).

3. Experiments and results

The computational experiments were designed to address the following question: How does the performance of protein function classifiers based on motifs generated *automatically* using a multiple sequence alignment based motif discovery tool such as MEME compare with that of classifiers based on motifs from the expert-curated PROSITE database? Can classification rules generated using this approach pick out a small subset of structurally or functionally significant sequence features (motifs) from among a large set of candidate motifs?

A representative decision tree and the corresponding rules are shown in Fig. 2.

Two key measures of classifier performance were used in this study: *Precision* and *Recall.* In intuitive terms, *precision* of a classifier (whether it is a rule that checks for presence of a single motif or applies a more complex rule) measures the degree to which the classifier is able to pick out members of a class of interest while rejecting all other instances. *Recall* measures the extent to which the classifier is able to identify all members of the class of interest (perhaps at risk of including some instances that do not belong to the class).

An instance assigned by a classifier to a specific class is said to be a *true positive* with respect to that class if it in fact belongs to that class. An instance is said to be a false positive with respect to a class if it is assigned to that class by the classifier, but in fact belongs to a different class. True negatives and false negatives can be defined in an analogous fashion. Let α be a classifier and c a class. Let $\mathrm{TP}_\alpha(c)$, $\mathrm{TN}_\alpha(c)$, $\mathrm{FP}_\alpha(c)$, and $\mathrm{FN}_\alpha(c)$ respectively be the number of true positives, true negatives, false positives, and false negatives produced by the classifier α for class c on a given test set. Then the precision of classifier α on class c (estimated using the given test set) is given by $\mathrm{TP}_\alpha(c)/(\mathrm{TP}_\alpha(c)+\mathrm{FP}_\alpha(c))$ and recall by $\mathrm{TP}_\alpha(c)/(\mathrm{TP}_\alpha(c)+\mathrm{FN}_\alpha(c))$. The *accuracy* of the classifier α for class c is estimated by $(\mathrm{TP}_\alpha(c)+\mathrm{TN}_\alpha(c))/N$ where N is the total number of instances tested. In our study, estimates were averaged over 5-fold cross-validation runs. Note that an ideal classifier has both precision and recall of 1 for each class. Overall precision and recall for a classifier can be obtained by calculating the overall average of precision and recall for all classes.

In the experiments described below, we performed two types of comparison. First, in separate experiments using either MEME and PROSITE motifs, we compared the accuracy, precision and recall of decision tree classifiers based on a *combination* of motifs with that of classifiers which use the presence or absence of the *single best motif* for each class as the only criterion for classification. We define the *single best motif* for a family or clan as the motif with the highest value for the product of precision and recall for that family or clan using the entire data set. This scoring was used because having high recall and

