DISCOVERING PROTEIN FUNCTION CLASSIFICATION RULES FROM REDUCED ALPHABET REPRESENTATIONS OF PROTEIN SEQUENCES

Carson M. Andorf, Drena L. Dobbs, Vasant G. Honavar[1]
Artificial Intelligence Laboratory
Department of Computer Science and
Graduate Program in Bioinformatics and Computational Biology
Iowa State University
Ames, IA 50011, USA
www.cs.iastate.edu/~honavar/aigroup.html

ABSTRACT

The paper explores the use of reduced alphabet representations of protein sequences in the data-driven discovery of data-driven discovery of sequence motif-based decision trees for classifying protein sequences into functional families. A number of alternative representations of protein sequences (using a variety of reduced alphabets based on groupings of amino acids in terms of their physico -chemical properties were explored in addition to the 20-letter amino acid alphabet. Classifiers were constructed using motifs generated using a multiple sequence alignment based motif discovery tool (MEME). Results of experiments on a data set of eleven protease families show that the classification performance of the resulting decision trees based on several reduced alphabets (e.g., a 7-letter alphabet based on groupings of amino acids based on their mass and charge, a 5-letter alphabet based on a random grouping of the 20 amino acids into 5 groups) is comparable to that of trees based on the 20-letter amino acid alphabet. The results also show that the sequence motifs based on different alphabets capture regularities in different portions of the sequences. This raises the possibility that the use of different alphabets might provide different, but complementary insights into protein structure-function relationships.

1. BACKGROUND AND INTRODUCTION

Assigning putative functions to protein sequences remains one of the most challenging problems in functional genomics. Our previous work [Wang et al., 2001; 2002] has shown that decision trees generated from sequence motif based representation of protein sequences successfully perform protein function classification. The success of motif based approaches to construction of protein function classifiers raises several interesting questions: What makes specific motifs or combinations of motifs serve as good predictors of protein function? Can we explain their success in terms of specific physico-chemical properties of the amino acids involved? Can we gain some useful insights into protein structure -- function relationships by exploring sequence regularities that are good predictors of function?

All data-driven knowledge acquisition techniques including the techniques for discovery of protein structure-function relationships outlined above have one thing in common: They search for patterns in the data that are predictive of specific classes of interest (e.g., functional families). The regularities that are found depend on: the choice of the data representation (e.g., representation of proteins by their amino acid sequences or their motif composition); knowledge representation (e.g., rules that capture the presence or absence of combinations of motifs); and the criteria used to select specific regularities from among a large number of candidates [Mitchell, 1997]. Thus, it is interesting to explore alternative alphabets for representation of the protein sequences to be classified.

There are several examples of the use of reduced alphabet representations of protein sequences based on alphabets that correspond to groupings of the 20 amino acids into categories based on specific physico-chemical properties of amino acids (e.g., hydrophobicity) in studies of protein recognition and protein folding [Chan, 1999; Murphy et al., 2000], protein synthesis [Riddle et al., 1997], phylogeny [Naylor and Brown, 1998], sequence matching and retrieval [Smith and Smith, 1990], among others.

Against this background, this paper explores data-driven construction of motif-based decision trees for protein function classification using motifs discovered from reduced alphabet representations of protein sequences. Different reduced alphabets correspond to groupings of the 20 amino acids into categories based on specific physico-chemical properties (e.g., hydrophobicity, charge, volume, surface area, solubility), or combinations of properties of amino acids or evolutionary information.

2. AUTOMATED DISCOVERY OF PROTEIN FUNCTION CLASSIFIERS USING REDUCED ALPHABET PROTEIN SEQUENCE REPRESENTATIONS

Data Set Used

Protease families were chosen for this study because many of them are well-characterized, with known structures and functions [Barrett et al., 1998]. The data set used in this study consisted of proteins from eleven families of proteases (S1, S2A, S2B, S2C, S3, S6, S8, S9A, S9B, S9C, and S10) in the MEROPS database (http://www.merops.co.uk/) [Rawlings and Barrett, 1993]. The eleven families represent a subset of serine-type peptidases (proteases) from MEROPs and include a total of 206 proteins. The largest family, S1, contained 66 proteins and the smallest family S8 contained 5 proteins The protein sequences for members of each of the eleven protease families were obtained from the SWISS-PROT protein knowledgebase (http://www.expasy.ch/sprot/sprot-top.html) [Barioch and Apweiler, 2000].

