Data-Driven Generation of Decision Trees for Motif-Based Assignment of Protein Sequences to Functional Families

Dake Wang, Xiangyun Wang, and Vasant Honavar
Artificial Intelligence Research Laboratory
Department of Computer Science and
Graduate Program in Bioinformatics and Computational Biology
Iowa State University, Ames, IA 50011, USA

Drena L. Dobbs
Department of Zoology and Genetics and
Graduate Program in Bioinformatics and Computational Biology
Iowa State University, Ames, IA 50011, USA

This paper describes an approach to data-driven discovery of sequence motif-based models in the form of decision trees for assigning protein sequences to functional families. Unlike approaches that try to classify protein sequences based on presence of a single motif, 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 sequences with known functions is used to automatically construct decision trees that capture regularities that are sufficient to assign the sequences to their respective functional families. The accuracy of the resulting decision tree classifiers are then evaluated on an independent test set. Experiments using several protein data sets indicate that proposed approach matches or beats the technique of assigning protein sequences to functional families based on the presence of a single characteristic motif in terms of the accuracy of resulting classification.

Introduction

Function prediction of unknown proteins remains one of the most challenging problems in functional genomics. Early work on protein pattern recognition [Dayhoff et al., 1983] suggested that subsequences of amino acids may be conserved in a protein family. More recently, a variety of computational approaches have been developed for identification of short sequences of amino acids that are conserved within a family of closely related protein sequences. The interested reader is referred to [Hudak and McClure, 1999] for a comparison of several such motif detection methods). Since protein function is often correlated with highly conserved motifs, such motifs have recently been used in protein function prediction. Consequently, several databases have been developed that contain motifs or motif clusters. Examples of such databases include: Prosite [Hoffman et al., 1999], Pfam [Bateman et al., 2000], and Prints [Attwood and Beck, 1994] databases. When queried with a motif, the database returns a function associated with that motif, providing clues of possible functions of the protein containing that motif. However, many proteins usually contain more than one motif. It may in general be necessary to identify combinations of several motifs that need to be present (or even absent) in order to reliably assign a sequence to a functional family. Thus, there is a need for sophisticated tools for discovery of such sequence regularities that are predictive of protein function. Machine learning or data mining algorithms offer one of the most effective and practical apporaches to discovery of such a-priori unknown, predictive relationships from data. A variety of machine learning algorithms have been developed in the artificial intelligence and pattern recognition literature for data-driven induction of pattern classifiers [Mitchell, 1997; Bishop, 1995; Balakrishnan and Honavar, 1998]. Such classifiers may take the form of decision trees [Quinlan, 1992], artificial neural networks [Bishop, 1995], statistical models [Duda and Hart, 1973], grammars and automata [Parekh and Honavar, 2000], among others. The choice of the model and the inference algorithm are influenced by various factors including: the form and amount of data and prior knowledge that are available; the need for transparency of the learned model, etc. [Mitchell, 1997]. This paper describes an approach to data-driven automated discovery of decision trees for assigning protein sequences to functional families based on the motif composition of the sequences.

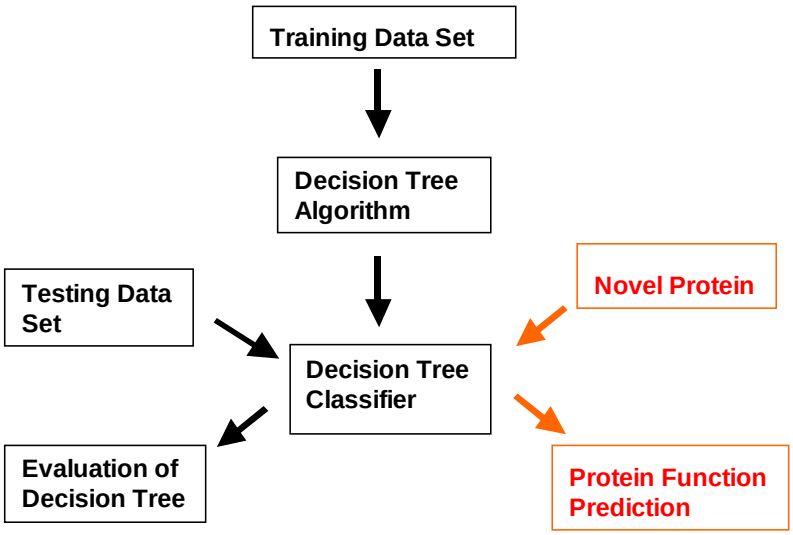


Figure 1. Overview of experiment design. Protein sequences with known functions are divided into a training dataset and a testing dataset, respectively. A decision tree algorithm is used to build a decision tree classifier using the training dataset. Classification accuracy of the decision tree is determined using the testing dataset. Finally, the decision tree will be used to assign noval protein sequences to known functional families based on the proteins' motif compositions.

