Cross-validation of protein structural class prediction using statistical clustering and neural networks

BRENT A. METFESSEL,¹ PETER N. SAURUGGER,² DONALD P. CONNELLY,¹
AND STEPHEN S. RICH¹

¹ Department of Laboratory Medicine and Pathology, University of Minnesota, Minneapolis, Minnesota 55455

² Department of Data Management, Plant Breeding Division, Pioneer Hi-Bred International, Johnston, Iowa 50131

(RECEIVED January 16, 1993; REVISED MANUSCRIPT RECEIVED March 17, 1993)

Abstract

We present an approach to predicting protein structural class that uses amino acid composition and hydrophobic pattern frequency information as input to two types of neural networks: (1) a three-layer back-propagation network and (2) a learning vector quantization network. The results of these methods are compared to those obtained from a modified Euclidean statistical clustering algorithm. The protein sequence data used to drive these algorithms consist of the normalized frequency of up to 20 amino acid types and six hydrophobic amino acid patterns. From these frequency values the structural class predictions for each protein (all-alpha, all-beta, or alpha-beta classes) are derived. Examples consisting of 64 previously classified proteins were randomly divided into multiple training (56 proteins) and test (8 proteins) sets. The best performing algorithm on the test sets was the learning vector quantization network using 17 inputs, obtaining a prediction accuracy of 80.2%. The Matthews correlation coefficients are statistically significant for all algorithms and all structural classes. The differences between algorithms are in general not statistically significant. These results show that information exists in protein primary sequences that is easily obtainable and useful for the prediction of protein structural class by neural networks as well as by standard statistical clustering algorithms.

Keywords: back-propagation neural network; cross-validation; Euclidean distance; learning vector quantization neural network; McNemar's test; protein structural class prediction

The protein structure prediction problem remains one of the most important problems in molecular biology. In general, the function of a gene is based on the function of the protein for which it encodes; in turn, the function of a protein is based on its structure—that is, the conformation in which the protein folds. Thus, protein structure plays a key role in cell biology, molecular biology, and biochemistry.

Presently, protein structures are determined using X-ray crystallographic and NMR techniques. The use of X-ray techniques requires crystallization of the protein, which can be quite difficult. In addition, X-ray diffraction patterns cannot always be interpreted. For patterns that are interpretable, it can take months or even years to determine the structure of a single protein. NMR does not require crystallization and can be an alternative struc-

ture determination method for relatively small proteins. However, many months may pass and a large amount of supercomputer time may be required before a structure is determined (Kinoshita, 1990).

As a result, although more than 80,000 loci exist in GenBank in 1992, all of which can be translated into protein primary sequences, only about 800 protein structures exist in the Brookhaven Protein Data Bank (Bernstein et al., 1977). Many of the structures in the Protein Data Bank are minor modifications of previous structures such that only about 200 truly unique structures presently exist in the Data Bank. The protein structure prediction problem takes on even more importance when one considers all the sequence data that will be generated by the Human Genome Project. Unless new methods to predict protein structure are developed in the near future, the gap between known sequences and structure will become overwhelming.

The prediction of protein structural class is an important component of the overall protein structure prediction

Reprint requests to: Brent Metfessel, O'Pin Systems, International Plaza, Suite 635, 7900 International Drive, Bloomington, Minnesota 55425.

problem (Levitt & Chothia, 1976). Several early studies have tried to predict protein structural class directly from the sequence (Sheridan et al., 1985; Klein, 1986; Klein & Delisi, 1986; Nakashima et al., 1986). These studies have used discriminant analysis or other statistical distance measures to predict structural class from the amino acid composition. Rough measures of hydrophobic residue periodicity were sometimes also included. Some of these investigations have claimed to predict the structural class of a protein with up to 83% accuracy.

These studies have been difficult to compare to each other for several reasons. First, the protein sets used vary widely in both size and extent of sequence homology. Second, the number of structural classes in which to categorize the proteins has varied from study to study, ranging from three to five structural classes. Third, criteria for dividing proteins into their structural classes has differed. For example, Nakashima et al. (1986) classified a protein as "alpha-type" if the protein is known to contain greater than 15% alpha-helices and less than 10% beta-strands. Klein and Delisi (1986), on the other hand, required their alpha-class proteins to contain at least 40% alpha-helices and not more than 5% beta-strands. Fourth, cross-validation methods have varied widely, ranging from leave-one-out methods to single-test-set methods. Nakashima et al. (1986) used no cross-validation. In a later study, Chou (1989) also attempted to predict protein structural class using statistical methods based on differential amino acid frequencies. Using a set of 64 proteins, his algorithm correctly predicted 84% of the all-alpha proteins, 80% of all-beta proteins, 79% of alpha + beta, and 75% of alpha-beta proteins. Chou obtained 12 proteins not in the original database and reached a classification accuracy of 83%, but no information is given with respect to protein selection.