[1] This research was supported in part by grants from the National Science Foundation (9982341, 9972653), the Carver Foundation, and Pioneer Hi-Bred, Inc. and has benefited from interactions with Diane Schroeder, Dr. Xiangyun Wang, Zhong Gao, and Changhui Yan, of the Iowa State University Artificial Intelligence Research Laboratory.

Reduced Alphabet Representations of Protein Sequences

We tested four different types of reduced alphabets, corresponding to:

1. Groupings of amino acids based on a single physico-chemical characteristic of each amino acid (one-tuple representation) e.g., hydrophobicity;
2. Groupings of amino acids based on simultaneous consideration of two amino acid characteristics for each amino acid (two-tuple representation) e.g., mass and volume;
3. Groupings defined by Murphy et al., [2000] based on a Blosum50 substitution matrix [Henikoff and Henikoff, 1992
4. Groupings based on randomly generated partitions of the set of amino acids

Amino Acid Alphabet	Hydrophobicity	Charge	Volume	Mass
A (Alanine)	Hydrophobic	No Charge	Small	Small
C (Cysteine)	Hydrophobic	No Charge	Medium	Medium
D (Aspartic Acid)	Hydrophilic	Negative	Medium	Medium-Large
E (Glutamic Acid)	Hydrophilic	Negative	Medium-Large	Medium-Large
F (Phenylalanine)	Hydrophobic	No Charge	Large	Large
G (Glycine)	Hydrophobic	No Charge	Small	Small
H (Histidine)	Hydrophilic	Positive	Medium-Large	Medium-Large
I (Isoleucine)	Hydrophobic	No Charge	Medium-Large	Medium-Large
K (Lysine)	Hydrophilic	Positive	Medium-Large	Medium-Large
L (Leucine)	Hydrophobic	No Charge	Medium-Large	Medium-Large
M (Methionine)	Hydrophobic	No Charge	Medium-Large	Medium-Large
N (Asparagine)	Hydrophilic	No Charge	Medium	Medium-Large
P (Proline)	Hydrophobic	No Charge	Medium	Medium
Q (Glutamine)	Hydrophilic	No Charge	Medium-Large	Medium-Large
R (Arginine)	Hydrophilic	Positive	Medium-Large	Large
S (Serine)	Hydrophilic	No Charge	Small	Medium
T (Threonine)	Hydrophilic	No Charge	Medium	Medium
V (Valine)	Hydrophobic	No Charge	Medium-Large	Medium
W(Tryptophan)	Hydrophobic	No Charge	Large	Large
Y (Tyrosine)	Hydrophobic	No Charge	Large	Large

Table 1: A table of physico-chemical properties of amino acids [Kyte and Doolittle; 1982, Taylor 1986; Zamyatin, 1972; Lide, 2001] used to generate reduced alphabets.

An one-tuple representation is a many-to-one mapping of amino acids to a new alphabet corresponding to a single amino acid characteristic. Physico-chemical properties of amino acids (e.g., hydrophobicity, charge, volume, mass, surface area, solubility) are often used as the basis for generating reduced alphabets [Taylor 1986]. To generate a one-tuple representation based on these properties, we must assign each amino acid a discrete value for the property chosen. A simple way to do this is to cluster the amino acids based on their (continuous) values for that property. For example, amino acids can be divided into two groups on the basis of their hydrophobicity (hydrophobic and hydrophilic) or four groups on the basis of mass (small, medium, medium large, and large). To create the new representation, each amino acid is mapped to its corresponding value for the chosen property. For simplicity, we represented each amino acid in the "hydrophobic" group with the letter R (the single letter code for Arginine, the most hydrophobic amino acid) and each amino acid in the "hydrophilic" group with the letter H (for Histidine, the most hydrophilic amino acid). Thus, for hydrophobicity, the 20-letter amino acid alphabet was replaced with a new 2-letter alphabet {H,R} for representing protein sequences. For grouping the amino acids based on one-tuple physico-chemical properties, we considered four properties: hydrophobicity, charge, volume, and mass. Table 1 shows the one-tuple physico-chemical property assignments for each of the 20 naturally occurring amino acids. Amino acids were divided into two groups on the basis of hydrophobility [Kyte and Doolittle,1982], forming a new two-letter alphabet {H,R}. Three groups were generated on the basis of *charge* [Taylor, 1986]: positively charged, negatively charged, and uncharged, creating a three letter alphabet {P,N,U}. Volume [Zamyatin, 1972] and mass [Lide, 2001] were each mapped to four discrete values: small, medium, medium large, and large, creating four letter alphabet {S, M, A , L} based on these properties. The discrete values were determined by clustering [Sokal and Michener, 1958].