Computational Problem and Approach

The basic computational problem we seek to address is the following: Given a database or training set of amino acid sequences that code for 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 basic approach is illustrated in Figure 1.

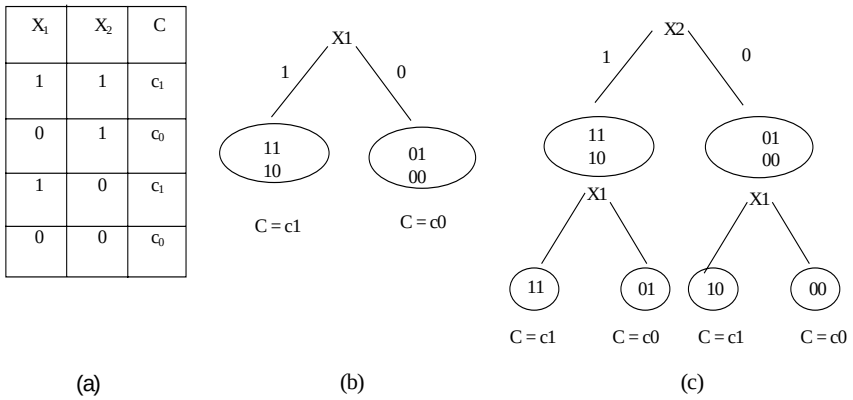
Data Representation

The first step in this process is the preparation of a data set. 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.* The choice of attributes plays a critical role in the data mining process. 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 or the set of motifs identified by running a suitable motif-finding programs on the set of sequences. Our choice of motif-based representation of sequences is inspired by the success of a similar vocabulary-based representation of documents in text classification [Salton, 1983].

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 *i*th 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 function of the corresponding protein. This training set can be 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. This process is illustrated in Figure 1.

Inducing Pattern Classifiers

Decision trees [Quinlan, 1992] offer an attractive model for pattern classification. A number of algorithms are available for inducing decision trees from training sets. One such algorithm recursively selects attributes (using an *information gain* criterion) that *best* partition the training data into various classes until the resulting tree assigns each of the instances in the training set to their respective (known) classes [Figure 2]. The transparency of the resulting trees make it easy for humans to examine and to explore the significance of the regularities extracted from the data. In this study, we used a decision tree learning algorithm [Quinlan, 1992] for building protein sequence classifiers. However, the basic approach would work unchanged with the decision tree learning algorithm replaced with one of the other data-driven algorithm for constructing pattern classifiers [Mitchell, 1997; Balakrishnan and Honavar, 1998].



X_1	X_2	C
1	1	c_1
0	1	c_0
1	0	c_1
0	0	c_0

Figure 2. Construction of decision trees. Given a training dataset, the decision tree learning algorithm [Quinlan, 1992] generates a *simple* tree that correctly classifies the training examples. Among the two trees shown, the first (shown in b) is simpler than the second (shown in c). In this case, the training set has 4 examples belonging to one of two classes (c0 or c1). Each pattern is represented using two binary attributes x1 and x2. Decision tree shown in (b) tests on attribute x1 first. The leaf nodes of the tree have class labels associated with them. Thus, once the tree is constructed, a pattern can be assigned to the appropriate class by checking the value of the corresponding attributes and following the corresponding branches starting at the root of the tree.

Data Preparation

The Prosite database (http://www.expasy.ch/cgi-bin/prosite-list.pl) contains over 1100 entries. Each entry describes a function shared by some proteins. In this paper, one Prosite documentation entry corresponds to a protein class. For example, the entry PDOC00662 corresponds to class "MCM family signature and profile". Specifically, the protein classes considered in this study are: PDOC00662 (a class of DNA or RNA associated proteins), PDOC00064 (a class of oxidoreductases), PDOC00670 (a class of transferases), PDOC50007 (a class of hydrolases), PDOC00154 (a class of isomerases), PDOC00343 (a class of structural proteins), PDOC00561 (a class of receptors), PDOC00224 (a class of cytokines and growth factors), PDOC00791 (a class of protein secretion and chaperones), and PDOC00271 (a class included in the catch-all "Others" category). For clarity of presentation, the Prosite documentation ID, i.e., the PDOCxxxxx number, was used to represent that class. Similarly, the Prosite access number, i.e., the PSxxxxx number, was used to represent that motif pattern or profile. In the Prosite database, a protein motif can be either a regular expression (defined over the 20 amino acid alphabet), called pattern, or a weighted matrix (built on alignment of multiple protein sequences), called profile. Each Prosite documentation entry lists not only the true positive proteins (the protein that indeed belong to that class), but also false positive proteins (the proteins that *do not* belong to that class but contain the characteristic motif(s) associated with that class), false negatives, potential hits, and unknowns. The proteins used in this paper include the true positive, false positive, and false negative proteins. Whereas the true positive and false negative proteins were labeled with their