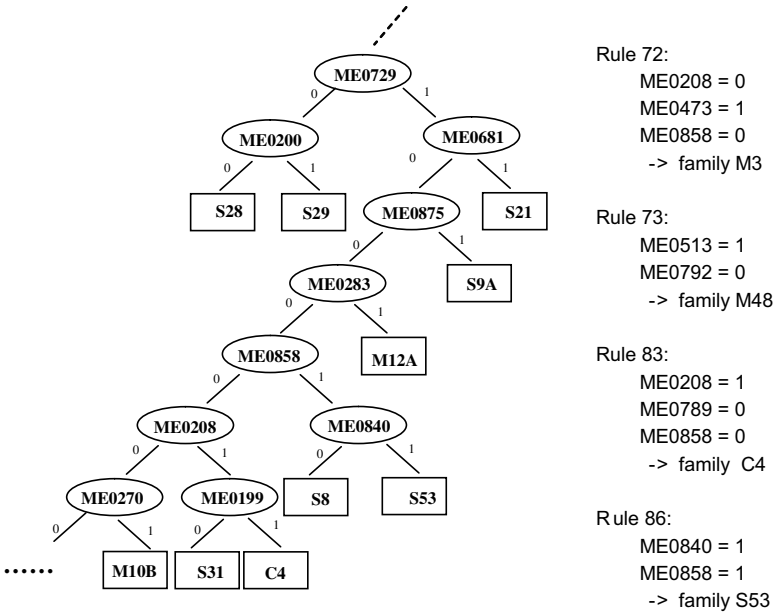

Fig. 2. A decision tree and rule sets for classifying protease sequences into functional families. The left panel shows a portion of typical decision tree. Each oval represents an internal node for testing the presence (1) or absence (0) of a MEME-generated motif (e.g., ME0840). Rectangles represent leaf nodes which indicate the MEROPS family to which the sequence was assigned (e.g., S53, a serine protease family). The right panel shows corresponding rule sets for several families. For example, Rule 86 states that if a sequence contains both motifs ME0840 and ME0858, it belongs to the S53 family.

low precision (and vice versa) is useless for classification. In each case, we compared the performance precision, recall, and accuracy of rules derived from decision tree classifiers with that of the single best motif. Second, we compared the performance of rules based on automatically generated MEME motifs (using empirically determined optimal parameter settings) with that of rules based on motifs from the expert-curated PROSITE database.

Classification performance results obtained using decision tree classifiers based on MEME motifs vs. classification using single best motifs are shown in Table 1. In assigning protein sequences to *families*, rules extracted from decision tree classifiers had accuracy comparable to that of single best motifs (91.8% vs. 91.0%) but precision higher than that of single best motifs (94.6% vs. 85.2%). On the other hand, the recall of the rules was somewhat lower than that of single best motifs (92.8% vs. 96.4%). In assigning sequences to *clans*, decision tree classifiers performed significantly better in assigning peptidases to MEROPS clans than single best motifs, both in terms of accuracy (90.4% vs. 43.1%) and recall (90.8% vs. 65.2%), and had comparable precision (92.0% vs.

Table 1
Comparison of classification performance of family and clan rules based on MEME motifs with that of single best motifs, i.e., motifs with the largest (precision × recall)

	MEME motifs per rule set	Accuracy (%)		Precision (%)		Recall (%)	
		Rules	Best motif	Rules	Best motif	Rules	Best motif
Families	2.9	91.8	91.0	94.6	85.2	92.9	96.4
Clans	9.2	90.4	43.1	92.0	88.5	90.8	65.2

Column 1 shows the average number of motifs per rule. The percentage precision and recall figures for family (clan) correspond to averages taken over families (clans). Percentage accuracy is computed over the entire test sample. All of the results represent estimates based on 5-fold cross-validation.

88.5%). These results show that the decision tree algorithm is able to successfully identify combinations of motifs that capture shared features of diverse sets of proteins belonging to a clan. In contrast, families tend to comprise more closely related sequences than clans, and hence, single best motifs often perform as well as rules.

3.1. Decision trees using PROSITE motifs

Table 2 shows results of an analogous set of experiments carried out using PROSITE motifs to build decision tree classifiers. For *families*, rules generated by decision tree classifiers had somewhat lower accuracy (77.4%) than single best motifs (84.9%), but higher precision (88.9% vs. 75.7%) and recall (84.4% vs. 81.0%). For *clans*, decision tree classifiers performed significantly better than single best motifs, in terms of accuracy (88.0% vs. 75.3%), precision (98.4% vs. 92.9%) and recall (83.2% vs. 73.5%). Closer examination of the rule set for MEROPS family S1 with 6–7 motifs on average in its rule shows that the rule set significantly outperforms single best motif in terms of accuracy (98.3% vs. 89.7%), precision (97.6% vs. 89.7%), and recall (97.0% vs. 89.7%). These

Table 2
Comparison of classification performance of family and clan rules based on PROSITE motifs with that of single best motifs, i.e., motifs with the largest (precision × recall)

	PROSITE motifs per rule set	Accuracy (%)		Precision (%)		Recall (%)	
		Rules	Best motif	Rules	Best motif	Rules	Best motif
Families	2.9	77.4	84.9	88.9	75.7	84.4	81.0
Clans	12.0	88.0	75.3	98.4	92.9	83.2	73.5

Column 1 shows the average number of motifs (checked for presence or absence) per rule. The percentage precision and recall figures for family (clan) are correspond to averages taken over families (clans). Percentage accuracy is computed over the entire test sample. All of the results represent estimates based on 5-fold cross-validation.

results are consistent with the pattern observed in the case of MEME motifs: the more diverse the set of sequences that are to be assigned to a group the more likely it is that rules outperform single best motifs.