A two-tuple representation is a many-to-one mapping of amino acid identities to a new alphabet corresponding to two of its characteristics. We used a procedure analogous to that described above to create new alphabets based on two physico-chemical properties for each amino acid. For example, a new alphabet created using the properties hydrophobicity {H,R} and mass {S,M,A,L} has eight characters ($|H| \times |M| = 2 \times 4 = 8$). For example, alanine has the properties of being hydrophilic and small, so it would be mapped to (R-S). When the twenty amino acids are mapped according to these two properties, and alphabet of actual size seven is generated (shown in Table 2a) because no amino acid is both hydrophobic and small. Sequence representations based on partitions of amino acids using combinations of values for any k characteristics can be generated in a similar fashion. However, when k exceeds 2, the resulting alphabets approach the original 20-letter amino acid alphabet. Table 2b shows a list of the six possible two-tuple alphabets (based on the four physico-chemical properties), along with their sizes and compositions.

Amino acid substitution matrices based on multiple sequence alignments from large databases are very widely used in sequence analysis. The Blosum50 substitution matrix [Henikoff and Henikoff, 1992] was used to generate reduced amino acid alphabets that of sizes ranging from two to fifteen for protein fold recognition using global sequence alignment [Murphy et al., 2000]. Table 3 shows the Blossum50 matrix-based amino acid groupings used in our study, which correspond to the reduced alphabets defined in the Murphy study [Murphy et al., 2000].

Amino Acid Alphabet	Hydrophobicity/Mass Alphabet
C (Cystine) S (Serine) T (Threonine)	(R-M) (Hydrophobic – Medium)
D (Aspartic Acid) E (Glutamic Acid) H (Histidine) K (Lysine) N (Asparagine) Q (Glutamine)	(R-ML) (Hydrophobic-Medium/Large)
R (Arginine) Y (Tyrosine)	(R-L) (Hydophobic –Large)
A (Alanine) G (Glycine)	(H-S) (Hydrophilic – Small)
P (Proline) V (Valine)	(H-M) (Hydrophilic – Medium)
I (Isoleucine) L (Leucine) M (Methioine)	(H-ML) (Hydrophilic – Medium/Large)
F (Phenylalanine) W(Tryptophan)	(H-L) (Hydrophilic – Large)

Table 2a: Many-to-one mapping of amino acid identity to Hydrophobic/Hydrophilic – Mass Alphabet.

Two-tuple Alphabet	Alphabet
Hydrophobicity – Charge Size = 6, Size in practice = 4	{Hydrophobic-Positive, Hydrophobic-Negative, Hydrophobic-Uncharged, Hydrophilic-Positive, Hydrophilic-Negative, Hydrophilic- Uncharged}
Hydrophobicity – Volume Size = 8, Size in practice =8	{Hydrophobic- Small, Hydrophobic-Medium, Hydrophobic-Medium/Large, Hydrophobic-Large, Hydrophilic-Small, Hydrophilic-Medium, Hydrophilic-Medium/Large, Hydrophilic-Large}
Hydrophobicity – Mass Size = 8, Size in practice =7	{Hydrophobic- Small, Hydrophobic-Medium, Hydrophobic-Medium/Large, Hydrophobic-Large, Hydrophilic-Small, Hydrophilic-Medium, Hydrophilic-Medium/Large, Hydrophilic-Large}
Charge – Volume Size = 12, Size in practice =7	{Positive-Small, Positive-Medium, Positive-Medium/Large, Positive-Large, Negative-Small, Negative-Medium, Negative-Medium/Large, Negative-Large, Uncharged-Small, Uncharged-Medium, Uncharged-Medium/Large, Uncharged-Large}
Charge – Mass Size = 12, Size in practice =7	{Positive-Small, Positive-Medium, Positive-Medium/Large, Positive-Large, Negative-Small, Negative-Medium, Negative-Medium/Large, Negative-Large, Uncharged-Small, Uncharged-Medium, Uncharged-Medium/Large, Uncharged-Large}
Volume – Mass Size = 16, Size in practice =8	{Small- Small, Small-Medium, Small-Medium/Large, Small-Large, Medium-Small, Medium-Medium, Medium-Medium/Large, Medium-Large, Medium/Large-Small, Medium/Large-Medium, Medium/Large-Medium/Large, Medium/Large-Large, Large- Small, Large-Medium, Large-Medium/Large, Large-Large}

Table 2b: A table of the two-tuple alphabets and their size. Size refers to the actual size if every two-tuple combination is used. Size in practice is the size of the two-tuple combinations that actually occur in the data set.