A recent attempt at protein structural class prediction (Zhang & Chou, 1992) used the maximum component coefficient principle to predict the structural class of the proteins in the Chou (1989) data set, based on the frequencies of 20 amino acid types. This algorithm correctly predicted 100% of the all-alpha proteins and obtained an overall accuracy rate of 82.8%, which was about 3.1% higher than the general accuracy rate reported by Chou (1989), but no cross-validation of the results was presented.

The investigation to be described differs from these previous studies in that: (1) back-propagation as well as other neural network models are used, and the results are compared to those from a statistical distance metric; and (2) randomized cross-validation and statistical analysis is used to evaluate the performance of the methods on the 64 proteins used in the two most recent studies.

Unlike the two most recent studies using the Chou (1989) protein set, six specific hydrophobic patterns are added to the list of structural class predictors. This is done because residue hydrophobicity represents the major driving force behind protein folding (Dill, 1990). A protein

tends to fold so that water is excluded from the center of the molecule, making the protein's core very compact and hydrophobic.

According to the theory of Lim (1974), the hydrophobic patterns $(i, i + 2)$ and $(i, i + 2, i + 4)$ —that is, a hydrophobe occurring at every other amino acid along the chain—tend to occur more frequently in beta-sheets, while the patterns $(i, i + 3)$, $(i, i + 3, i + 4)$, and $(i, i + 1, i + 4)$ tend to occur more often in alpha-helices. The latter concept is an extension of the concept of the "helical wheel" or amphipathic alpha-helix (Schiffer & Edmundson, 1967), where the helix has a hydrophilic side (with amino acid side chains extending into the solvent) and a hydrophobic side (where side chains extend into the non-polar core of the protein).

In this investigation, the frequencies of 20 amino acids and six hydrophobic residue patterns $(i, i + 2; i, i + 3; i, i + 2, i + 4; i, i + 1, i + 4; i, i + 3, i + 4; i, i + 5)$ are used for the prediction of protein structural class by three methodologies: a modified Euclidean statistical clustering algorithm (Widman & Pierce, 1990), a back-propagation neural network (Rumelhart et al., 1986), and a learning vector quantization (LVQ) network (Kohonen et al., 1988).

The intent of this investigation is to discover: (1) whether the selected primary sequence information is useful for the prediction of high-level protein structural classes in a randomized cross-validation paradigm, and (2) whether a significant difference exists in the ability of the above three algorithms to utilize the information in predicting protein structural class.

Results

The three algorithms were trained on 56 training examples and tested on 8 examples, in each of 24 training and testing sessions. The training and testing sets were created from the 64 proteins used originally by Chou (1989) (Table 1). The inputs representing each protein sequence in the training and test sets were determined from the amino acid frequencies as well as the frequency of occurrence of the above hydrophobic patterns (per 100 residues). These input elements are also termed "predictors," because they are used as the bases for the prediction of structural class by each algorithm.

Results for 26 inputs

The results of the study using all 26 predictors are summarized in Table 2. Preliminary results have been reported (Metfessel & Saurugger, 1993). The results reported here for the back-propagation network differ slightly due to a more thorough examination of neural network hidden unit parameters and a small change in the sequence database representation for multiple-subunit proteins. The

Table 1. Protein sequence files used in this study along with Brookhaven PDB or PIR identification codes