Comparison of data shown in Tables 1 and 2 indicates that the rules constructed for classifying peptidase sequences into the corresponding MEROPS *families* using MEME mofis performed better (avg. accuracy 91.8%, precision 94.6% and recall 92.9%) than rules constructed using PROSITE motifs (avg. accuracy 77.4%, precision 88.9% and recall 84.4%). Thus, using the parameter settings chosen in this study, MEME motifs appear better suited for distinguishing closely related members of peptidase families (according to MEROPS classification) than the available PROSITE motifs. This is not surprising because PROSITE does not include characteristic motifs for many of the peptidase families used in our data set. Motifs in PROSITE are limited to those that have been identified as characteristic *signatures* of specific protein families. In contrast, because MEME motifs are generated automatically from a given set of sequences, they can capture a broader range of regularities among the chosen set of sequences. Therefore, decision trees and rules constructed using MEME motifs may have more flexibility than rules constructed using PROSITE motifs for characterizing functional families at different hierarchical classification levels.

The performance of classification rules for assigning peptidases to MEROPS *clans* (each clan typically contains several related families) constructed using MEME motifs (avg. accuracy 90.4%, precision 92.0% and recall 90.8%) was comparable to that of rules constructed using PROSITE motifs (average accuracy 88.0%, precision 98.4% and recall 83.2%). The classification rules constructed from PROSITE motifs used more motifs (12.0 per clan) than the rules constructed from MEME motifs (avg. 9.2 per clan).

3.2. Performance of rules based on MEME motifs on MEROPS families with a corresponding PROSITE classification

In light of the preceding discussion, it is interesting to ask: How does a fully automated method of constructing decision tree classifiers for assigning protein sequences to functional families (using MEME motifs) compare with an approach that relies on motifs that have been identified and associated with specific functional families using great deal of expert knowledge in the case of functional families that are represented in PROSITE. Table 3 shows how the rule sets performed on the subset of families and clans that had a corresponding PROSITE family.

One difficulty in comparing the performance of classifiers constructed using PROSITE motifs with those constructed using MEME motifs is that ProfileScan (used to identify PROSITE motifs in data sets) and MAST (used to identify MEME motifs) use different types of parameters to control the stringency of

Table 3
Performance of family rule sets for the subset of MEROPS families with a corresponding PROSITE family (MEROPS families S1 S2B S10 S14 S16 C12 C14 C15 C19 A8 M17 M24B M22)

		PROSITE	MEME
Accuracy (%)	Best motif	92.0	91.4
	Rule set	95.7	85.3
Precision (%)	Best motif	99.8	91.1
	Rule set	99.8	95.6
Recall (%)	Best motif	92.0	93.7
	Rule set	93.1	88.9
Motifs per rule set		1.6	1.2

All entries represent averages over the families. Percentage accuracy is computed over the entire test sample. All of the results represent estimates based on 5-fold cross-validation.

motif matches. ProfileScan offers two choices for motif matching: *weak match* and *strong match*. We used the weak match setting based on preliminary experiments which showed that it yielded better results for the peptidase data set. The performance of MAST is sensitive to *p*-value used. For example, using *p*-value of 0.0001, the MEME motif based decision tree accuracy was 78%, but reducing the *p*-value by a factor of 10 caused the MEME-based decision tree accuracy to increase to 96%.

3.3. Structural and functional significance of the classification rules constructed using MEME motifs

The results presented in previous sections show that decision trees constructed using relatively short (12 amino acids long) motifs can classify peptidase sequences into MEROPS families and clans with high accuracy, precision, and recall. This suggests the possibility that the resulting *automatically generated* classification rules capture sequence regularities that correspond to structurally and/or functionally significant aspects of protein structure. Hence, it is interesting to examine the motifs frequently used in decision tree rule sets in the context of the three-dimensional structure of several representatives of peptidase families with known structures and functions. We chose to examine the C14 family (Caspase family) for which three-dimensional structural information is available in the PDB database [10]. Caspases play critical roles in programmed cell death or apoptosis [16]. The structure of a representative member of the C14 family, human Caspase 1 (PDB ID: 1BMQ), is shown in Fig. 3(a). Two key catalytic residues of the enzyme are His237 and Cys285; Arg179 and Arg341 contribute to substrate binding. Mutations of either of the two catalytic residues or Arg179 have been shown to abolish caspase activity [39].