respective correct class labels, all false positive proteins were classified as "Other", represented with "PS00000". Five hundred and eighty five proteins were collected from the 10 classes mentioned above.

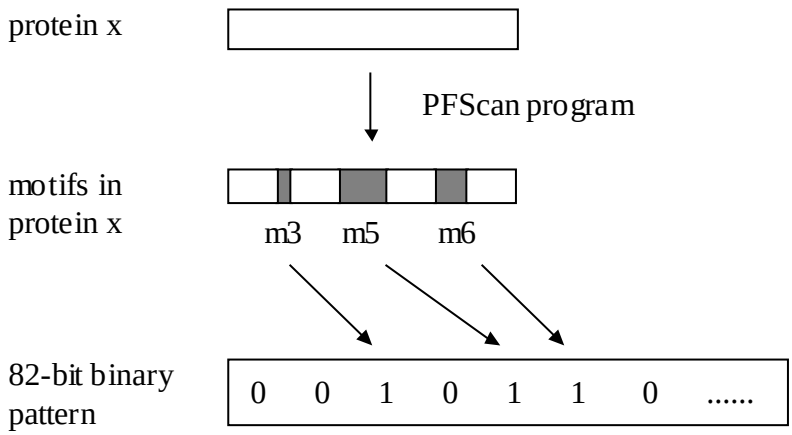


Figure 3. Illustration of data preparation from protein sequences. Protein x was processed by PfScan program to determine its motif composition(as shown by motif a, b, c and d here). A total number of 82 motifs were used in our work. Thus a 82 elements vector was used to represent the presence(1) or absence(0) of motifs in fixed order.

Each protein class can be characterized by one or more characteristic motif patterns and/or profiles. For example, class PDOC00670 has two characteristic motifs, PS00856 a pattern and PS50052 a profile. Protein sequences containing any of the characteristic motifs of a functional class were collected and labeled as belonging to that class. Each collected protein was then processed by the profileScan program (http://www.isrec.isb-sib.ch/software/PFSCAN_form.html) to determine its motif composition. Only the motifs that were identified as *significant* matches by profileScan were chosen. This analysis identified additional motifs in the sequences besides the ones designated as the characteristic motifs for the family associated with each sequence. Thus, each protein sequence was represented using 82 binary attributes with each attribute denoting the presence or absence of the corresponding motif in the sequence (see Figure 3). Experiments in this paper were carried out mainly with proteins in this data set.

In addition, a set of 73 proteins collected from another five classes was also used in one experiment. The 5 protein families were: PDOC00360 (Poly [ADP-Ribose] Polymerase, PPZF), PDOC00295 (DNA Ligase, LIGASE), PDOC00605 (Guanine Releasing Factor, GRF), PDOC50003 (Cytoskeletal protein, CYTO), and PDOC00463 (Yeast Transcription activator, ACT). Using a procedure similar to the one described above for the data set with 10 protein families, each of the 73 protein sequences was represented using 10 binary attributes with each attribute denoting the presence or absence of a motif.

Experiments and Results

Generating Decision Tree Classifiers from Training Data

Five hundred and eighty five proteins belonging to one of the 10 classes or to the “Other” class (the false positive proteins from all of the 10 protein classes) were used in this experiment. Subsets of proteins were randomly picked from the 585-protein pool as the training samples. The sizes of the training sample sets were 11 (2% of the total sample set), 20, 29, 58, 117, 175, 234, 294, 351, and 585 proteins. For a given training set size, the experiment was

repeated three times using a different randomly sampled training set in each case. After a decision tree was built using a training set, all 585 proteins were used as the test set to determine classification accuracy of the resulting decision tree. The results (shown in Table 1) indicate that with only 10% of the total protein samples a decision tree could be constructed to classify proteins with an accuracy of 94%.