ID code	Protein name and source	ID code	Protein name and source
I. All-alpha proteins		III. Alpha-beta proteins	
PDB-3CPV	Calcium-binding parvalbumin B (carp)	PDB-2ACT	Actinidin (kiwi fruit)
PDB-156B	Cytochrome B562 (<i>E. coli</i>)	PDB-2B5C	Cytochrome B5 (bovine)
PDB-5CYT	Cytochrome c (albacore tuna)	PDB-1FDX	Ferredoxin (<i>P. aerogenes</i>)
PDB-2C2C	Cytochrome c2 (<i>R. rubrum</i>)	PDB-1HIP	Oxidized high-potential iron protein (<i>C. vinosum</i>)
PDB-155C	Cytochrome c ₅₅₀ (<i>P. dentrificans</i>)	PDB-1INS	Insulin (porcine)
PDB-05C1	Cytochrome c ₅₅₅ (<i>C. thiosulfatophilum</i>)	PDB-2LZM	Lysozyme (<i>E. coli</i> infected with bacteriophage T4)
PDB-1HMQ	Methemerythrin (<i>T. dyscritum</i>)	PDB-8LYZ	Lysozyme (hen egg white)
PDB-1HRB	Hemerythrin B (<i>G. gouldii</i>)	PDB-9PAP	Papain (papaya)
PDB-2MHR	Myohemerythrin (<i>T. zostericola</i>)	PDB-1P2P	Phospholipase A2 (porcine)
PDB-2MHB(A)	Methemoglobin alpha chain (horse)	PDB-1RNS	Ribonuclease S (bovine)
PDB-2MHB(B)	Methemoglobin beta chain (horse)	PDB-2SNS	Staphylococcal nuclease (<i>S. aureus</i>)
PIR-HAHU	Hemoglobin alpha chain (human)	PDB-2SSI	Streptomyces subtilisin inhibitor (<i>S. albobruiensis</i>)
PIR-HBHU	Hemoglobin beta chain (human)	PDB-7TLN	Thermolysin (<i>B. thermoproteolyticus</i>)
PIR-HGHUA	Hemoglobin gamma chain (human)	PDB-6PTI	Bovine pancreatic trypsin inhibitor
PDB-0HBG	Hemoglobin (glycera)	PDB-3ADK	Adenylate kinase (porcine)
PDB-2LHB	Cyano/methemoglobin V (sea lamprey)	PDB-6ADH	Alcohol dehydrogenase (horse)
PIR-GGIC3	Hemoglobin III (midge)	PDB-2CAB	Carbonic anhydrase B (human)
PDB-5MBN	Myoglobin (sperm whale)	PDB-1CAC	Carbonic anhydrase C (human)
PDB-1MBS	Myoglobin (seal)	PDB-5CPA	Carboxypeptidase A (bovine)
II. All-beta proteins		PDB-1CPB	Carboxypeptidase B (bovine)
PDB-2CHA	Alpha chymotrypsin A (bovine)	PDB-4DFR	Dihydrofolate reductase (<i>E. coli</i>)
PDB-3CNA	Concanavalin A (jack bean)	PDB-4FXN	Flavodoxin (<i>Clostridium</i> MP)
PDB-3EST	Elastase (porcine)	PDB-4GPD	D-glyceraldehyde-3-phosphate dehydrogenase (lobster)
PDB-1NXB	Neurotoxin B (Erabutoxin B) (sea snake)	PDB-2GDI	D-glyceraldehyde-3-phosphate dehydrogenase (<i>B. stearothermophilus</i>)
PDB-3FAB(H)	Lambda immunoglobulin Fab', heavy chain (human)	PDB-6LDH	Lactate dehydrogenase (dogfish)
PDB-3FAB(L)	Lambda immunoglobulin Fab', light chain (human)	PDB-3PGK	Phosphoglycerate kinase (baker's yeast)
PDB-1MCG	Bence-Jones immunoglobulin (human)	PDB-1RHD	Rhodanese (bovine)
PDB-1REI	Bence-Jones immunoglobulin, REI variable portion (human)	PDB-1SBT	Subtilisin BPN (<i>B. amyloliquefaciens</i>)
PDB-2APP	Penicillopepsin (<i>P. janthinellum</i>)	PDB-1SRX	Thioredoxin (<i>E. coli</i>)
PDB-2PAB	Prealbumin (human)	PDB-1TIM	Triose phosphate isomerase (chicken)
PDB-2SGA	Proteinase A (<i>S. griseus</i>)		
PDB-3SGB	Proteinase B (<i>S. griseus</i>)		
PDB-4RXN	Rubredoxin (<i>C. pasteurianum</i>)		
PDB-2SOD	Cu,Zn superoxide dismutase (bovine)		
PDB-4PTP	Beta trypsin (bovine)		

results for each algorithm represent the average for 24 runs (with each run performed on a different training/test set pair). The neural networks (a three-hidden-unit back-propagation network and a three-hidden-unit-per-output LVQ network) appeared to perform slightly better than the Euclidean statistic, with a 1.5–2.0% improvement in accuracy on the test sets. In order to evaluate the influence that the random sampling has on the prediction accuracy, the scores can be broken down into average scores for each group of 8 test sets, leading to a total of three scores for the 24 test sets. In this manner all 64 proteins are included once per score, and the range of scores reflects the variation caused by the composition of the training sets. For the modified Euclidean distance metric, the corresponding range of prediction accuracies is 70.3–

78.1% (from 45/64 correct to 50/64 correct). The back-propagation network results range from 73.4% to 78.1% (47/64 correct to 50/64 correct) and the LVQ prediction accuracy range is 73.4–79.7% (47/64 correct to 51/64 correct).