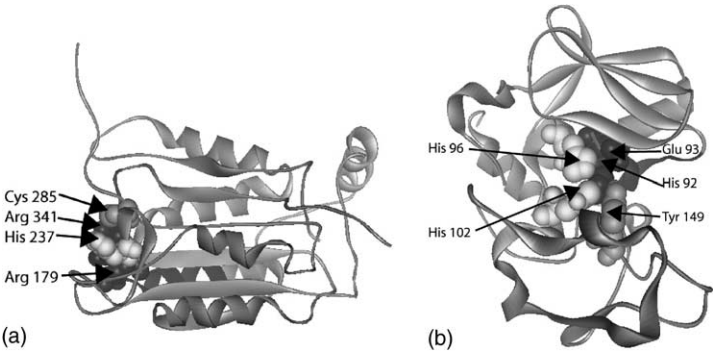

Fig. 3. (a) The three-dimensional structure of human Caspase-1 (MEROPS family C14), corresponding to PDB entry 1BMQ. The four labeled residues Arg 179, His 237, Cys 285, and Arg 341 are known to form the substrate binding pocket of the Caspase-1 enzyme [39]. Three of these residues (arg 179, His 237, and Cys 285) are located within the MEME-generated motifs frequently used by the decision tree classifier for the MEROPS family C14. These, motifs correspond to residues 179–190 (red), 228–239 (yellow), 276–287 (green). (b) The three-dimensional structure of Astacin (MEROPS family M12) from *A. astacus*, corresponding tp PDB entry 1QJJ. Five MEME-generated motifs selected by the decision tree algorithm for the MEROPS family M12 correspond to residues 83–94 (red), 96–107 (yellow), and 142–153 (green). The five labeled residues—His 92, His 96, Glu 93, His 102, Tyr149 that appear within the motifs have been shown to form the zinc binding pocket of the enzyme [11].

Examination of rule sets for the C14 family constructed based on MEME motifs (using maximum motif lengths ranging from 6 to 50 and a p-value of 10^{-5}) revealed that four motifs were found in 80% of the rules generated from 10 independent runs. When we examined the location of these motifs within the human Caspase-1 protein structure, we found that three of the four motifs are close to each other in the three-dimensional structure (Fig. 3(a)), and each of these three motifs contained one of three residues involved in the catalytic activity of Caspase-1: Arg179, His237 and Cys285. When we examined the rules constructed from motifs generated using different values for the maximum motif length parameter, we found that approximately 69% of the motifs most frequently appearing in the rule sets include active site residues. The presence of one or two of these three motifs was found to be sufficient to reliably separate Caspases from all the other peptidases families. The motif that covers Cys285 corresponds to the top ranked motif in caspase family in the output of the MEME program.

Similarly, examination of the rule sets generated for the MEROPS family M12 in relation to the three-dimensional structure of Astacin (PDB entry 1QJJ) showed that five residues that have been shown to form the zinc binding pocket His 92, His 96, Glu 93, His 102, Tyr 149 [11] are contained in the motifs most frequently used by the decision tree algorithm (see Fig. 3(b)).