Table 1. Effect of training set size on classification accuracy*.

Training set size	11	20	29	58	117	175	234	294	351	585
Accuracy (%)	59.1	71.7	82.7	94.0	95.1	96.5	98.4	99.0	98.7	99.8

* The results are the average of three independent runs.

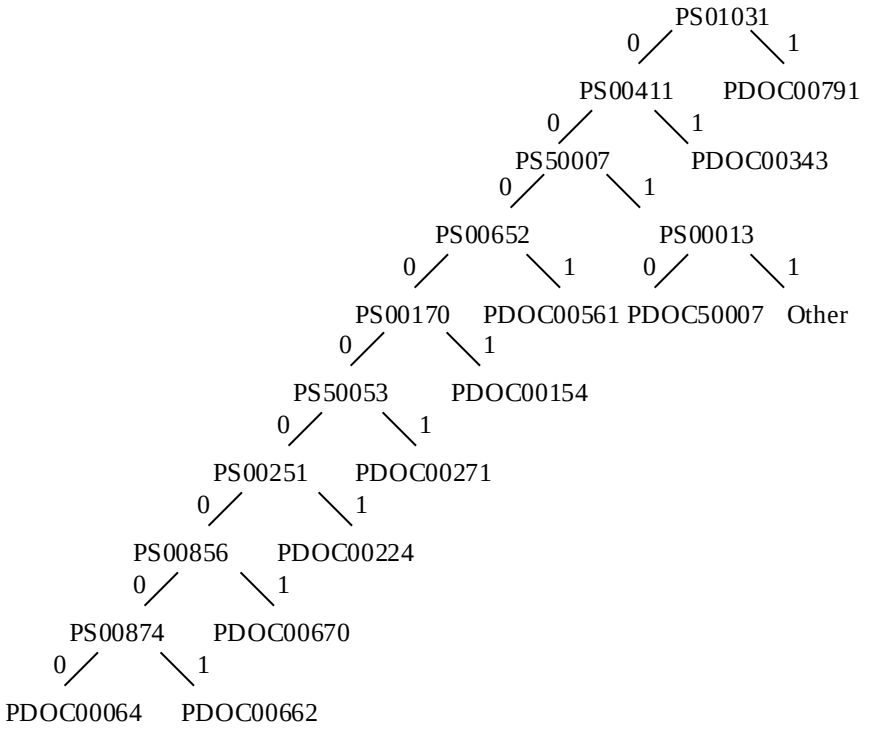


Figure 4. Illustration of a decision tree constructed with 58 training proteins. The root and internal nodes represent attributes (i.e., Prosite motifs), and the leaf nodes represents protein classes. The value "0" of the first attribute, PS01031, leads to the internal node PS00441, meaning if a protein contains motif PS01031, then it needs further evaluation with motif PS00411; the value "1" of the attribute PS01031 leads to the leaf node PDOC00791, meaning if a protein contains motif PS01031, then it belongs to class PDOC00791. Similarly, if a protein contains motif PS00411 (the value is "1"), it belongs to class PDOC00343; otherwise it needs to be evaluated further with motif PS50007, and so on. The "Other" class (PDOC00000) contains all false positive proteins (see Materials and Methods section).

Each protein class, defined according to a Prosite documentation entry (see above), is represented by one or more characteristic motifs. On the other hand, each motif is associated with a unique documentation entry i.e., a protein class. Analysis of the decision trees constructed in this experiment indicated that the characteristic motifs of a protein class played a critical role in classification (a tree built with all 58 proteins is shown in Figure 4). On the

surface, this might raise the question as to whether a decision tree offers anything beyond a simple query of the Prosite database with the characteristic motif. However, a closer examination of the decision trees generated by the algorithm indicates that there are situations in which the combinations of motifs that are used by the decision tree for separating the various families are different from the documented characteristic motifs for the corresponding families. Furthermore, the false positives generated by the decision tree are significantly fewer than those resulting from a Prosite search using the characteristic motif for each family. There were totally 13 false positive proteins from the 10 classes based on querying the Prosite database. The number of false positives resulting from the use of the decision tree trained using training sets of different sizes are shown in Table 2.

Table 2. Number of false positive classifications resulted from the decision tree program*.