In addition, the training set prediction accuracies differed between the methods, with the Euclidean distance metric obtaining an accuracy of 93.5%, while the neural networks learned 99.6–100% of the correct classes of the training proteins.

When McNemar's test is applied to the test set results, none of the differences between the three algorithms are statistically significant ($P > 0.05$). All Matthews correlations are statistically significant, with the coefficients for the all-beta class (back-propagation and LVQ neural net-

Table 2. Results of cross-validation using statistical and neural network methods for protein structural class prediction: 26 predictors

Algorithm	Prediction accuracy (%)		Matthews correlation coefficient ^a		
	Training	Testing	All-alpha	All-beta	Alpha-beta
Back-propagation	100	75.5	0.76**	0.58*	0.51**
LVQ	99.6	76.0	0.77**	0.61*	0.52**
Euclidean	93.5	74.0	0.65**	0.65**	0.49**

^a The Matthews correlation coefficients were calculated based on the test sets only. * $P < 0.05$. ** $P < 0.01$.

works) being significant to $P < 0.05$ and the other coefficients being significant to $P < 0.01$.

Simplification of predictors

The effects of simplifying the inputs to the algorithms were studied. Predictors where the mean frequencies for the three classes are similar enough to each other such that the mean frequency for each class is within one standard deviation of the mean frequency for both of the other classes were dropped from the list. As a result, the hydrophobic pattern ($i, i + 5$) as well as the amino acids Asp, Glu, Gly, Ile, Asn, Tyr, Trp, and Val were excluded from use for the next analysis, leaving a total of 17 predictors. Table 3 depicts the means and standard deviations of all 26 amino acid and hydrophobic pattern frequencies for each structural class. Some differences between the structural classes are noted in the frequencies of the "beta-sheet" hydrophobic patterns ($i, i + 2$) and ($i, i + 2, i + 4$), with the all-beta proteins having a somewhat higher frequency. A larger difference in pattern frequencies is noted in the ($i, i + 3$) and ($i, i + 1, i + 4$) patterns, the so-called "alpha-helix" patterns, with the all-alpha class having the highest frequencies.

In terms of amino acid frequencies, Ala, Lys, and Leu have the highest frequencies for the all-alpha proteins, and Ser and Thr are highest for the all-beta class. (Gly also has a high mean frequency for the all-beta class, but the high standard deviation makes it statistically a weak predictor.) The alpha-beta proteins often have amino acid frequencies intermediate to the other two structural classes. However, Arg seems to have a higher value in the alpha-beta proteins overall than in proteins of the other two classes.

Results for 17 inputs

The results from studies using the 17-predictor set are shown in Table 4. A two-hidden-unit back-propagation and a two-hidden-unit-per-output-unit LVQ network were utilized for this section of the study because these

Table 3. Means and standard deviations for the frequencies of each predictor for each structural class^a

Predictor ^b	Structural class					
	All-alpha		All-beta		Alpha-beta	
	Mean	SD	Mean	SD	Mean	SD
$i, i + 2$	3.77	1.27	5.20	1.10	4.94	2.05
$i, i + 3$	6.11	2.42	2.50	1.26	4.36	1.31
$i, i + 2, i + 4$	0.44	0.57	1.16	0.59	0.60	0.63
$i, i + 5$	1.29	0.75	1.25	1.07	1.69	0.96
$i, i + 1, i + 4$	2.21	1.17	0.81	0.51	1.27	1.34
$i, i + 3, i + 4$	1.52	0.96	0.55	0.63	1.52	0.85
Ala	11.65	4.40	6.64	3.10	8.96	3.86
His	4.33	2.51	1.82	1.37	2.15	1.41
Lys	10.47	2.86	4.53	2.51	6.69	3.04
Ser	4.72	2.43	11.38	4.22	6.99	2.73
Thr	4.98	1.86	9.06	2.46	5.56	1.90
Cys	0.97	0.73	3.35	3.24	3.27	3.95
Leu	8.76	3.62	5.86	2.11	7.00	2.74
Pro	3.35	1.22	4.98	1.95	4.31	1.85
Arg	2.11	1.09	2.49	1.61	3.74	2.31
Asp	6.89	2.55	5.08	4.22	5.80	2.08
Glu	5.43	2.71	3.77	3.04	5.76	2.80
Gly	7.94	3.09	10.65	3.77	8.42	2.37
Ile	3.74	2.54	4.52	1.77	5.43	2.56
Gln	2.51	1.38	4.27	3.11	3.07	1.51
Asn	3.97	1.72	4.44	1.75	4.99	2.30
Phe	4.98	2.47	3.01	1.24	3.36	1.49
Tyr	2.71	1.48	4.01	1.73	4.04	2.28
Trp	1.36	0.89	1.53	0.83	1.56	1.21
Val	6.75	2.71	7.95	2.45	7.23	2.76
Met	2.14	2.02	0.65	0.51	1.67	1.14