The fourth method for creating reduced alphabets is based on random partitioning of the set of 20 amino acids into a number of groups. The number of groups corresponds to the alphabet size. Once alphabet size (and hence number of groups in the partition) is fixed, each amino acid is randomly assigned to one of the groups with a probability equal to the reciprocal of the number of groups. A representative subset of the randomly generated reduced alphabets are shown in Table 4.

Datasets based on reduced alphabet representations of the original data set were generated using each of the reduced alphabets described above.

	Alphabet size				
Amino Acid Alphabet	15	10	8	4	2
L (Leucine)	Group 1	Group 1	Group 1	Group 1	Group 1
V (Valine)	Group 1	Group 1	Group 1	Group 1	Group 1
I (Isoleucine)	Group 1	Group 1	Group 1	Group 1	Group 1
M (Methioine)	Group 1	Group 1	Group 1	Group 1	Group 1
C (Cystine)	Group 2	Group 2	Group 1	Group 1	Group 1
A (Alanine)	Group 3	Group 3	Group 2	Group 2	Group 1
G (Glycine)	Group 4	Group 4	Group 2	Group 2	Group 1
S (Serine)	Group 5	Group 5	Group 3	Group 2	Group 1
T (Threonine)	Group 6	Group 5	Group 3	Group 2	Group 1
P (Proline)	Group 7	Group 6	Group 4	Group 2	Group 1
F (Phenylalanine)	Group 8	Group 7	Group 5	Group 3	Group 1
Y (Tyrosine)	Group 8	Group 7	Group 5	Group 3	Group 1
W(Tryptophan)	Group 9	Group 7	Group 5	Group 3	Group 1
E (Glutamic Acid)	Group 10	Group 8	Group 6	Group 4	Group 2
D (Aspartic Acid)	Group 11	Group 8	Group 6	Group 4	Group 2
N (Asparagine)	Group 12	Group 8	Group 6	Group 4	Group 2
Q (Glutamine)	Group 13	Group 8	Group 6	Group 4	Group 2
K (Lysine)	Group 14	Group 9	Group 7	Group 4	Group 2
R (Arginine)	Group 14	Group 9	Group 7	Group 4	Group 2
H (Histidine)	Group15	Group 10	Group 8	Group 4	Group 2

Table 3: Reduced alphabets of Murphy et al., 2000, based on Blosum50 substitution matrix [Henikoff and Henikoff, 1992].

Motif-based Representation of Protein Sequences

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.* We represent protein sequences using a suitable *vocabulary* of *sequence motifs* [Wang et al., 2001a; Wang et al., 2001b]. 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 data 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. In this study, the set of sequence motifs (the vocabulary) was obtained by running MEME – a multiple alignment based motif discovery program [Bailey, et.al., 1998] on sequence data for each of the peptidase families. The MAST (Motif Alignment and Search Tool) program was used to determine the motif composition of a protein sequence. Several perl scripts were used to transform the MAST output into the data format that can be used by the C4.5 program [Quinlan, 1992] for construction of decision trees for assigning protein sequences to the corresponding families.