Training set size	11	20	29	58	117	175	234	294	351	585
False positives	281.7	165.3	101.0	35.0	28.7	20.3	9.3	5.7	7.3	1.0

* The numbers represent averages over three independent runs

The results show that the number of false positive classifications using the decision tree falls below that resulting from Prosite search using characteristic motifs for training set sizes greater than or equal 234 (40% of the data set). The number of false positives approaches zero as the fraction of the data set used for training approaches 100%. This suggests that the decision tree program in fact discovers regularities among protein sequences that belong to a functional family that are not captured explicitly by their characteristic motifs as documented in the Prosite database.

To further explore this issue, a second data set of 73 protein sequences drawn from five classes (see Materials and Methods section for details) were used to build a decision tree. The protein classes were chosen such that there were significant overlaps among the families in terms of their motif composition. For example, motif PS50010 (GRF_DBL) is present in proteins belonging to both classes PDOC00605 (GRF) and PDOC00360 (PPZF). In this scenario, querying the Prosite database with a single characteristic motif would result in a high rate of false positives. However, the decision tree built by using randomly sampled training instances from this data set resulted in highly accurate assignment of sequences to the data set, the classification exceeded 95% when the size of the training set was greater than or equal to 22 (30% of the data set). When the trees were trained with 58 or more sequences (representing 80% or more of the data set) every sequence in the data set was correctly assigned to the corresponding functional family by the resulting decision tree. A sample decision tree constructed using a training set of size 58 is shown in Figure 5. The decision tree distinguishes proteins belonging to class PDOC00360 from those belonging to class PDOC00605 based on the presence of PS50064 (PARP_ZN_FINGER_2) motif in the former but not in the latter although both families contain the PS50010 (GRF) motif.

Table 3. Classification of proteins containing common motifs.*

Training set size	4	7	15	22	29	36	44	51	58	73
Accuracy (%)	62.1	81.3	89.5	95.4	97.3	98.2	95.9	99.1	100	100

* The data are the average of three independent runs.

Assigning Novel Sequences to Families Using a Trained Decision Tree Classifier

The previous experiments demonstrate the effectiveness of the proposed approach in constructing fairly accurate models that capture regularities that help to accurately classify sequences belonging to different functional families. The extracted regularities are in form of combinations of motifs that are present or absent in the respective sequences. The accuracy of the resulting classifier exceeds that obtained by querying the Prosite database with the characteristic motif for each family. However, the real utility of the data-driven approach to building classifiers for functional classification of protein sequences would be in assigning novel sequences (with unknown function) to one of the known functional families. Conclusive demonstration of this would entail verifying the predictions of the

classifier through biological experiments. However, we can assess the usefulness of the proposed approach in this context by systematic computational experiments where the predictions given by the decision tree are compared with the (known) correct classifications on a part of the data set that is not used in training. We randomly partitioned the data set of 585 proteins into 3 sets named split1, split2, and split3, each containing 195 proteins. One of the partitions was used to train the decision tree, the other two were used to test the resulting classifier. The results (shown in Table 4) demonstrate the ability of the decision tree to generalize effectively beyond the training data.

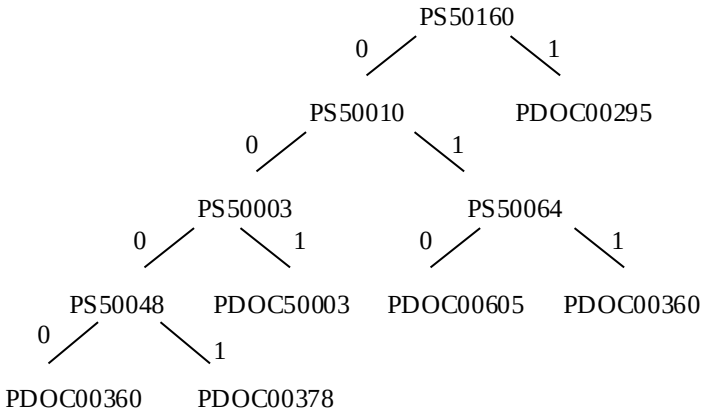


Figure 5. Illustration of a decision tree built with 58 training proteins. The classification accuracy using this tree was 100%. An internal node represents an attribute for classification, whereas a leaf node represents a protein class. In this figure, value "0" of attribute "PS50160" (motif DNA_LIGASE_A3) leads to the internal node PS50010 (motif GRF_DBL) whereas value "1" of "PS50160" leads to class PDOC00295 (LIGASE).

Table 4. Protein classification with split data sets.