Table 4
Percentage of active site motifs among the motifs used in the rule sets

Peptidase family	PDB ID	% of motifs with active site in the rule sets
M24A	1MAT	98
M20A	1CG2	87
M24B	1A16	83
M12A	1QJJ	78
C15	1A2Z	68
S21	1LAY	68
C14	1ICE	66
S8A	1BE6	64
M10A	2TCL	59
C12	1UCh	57
C1	1YAL	46
A2	4UPJ	43
S1A	2GMT	43
S14	1TYF	41
S26	1B12	39
C5	1AVP	37
M17	1BLL	33
S10	1CPY	30
A1	1F34	29
M10B	1AFO	28
M12B	1DTH	26
C2	1DFO	<25
C3	1HAV	<25
M13	1DMT	<25
M27	1F82	<25
M4	1TLP	<25
M8	1LML	<25
S3	2SNV	<25
S24	1UMU	<25
S29	1A1R	<25
A6	1F8V	<25

To investigate how often motifs that contain active sites show up in the rule sets, we examined a total of 31 MEROPS families with known active sites. Active site information was extracted from the MEROPS database. Table 4 shows the percentage of the motifs in the rule sets that correspond to a known active site for each of the families. For 10 of the families, the active site motifs account for more than 50% of the motifs used in the rule sets, with average of 73%. In 11 families, active site motifs account for 25–50% of the motifs in the rule sets with an average of 36%. In the remaining 10 families, the active site motifs account for less than 25% of the motifs used in the rule sets. When active sites are highly conserved and unique for a family, they provide a reliable source of information for discriminating that family from other families.

However, this is not always the case. Families that appear to have a common evolutionary origin (e.g., families belonging to the same clan) often have similar active sites. Thus, it is necessary to use information other than the active site motifs for telling such families apart. The families with relatively high fractions of active motifs in the rule sets often belong to clans with fewer families. For example, C15 is the sole member of clan CF, C14 is the only member from clan CD in the dataset, M24A and M24B are the only members of clan MG. Family M24B has the highest presence of active site motifs (98%) in its rule sets.

4. Summary and future directions

In this paper, we investigated the feasibility of a fully automated approach for protein function classification. In summary, we found that:

- Decision trees built using a motif-based representation of protein sequences constructed using MEME outperform decision trees constructed using PROSITE motifs in classifying proteases into corresponding MEROPS *families*.
- Decision tree classifier for clans significantly outperformed single best motif (defined as one having the largest product of precision and recall for a given clan): The difference in performance between decision trees and single best motifs was less dramatic in the case of families. Examination of the results for individual families showed that the more diverse of the sequences in a functional family, the greater the performance advantage offered by the decision trees.
- In several examples of proteins with known structure, the decision tree algorithm was able to identify combinations of motifs from different parts of the sequence that clustered together in three-dimensional structure and corresponded to a functionally significant structural feature (e.g., binding site) (see Fig. 3(a) and (b)). This is especially intriguing in light of the fact that no biological expertise or knowledge was used in identifying the motifs (other than the amino acid substitution matrix used by MEME) or in constructing the rules (other than the MEROPS family labels for the sequences in the training set).

The results presented in this paper have shown that a motif discovery algorithm such as MEME can provide a source of sequence features for automated, data-driven construction of decision trees or rules for classifying proteins into relevant functional families. These results suggest that the rules constructed using MEME motifs are especially good at characterizing sequence regularities (in the form of relatively short conserved sequence patterns) associated with closely related functional families. Thus, when adequate training data are available, data-driven discovery of protein sequence–function

relationships using automated motif identification and machine learning appears to complement if not offer a viable high throughput alternative for assigning putative functions to novel proteins by labor intensive expert annotation.

Translating the recent advances in high throughput data acquisition technologies in biological sciences into fundamental gains in scientific understanding of biological processes calls for the development of sophisticated computational tools for characterization and prediction of macromolecular structure–function relationships. The results reported here raise the possibility of using techniques similar to those employed in this study to explore and characterize protein structure–function relationships at multiple levels. More extensive studies with a broader range of proteins are needed to rigorously test whether rules constructed using decision trees or other similar machine learning algorithms can, using a purely data-driven automated approach, identify the sequence correlates of functionally significant three-dimensional structural features of proteins. It is intriguing to consider whether knowledge of such relationships mined from the data can be effectively incorporated into ab initio approaches to structure prediction.

References

[1] S.F. Altschul, T.L. Madden, A.A. Schaffer, D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (1997) 3389–3402.