Alphabet	Size	Mappings	Mean Error
Random 2a	2	{P,T,V,E,H,I,K,Q,R,F,W,Y} ; {A,G,S,C,D,N,L,M}	57.08
Random 2e	2	{G,I,Q,H,D,N,R,A,F,C} ; {T,K,S,P,M,V,W,L,Y,E}	62.56
Random 3a	3	{T,D,N,L,W} ; {C,V,E,H,I,K,M,F,Y} ; {A,G,S,P,Q,R}	39.80
Random 3d	3	{N,I,K,Q,H,D,M} ; {P,S,V,A,C,F} ; {L,Y,G,W,E,T,R}	41.98
Random 4a	4	{G,T,H,R} ; {S,E,Q,F,Y} ; {D,N,I,K,L,M} ; {A,C,P,V,W}	5.02
Random 4e	4	{W,Q,S,V,N,E} ; {A,P,Y,C} ; { F,I,G,L,M} ;{H,R,K,D,T}	14.24
Random 5a	5	{H,M} ; {P,T,D,N,K,L,R,F,W} ; {A,G,S} ; {Q,Y} ; {C,V,E,I}	4.23
Random 5e	5	{G,Q,H,S,P} ; {D,Y,R} ; {I,N,C,L,E} ; {F,K,M,V} ; {A,W,T}	8.41
Random 6a	6	{D,N,M,R,F} ; {S,H,K} ; {A,G,V,Q,W,Y} ; {C,T} ; {P,E,I} ; {L}	8.64
Random 6c	6	{M,Q} ; {C,T,L,W,A} ; {P,N,V} ; {H,Y,K,D,E} ; {R,F} ; {G,S,I}	9.73
Random 7a	7	{T,V,E,I} ; {A,K,Q} ; {G,C} ; {P} ; {S,W} ; {N,H,L,M,Y} ; {D,R,F}	8.06
Random 7e	7	{N,I,W,E} ; {P,L,G} ; {Q,T,R} ; {C,F,D} ; {M} ; {K,S,V,Y} ; {H,A}	6.88
Random 8a	8	{W,Y,A,C} ; {V,Q,S,N} ; {F} ; {P,H,K,R,D} ; {T,I,G} ; {M} ; {E} ; {L}	4.19
Random 8b	8	{Q,N,K} ; {H,I,M} ; {T,Y,A} ; {D} ; {W,V} ; {R,G} ; {E} ; {C,S,F,P,L}	8.82
Random 9a	9	{T} ; {Q} ; {V,E,Y,G,N} ; {I,R,W,A} ; {C} ; {H,F} ; {P,K,M} ; {S} ; {D,L}	8.77
Random 9c	9	{R} , {H,T} ; {G,D,L} ; {I,P} ; {Q,Y,A} ; {N,M} ; {W,F,V} ; {S} , {C,K,E}	6.89
Random 10a	10	{I,K} ; {N} ; {W,C} ; {G} ; {Q,A} ; {R} ; {T,V} ; {E,F,D,M} ; {P,Y,S,L} ; {H}	10.71
Random 10e	10	{V} ; {H,D,R} ; {P,T} ; {N,S,E} ; {Q,G} ; {L} ; {M,W} ; {I,C} ; {A,F} ; {K,Y}	5.37

Table 4: A table of the random alphabets. The size of the alphabet, groupings of amino acids that define the corresponding alphabet, the mean error of the decision tree classifier based on the alphabet are shown for several representative randomly generated reduced alphabets. Size refers to the actual number of different partitions of the twenty letter amino acid alphabet. The reported error estimates are based on 50 independent runs of the decision tree learning algorithm using a randomly sampled 2/3 of the data set for training and the remaining 1/3 for testing.

Because different choices of alphabet result in different representations of protein sequences, and motifs are constructed from aligned sequences, we generated a distinct data set corresponding to each reduced alphabet. Once a data set is constructed, a subset of the data set (training set) can be used to train a classifier which can then be used to assign sequences to one of the several functional families represented in the training set.

Data Driven Generation of Decision Tree Classifiers

In this study, we used the C4.5 family of decision tree algorithms [Quinlan, 1992] for building protein sequence classifiers. It 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 were then converted into rules for further analysis. Each rule is of the form ``if *condition* then *class"* where condition checks for a combination of motifs which need to be present (or absent) to reliably predict the classification of the corresponding protein.

3 EXPERIMENTS AND RESULTS

The computational experiments were motivated by the following questions:

- How do the protein function classifiers based on reduced alphabet representations of protein sequences compare with those based on the original 20-letter amino acid alphabet in terms of classification accuracy? How do the different reduced alphabet representations of protein sequences compare with one another?
- Do the motifs identified by the decision tree learning algorithms correspond to more or less the same portions of the sequences in the data set independent of the alphabet used to represent the sequences? Or can different choices of the alphabet uncover sequence regularities useful for function prediction from different parts of the sequences?
- Can different choices of alphabet for sequence representation provide different insights regarding the underlying protein structure function relationships? Are some reduced alphabets more natural than others in capturing sequence regularities that are predictive of protein function?

Alphabet	Size	Mean Error	Standard Deviation
Hydrophobicity	2	43.16	10.2
Charge	3	32.18	8.21
Volume	4	7.18	3.17
Mass	4	10.13	4.54
Hyrophobicity-Charge	4	12.03	7.70
Hyrophobicity-Mass	7	8.22	2.51
Charge-Volume	7	7.07	2.69
Charge-Mass	7	4.14	3.91
Volume-Mass	8	10.7	3.80
Hyrophobicity-Volume	8	7.61	2.81
Random 2	2	55.34	4.01
Random 3	3	40.23	3.41
Random 4	4	10.52	3.65
Random 5	5	9.45	3.76
Random 6	6	8.60	3.94
Random 7	7	8.02	2.91
Random 8	8	7.84	2.37
Random 9	9	7.73	4.32
Random 10	10	7.59	3.43
Blosum50 2	2	48.46	13.4
Blosum50 4	4	10.06	4.00
Blosum50 8	8	6.74	3.25
Blosum50 10	10	4.93	2.75
Blosum50 15	15	4.12	2.46
Amino Acid	20	5.87	3.89
Average		15.05	4.49