Training set	Split1	Split2	Split3
Test Accuracy	97.7 %	96.4 %	98.7 %

Summary and Discussion

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. Machine learning [Mitchell, 1997] currently offers one of the most effective and practical approaches to data-driven knowledge acquisition. Decision tree learning algorithm [Quinlan, 1992] represents one of the simplest and most commonly used machine learning algorithms for data-driven induction of classifiers. In this paper, we have presented an application of the decision tree learning algorithm for building protein sequence classifiers for assigning protein sequences to one of several functional families using a training set of sequences that are labeled with their corresponding functional families. The experimental results presented in this paper show that resulting decision tree classifiers are able to *generalize* well on test *sequences* that were not part of the training set. Furthermore, the decision trees provide more accurate models of protein functional families than those based on *characteristic motifs* for some of the families documented in the Prosite database. Examination of the resulting decision trees indicates that the algorithm is able to discover from the data, the *presence or absence of combinations of subsets of motifs* that distinguish sequences belonging to each functional family from sequences belonging to other functional families represented in the training data. In particular, the decision trees are able to identify *interactions* among motifs that can be quite far apart from each other with respect to their positions in the

sequence. Such interactions might have a critical influence on the 3-dimensional structure and function of the protein.

Like any data driven technique, the proposed approach relies on the availability of representative sequences corresponding to proteins with known function for building the classifier. When such data is available, the proposed approach can be quite effective in assigning putative functions to novel sequences. This can serve as a useful source of information for guiding focused biological experiments. Our current work is aimed at the development of algorithmic and systems solutions for integrated computational analysis and prediction of macro-molecular structure-function relationships using multiple heterogeneous, information sources (e.g., phylogenetic analysis, gene expression data) [Honavar et al., 1998b]. Some directions for further research include: systematic comparison of different machine learning algorithms for building predictors of protein function from sequence data; evaluation of the effectiveness of alternative approaches to motif detection in conjunction with different learning algorithms for building such predictors; and integration of the resulting tools with visualization routines for exploratory analysis of macro-molecular structure-function relationships.

Acknowledgments

This research was supported in part by grants from the National Science Foundation (9982341, 9972653), the Carver Foundation, Pioneer Hi-Bred, Inc. This research has benefited from interactions with Dr. Amy Andreotti, Dr. Gavin Naylor, Dr. Kai-Ming Ho, Dr. Les Miller, Dr. James Morris of the Iowa State University Protein Structure Group.

References

1. Dayhoff, M. O.; Barker, W. C.; Hunt, L. T. "Establishing Homologies in Protein Sequences," *Methods in Enzymology*, **91,** 524 (1983).
2. Hudak, J. and McClure, M.A. (1999) A Comparative Analysis of Computational Motif Detection Methods. Pacific Symposium on Biocomputing 4:138-149 (1999).
3. Hofmann K., Bucher P., Falquet L., Bairoch A. (1999) *The PROSITE database, its status in 1999* Nucleic Acids Res. 27:215-219.
4. Bateman, A. Birney, E.., Durbin, R., Eddy, S., Howe, K., and Sonnhammer, E. (2000) *Nucleic Acids Research*, 28:263-266.
5. Attwood, T.K. and Beck, M.E. (1994) PRINTS - A protein motif finger- print database. Protein Engineering, 7 (7), 841-848.
6. Mitchell, T. (1997). *Machine Learning.* New York: McGraw Hill.
7. Bishop, C. (1995) *Neural Networks for Pattern Recognition*. New York: Oxford University Press
8. Honavar, V., Parekh, R. and Yang, J. (1999). Machine Learning. Invited article. In: *Encyclopedia of Electrical and Electronics Engineering*, Webster, J. (Ed.), New York: Wiley.
9. Quinlan, J. R. (1992) C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo, CA.
10. Duda, R. & Hart, P. (1973) *Pattern Classification and Scene Analysis.* New York: Wiley.
11. Parekh, R. and Honavar, V. (2001) DFA Learning from Simple Examples. *Machine Learning.* In press.
12. Baldi, P. and Brunak, S. (1998) Bioinformatics: the Machine Learning Approach. Cambridge, MA: MIT Press.
13. Salton, G. and McGill, M. (1983). *Introduction to Modern Information Retrieval.* New York: McGraw Hill.
14. Honavar, V., Miller, L. and Wong, J. (1998) Distributed Knowledge Networks. In: Proceedings of the IEEE Information Technology Conference. Syracuse, NY.
15. Balakrishnan, K. and Honavar, V. (1998) Intelligent Diagnosis Systems. *Journal of Intelligent Systems.*. Vol. 8. No.3/4. pp. 239-290.