[2] R. Apweiler, T.K. Attwood, A. Bairoch, A. Bateman, E. Birney, M. Biswas, P. Bucher, L. Cerutti, F. Corpet, M.D.R. Croning, R. Durbin, L. Falquet, W. Fleishchmann, J. Gouzy, H. Hermjakob, N. Hulo, I. Jonassen, D. Kahn, A. Kanapin, Y. Karavidopoulou, R. Lopez, B. Marx, N.J. Mulder, T.M. Oinn, M. Pagni, F. Servant, C.J.A. Sigrist, E.M. Zdobnov, The InterPro database, an integrated documentation resource for protein families, domains and functional sites, Nucleic Acids Res. 29 (2001) 37–40.

[3] T.K. Attwood, M.D.R. Croning, D.R. Flower, A.P. Lewis, J.E. Mabey, P. Scordis, J. Selley, W. Wright, PRINT-S: the database formerly known as PRINTS, Nucleic Acids Res. 28 (2000) 225–227. Available from <http://bioinf.man.ac.uk/dbbrowser/prints/>.

[4] T.L. Bailey, M.E. Baker, C.P. Elkan, W.N. Grundy, MEME, MAST, and Meta-MEME: new tools for motif discovery in protein sequences, in: Pattern Discovery in Biomolecular Data, Oxford University Press, Oxford, 1999, pp. 30–54.

[5] D. Baker, A. Sali, Protein structure prediction and structural genomics, Science 294 (2001) 93–96.

[6] P.F. Baldi, S. Brunak, Bioinformatics: The Machine Learning Approach, The MIT Press, Cambridge, MA, 2001.

[7] W.C. Barker, J.C. Garavelli, Z. Hou, H. Huang, R.S. Ledley, M.P. McGarvey, H.W. Mewes, B.C. Orcutt, F. Pfeiffer, A. Tsugita, C.R. Vinayaka, C. Xiao, L.L. Yeh, Wu. Cathy, Protein information resource: a community resource for expert annotation of protein data, Nucleic Acids Res. 29 (2001) 29–32. Available from <http://pir.georgetown.edu/>.

[8] A.J. Barrett, N.D. Rawlings, J.F. Woessner, Handbook of Proteolytic Enzymes, Academic Press, New York, 1998.