We used the estimated *error* of classification (on test data not used for training the classifiers) as the primary performance measure to evaluate protein function classifiers generated using alternative alphabets for protein sequence representation. Estimated error of a classifier is computed as the percentage of instances in a *test* data set that are classified incorrectly. The error estimates were averaged over 50 independent runs of the algorithm, for each choice of the alphabet. Each run randomly selected two thirds of the instances in data for training the classifier and the remaining one third of the data as for testing the classifier (and estimating error). The error estimates (mean and standard deviations) computed from 50 independent runs for each choice of alphabet are summarized in Table 6.

Using the standard 20-letter amino acid alphabet resulted in an average error rate of about 5.9%. At alphabet sizes of two or three, the classification error is at least 5-fold higher than that of the 20-letter alphabet. In contrast, several alphabet choices where the alphabet size is greater than or equal to 4 result in fairly low error rates. For example, a 4-letter alphabet based on volume results in an error rate of 7%. A 7-letter alphabet based on charge and mass yields an error rate of approximately 4%. Interestingly, several of the best-performing randomly generated alphabets (e.g., Random 8a, Random 5a) resulted in error rates comparable to those based on physico-chemical properties or the BLOSUM-50 matrix. In particular, note that several reduced alphabets (i.e., those based on charge-mass, Blosum-10 and Blosum-15, and some of the better-performing random alphabets) gave error rates that were compared favorably with that obtained using the 20-letter amino acid alphabet (5.9%). Preliminary examination of some of the best-performing random alphabets suggests that the corresponding amino acid groupings do not have an obvious relationship to the groupings based on physico-chemical properties or the BLOSUM-50 matrix. A more thorough analysis of the properties shared by the amino acids is in progress.

Table 6: The size, mean error, and standard deviation for each of the different alphabets. Size refers to the actual number of different partions of the twenty letter amino acid alphabet. Mean error shows the mean error over 50 individual runs of building decision trees using motifs of these reduced random alphabets. Standard deviation refers to the standard deviation between the 50 individual runs.

Figure 2 shows how the error rate of the decision trees varies with alphabet size. The classification error drops rapidly as alphabet size reaches 4 (regardless of the particular choice of the alphabet).

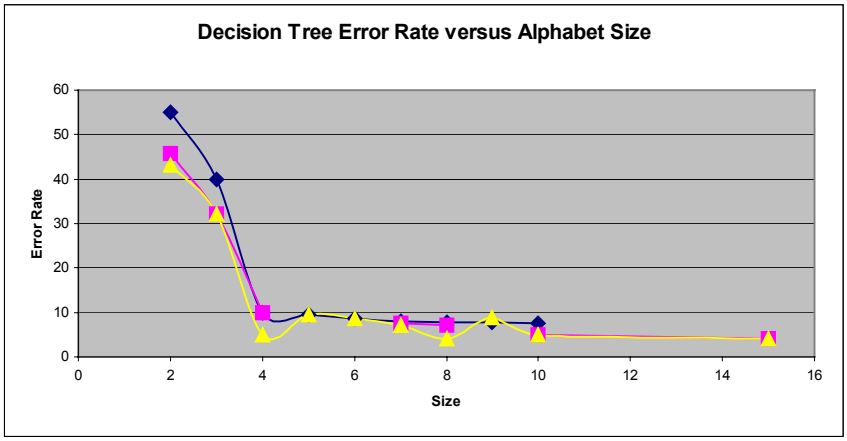

Figure 2: Plot of alphabet size versus error for different alphabet types. The purple line shows the results for the random alphabets. The pink line shows the results for alphabets constructed using biological knowledge. The yellow line shows the results for the alphabets that resulted in the lowest error rate for each alphabet size.

In light of the results presented above, it is interesting to examine whether different alphabet choices yield sequence motifs that essentially *cover* the same or different parts of the protein sequences in question. If the motifs picked out by the decision trees resulting from different choices of alphabet *cover* relatively disjoint portions of the sequences in question, it raises the possibility that different alphabets might provide different complementary insights into the underlying structure-function relationships. Hence, we examined the distribution of motifs that show up most often in decision trees for several different choices of alphabet.

Figure 3 shows the 3-dimensional structure of Serine Protease 2, from the S2A family. Motifs identified by the decision trees generated using 3 different alphabets (the 4-letter hydrophobicity-charge, 4-letter mass, and a 4-letter random alphabet, each of which had an error rate of approximately 10% in classifying proteins into one of the 11 families) are shown superimposed on the structure. Note that the sequence motifs picked out by the decision trees with comparable error rates but based on different alphabets *cover* different parts of the sequence and correspond to different parts of the structure.

The hydrophobiciy-charge alphabet picked a motif that extended into the core of the protein (Figure 3-a), the mass alphabet chose a motif that was located on the surface of the protein (Figure 3-b), and the random alphabet picked a motif that corresponds to a site that is distinct from the previous two (Figure 3-c). Yet each of these three decision trees correctly classified 100% of the S2A proteins (although their overall error rate for the 11 families was approximately 10%). These results make sense in light of our knowledge of protein structures: hydrophobicity is a critical property for the core of the protein; properties such as volume and mass affect the surface shape of the protein which plays a critical role in protein function (since it largely determines binding and docking). This suggests that different choices of the alphabet can uncover different sequence regularities that are predictive of protein function based on different properties or combinations of physico-chemical properties of amino acids that are conserved to different degrees at different portions of the proteins.

Figure 3-a

Figure 3-b

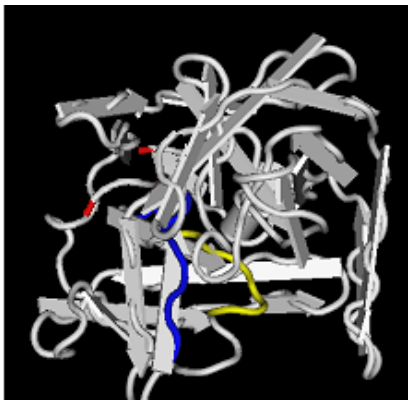

Figure 3-c

Figure 3a-c: Three different motifs from Serine Protease 2 from the S2A family (PDB ID: 2SFA). See www.cs.iastate.edu/~honavar/papers/carson-cbgi02.pdf for the color image. Red indicates the starting and end terminal positions . Yellow indicates the active site, starting in position 31 with a length of 6. Blue indicates the motif. Figure 3-a shows the motif from the Hydrophicity-Charge alphabet. It starts in position 138 and has a length of 12 amino acids. Figure 3-b shows the moitf from the Mass alphabet. It starts in position 119 and has a length of 12 amino acids. Figure 3-cshows the motif from the Random 4 alphabet. It starts in position 13 and has a length of 12 amino acids.

3. SUMMARY AND DISCUSSION

Conserved sequence motifs in protein families constitute an important source of information for understanding protein structure-function relationships. Reduced amino acid alphabets constructed by grouping amino acids on the basis of the values of their physico-chemical properties have proven useful in studies of proteins in a number of different contexts. For instance, reduced alphabets have been shown to be

useful in studies of protein folding and protein recognition [Chan, 1999; Murphy et al., 2000], protein design [Riddle et al., 1997], phylogeny [Naylor and Brown, 1998], sequence matching and retrieval [Smith and Smith, 1990], among others. Schafmeister et al. [1997] showed that a 108 amino acid, 4 helix bundle protein could be synthesized using a 7-letter reduced alphabet sequence. Riddle et al. [1997] showed that a functional β-sheet protein (the SH3 domain) can be largely encoded by a 5-letter reduced amino acid alphabet but not by a 3-letter alphabet. Murphy et al. [2000] estimate that foldable sequences for most proteins can be represented using a 10 or 12 letter reduced alphabet. Using information-theoretic arguments, Romero et al. [1999] show that the minimal alphabet size necessary for specifying globular proteins that occur in nature is 10. Interestingly, it is also believed that early protein evolution operated on a universe of proteins based on a relatively small alphabet of amino acids [Riddle et al., 1997]. The results presented in this paper demonstrate that sequence motif-based protein classifiers reduced alphabet representations of constructed from protein sequences can match and in some cases even outperform those constructed from the 20-letter amino acid representations of sequences. In particular, we find that alphabet sizes of 4 or 5 suffice for reliably assigning protease sequences into one of the 12 protease families. Surprisingly, we find that on the data set used in this study, some of the random alphabets (based on random partitions of the 20-letter alphabet) perform at rates near those based on physico-chemical properties. Alphabet size, regardless of the particular choice of the alphabet, is strongly correlated with the performance of the decision tree classifier. Preliminary results show that different choices of the alphabet might provide different, but complementary insights into the principles that underlie protein-structure relationships.