[9] A. Bateman, E. Birney, R. Durbin, S.R. Eddy, K.L. Howem, E.L.L. Sonnhammer, The Pfam protein families database, Nucleic Acids Res. 28 (2000) 263–266.
[10] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The protein data bank, Nucleic Acids Res. 28 (2000) 235–242. Available from <http://www.rcsb.org/pdb/>.
[11] J.S. Bond, R.J. Beynon, The astacin family of metalloendopeptidases, Protein Sci. 4 (1995) 1247–1261.
[12] P. Bork, C. Ouzounis, C. Sander, From genome sequences to protein function, Curr. Opin. Struct. Biol. 4 (1994) 393–403.
[13] C. Chothia, M. Gerstein, Protein evolution. How far can sequences diverge? Nature 85 (6617) (1997) 579–581.
[14] C. Chothia, A.M. Lesk, The relation between the divergence of sequence and structure in proteins, EMBO J. 5 (4) (1986) 823–826.
[15] M.O. Dayhoff, W.C. Barker, L.T. Hunt, Establishing homologies in protein sequences, Methods Enzymol. 91 (1983) 524.
[16] W.C. Earnshaw, L.M. Martins, S.H. Kaufmann, Mammalian caspases: structure, activation, substrates, and functions durng apoptosis, Annu. Rev. Biochem. 68 (1999) 383–424.
[17] D. Eisenberg, E.M. Marcotte, I. Xenarios, T.O. Yeates, Protein function in the post-genomic era, Nature 405 (6788) (2000) 823–826.
[18] L. Falquet, M. Pagni, P. Bucher, N. Hulo, C.J. Sigrist, K. Hofmann, A. Bairoch, The PROSITE database, its status in 2002, Nucleic Acids Res. 30 (2002) 235–238. Available from <http://www.expasy.ch/prosite/>.
[19] J.S. Fetrow, N. Siew, J.A. Di Gennaro, M. Martinez-Yamout, H.J. Dyson, J. Skolnick, Genomic-scale comparison of sequence- and structure-based methods of functionprediction: does structure provide additional insight? Protein Sci. 10 (5) (2001) 1005–1014;
L. Holm, C. Sander, Dali/FSSP classification of three-dimensional protein folds, Nucleic Acids Res. 25 (1) (1997) 231–234.
[20] J.G. Henikoff, E.A. Greene, S. Pietrokovski, S. Henikoff, Increased coverage of protein families with the blocks database servers, Nucl. Acids Res. 28 (2000) 228–230.
[21] L. Holm, C. Sander, Mapping the protein universe, Science 273 (1996) 595–602.
[22] J. Hudak, M.A. McClure, A comparative analysis of computational motif detection methods, Pacific Symp. Biocomput. 4 (1999) 138–149.
[23] T. Jaakkola, M. Diekhans, D. Haussler, A discriminative framework for detecting remote protein homologies, J. Comp. Biol. 7 (2000) 95–114.
[24] R.D. King, A. Karwath, A. Clare, L. Dehaspe, The utility of different representations of protein sequence for predicting functional class, Bioinformatics 17 (2001) 445–454.
[25] L. Lo Conte, B. Ailey, T.J. Hubbard, S.E. Brenner, A.G. Murzin, C. Chothia, SCOP: a structural classification of proteins database, Nucleic Acids Res. 28 (1) (2000) 257–259.
[26] T. Mitchell, Machine Learning, McGraw Hill, New York, 1997.
[27] A.F. Neuwald, J.S. Liu, C.E. Lawrence, Gibbs motif sampling: detection of bacterial outer membrane protein repeats, Protein Sci. 4 (1995) 1618–1632.
[28] C.A. Orengo, A.E. Todd, J.M. Thornton, From protein structure to function, Curr. Opin. Struct. Biol. 9 (3) (1999) 374–382.
[29] F.M. Pearl, N. Martin, J.E. Bray, D.W. Buchan, A.P. Harrison, D. Lee, G.A. Reeves, A.J. Shepherd, I. Sillitoe, A.E. Todd, J.M. Thornton, C.A. Orengo, A rapid classification protocol for the CATH Domain Database to support structural genomics, Nucleic Acids Res. 29 (1) (2001) 223–227.
[30] W.R. Pearson, Flexible sequence similarity searching with the FASTA3 program package, Methods Mol. Biol. 132 (2000) 185–219.
[31] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo, CA, 1992.

[32] N.D. Rawlings, A.J. Barrett, MEROPS: the peptidase database, Nucleic Acids Res. 28 (2002) 323–325. Available from <http://www.merops.co.uk/>.
[33] I. Rigoutsos, A. Floratos, C. Ouzounis, Y. Gao, L. Parida, Dictionary building via unsupervised hierarchical motif discovery in the sequence space of natural proteins, Proteins 37 (2) (1999) 264–277.
[34] B. Rost, Twilight zone of protein sequence alignments, Protein Eng. 12 (2) (1999) 85–94.
[35] R. Samudrala, Y. Xia, M. Levitt, E.S. Huang, A combined approach for ab initio construction of low resolution protein tertiary structures from sequence, Pacific Symp. Biocomput. 4 (1999) 505–516.
[36] K.A.T. Silverstein, E. Shoop, J.E. Johnson, E.F. Retzel, MetaFam: a unified classification of protein families. I. Overview and statistics, Bioinformatics 17 (2001) 249–261.
[37] J. Skolnick, J.S. Fetrow, From genes to protein structure and function: novel applications of computational approaches in the genomic era, Trends Biotechnol. 18 (1) (2000) 34–39.
[38] D. Wang, X. Wang, V. Honavar, D. Dobbs, Data-driven generation of decision trees for motif-based assignment of protein sequences to functional families, in: Proceedings of the Atlantic Symposium on Computational Biology, Genome Information Systems & Technology, 2001.
[39] K.P. Wilson, J.F. Black, J.A. Thomson, E.E. Kim, J.P. Griffith, M.A. Navia, M.A. Murcko, S.P. Chambers, R.A. Aldape, S.A. Raybuck, D.J. Livingston, Structure and mechanism of Interleukin-1β converting enzyme, Nature 370 (1994) 270–275.