Some directions for future work include:

- Exploration of additional alphabets for protein sequence representation based on properties of amino acids other than the ones examined in this paper. Some obvious properties include: surface area, solubility, bulkiness, refractivity, polarity, three-dimensional structure, etc. In the case of continuous valued properties, it would be interesting to systematically vary the the alphabet size as well as the scheme used for quantization of the continuous value into discrete bins.
- Systematic study of structure-function relationships discovered from reduced alphabet representations of protein sequences over a much broader set of protein families to examine the general applicability of the results reported here based on a study of 11 protease families.
- More in-depth examination of the results, addressing questions such as: Why are certain properties more conserved at particular sites and not others? Why are some active regions being picked out by decision trees generated using some alphabets and not others?
- Application of approaches similar to those used in this study to the discovery of sequence features based on different reduced alphabet representations of protein sequences that correspond to functionally significant 3-dimensional structural features of proteins
- Integration of the resulting tools with visualization routines for exploratory analysis of macro-molecular structure-function relationships.

References

1. Bailey, T. L. and Gribskov, M. (1998) Combining evidence using p-values: application to sequence homology searches", Bioinformatics, 14(48-54).
2. Bairoch A., and Apweiler R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res. 28:45-48. (http://www.expasy.ch/sprot/sprot-top.html)
3. Barrett, A.J., Rawlings, N.D., and Woessner, J.F. (1998). Handbook of proteolytic enzymes. New York: Academic Press.
4. Chan, H.S. (1999). Folding Alphabets. *Nature Structural Biology* **6,** 994 – 996
5. Lide, D. R. CRC Handbook of Chemistry and Physics, CRC Press, Inc., Cleveland, Ohio 58 (2001).
6. Henikoff, S. and Henikoff, J.G. (1992) Proc. Natl Acad. Sci. USA, 89, 10915 – 10919.
7. Kyte, J. and Doolittle, R. (1982) A Simple Method for Displaying the Hydropathic Character of a Protein, J. Mol Biol. 157 105-132.
8. Mitchell, T. (1997). Machine Learning. New York: McGraw Hill.
9. Murphy, Lynne, Wallqvist, Anders and Levy, Ronald R. (2000) Simplified amino acid alphabets for protein recognition and implications for folding. Protein Engineering vol. 13, no. 3, pp149-152.
10. Naylor, G.J.P. and W.M. Brown. 1998. Amphioxus Mitochondrial DNA, chordate phylogeny, and the limits of inference based on comparisons of sequences. i 47(1):61-76
11. Quinlan, J. R. (1992). C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo, CA.
12. Rawlings, N.D. and Barrett, A.J. (1993) Evolutionary families of peptidases. Biochem. J. 290:205-218. (http://www.merops.co.uk/)
13. Riddle, D.S., Santiago, J.V., Bray-Hall, S.T., Doshi, N., Grantcharova, V.P., Yi, Q., and Baker, D. (1997). Functional Rapidly Folding Proteins from Simplified Amino Acid Sequences. Vo. 4. No. 10. pp. 805-809
14. Romero, P., Obradovic, Z., and Dunker, A.K. (1999). Folding minimal sequences: The lower bound for sequence complexity of globular proteins. FEBS Letters 462: 362-367.
15. Schafmeister, C.E., LaPorte, S.L., Mierke, L.J.W., and Stroud, R. M. (1997). Nat. Struct. Biol. 4, 1039-1046
16. Smith, R.F., and Smith, T.F. (1990). Automatic generation of primary sequence patterns form sets of related protein sequences. *PNAS*, 87:118-122.
17. Sokal, R. R. and Minchener, C. D. (1958) A statistical method for evaluating systematic relationships. University of Kansas Scientific Bulletin 28:1409 – 1438.
18. Taylor, W.R. (1986) The Classification of Amino Acid Conservation, J. Theor. Biol. 119, 205-218.
19. Wang, D., Wang, X., Honavar, V., and Dobbs, D. (2001). 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.
20. Wang X., Schroeder, D., Dobbs, D., and Honavar, V. (2002). Data-Driven Discovery of Protein Function Classifiers: Decision Trees based on MEME motifs outperform PROSITE patterns and profiles on Peptidase families. In: Proceedings of the Conference on Computational Biology and Genome Informatics (CBGI-02), In press.
21. Zamyatin, A. A. (1972) Protein Volume in Solution, Prog. Biophys. Mol. Biol. 24107-123.