^a Numerical values are given in terms of the number of occurrences per 100 residues. Predictors that were excluded from the analysis using 17 predictors are underlined.

^b The first six row headers designate hydrophobic residue patterns based on a modification of the Rose et al. (1985) criteria.

networks were the simplest networks that gave the highest prediction accuracies. The training set prediction accuracies for the Euclidean statistical algorithm show the greatest differences from the performance on the 26-input

Table 4. Results of cross-validation using statistical and neural network methods for protein structural class prediction: 17 predictors

Algorithm	Prediction accuracy (%)		Matthews correlation coefficient ^a		
	Training	Testing	All-alpha	All-beta	Alpha-beta
Back-propagation	100	78.1	0.75**	0.68**	0.56**
LVQ	99.0	80.2	0.78**	0.72**	0.60**
Euclidean	90.2	76.6	0.67**	0.70**	0.55**

^a The Matthews correlation coefficients were calculated based on the test sets only. ** $P < 0.01$.

sets, with a decrease of 3.3% (90.2%). All algorithms depict improvement in test set prediction accuracies, with the LVQ network showing the greatest improvement, obtaining a prediction accuracy of 80.2% (4.2% greater than the average result for its runs with the 26-input sets).

The range of the average prediction accuracies for each group of eight test files is 75.0–78.1% (48–50 correct) for the modified Euclidean statistic, 73.4–84.4% (47–54 correct) for the back-propagation network, and 76.6–82.8% (49–53 correct) for the LVQ network. All Matthews correlation coefficients in Table 4 are significant to $P < 0.01$. (The all-beta class, having many fewer members than the alpha-beta class, requires a higher correlation coefficient for significance than that needed for the alpha-beta class.)

Using McNemar's test, none of the improvements on the test sets using 17 predictors are statistically significant as compared to 26 predictors. When comparing the different test set results within the 17-predictor space, none of the algorithms performed significantly better than the others.

The Matthews correlation coefficients, while generally higher for those methods that obtained higher test set accuracies, show similar trends for both the 26-predictor and 17-predictor studies, with all-alpha > all-beta > alpha-beta for the back-propagation and LVQ neural network algorithms, and all-alpha = all-beta > alpha-beta for the Euclidean distance metric, but none of these trends are statistically significant when the two-sample test of (independent) proportions is applied. The relatively high number of false positives for the alpha-beta class lowers its Matthews correlation coefficient; this factor is not accounted for in the two-sample test of proportions.

Number of neural network hidden units

Table 5 shows the variations in the back-propagation prediction accuracies as a function of the number of hidden units in the network. Within each predictor (input) space category, the network results do not vary by more than 2% in their prediction accuracies with the exception of the network with no hidden units (two-layer networks), which for both input space categories show a 4.7% drop in prediction accuracy from the best score. The reported results for Tables 2 and 4 were obtained from the highest scoring networks; where a tie exists, the network with the least number of hidden units was chosen. The "optimal" number of hidden units per output unit for the LVQ network was chosen in a similar fashion.

Independence of the algorithms

Intuitively, when comparing two algorithms, independence can be tested by finding out how many of the proteins that were noted as incorrect predictions were predicted incor-

Table 5. Prediction accuracy for back-propagation networks with different numbers of hidden units

No. of hidden units	Prediction accuracy (%) (test sets)	
	26 inputs	17 inputs
0	70.8	73.4
2	74.5	78.1
3	75.5	77.6
4	74.5	77.6
5	75.5	76.6
6	75.5	78.1
7	74.0	77.6
9	75.5	78.1

rectly by both algorithms. The measure used is the "percent coincidence," which is the value obtained when the total number of proteins predicted incorrectly by both algorithms is divided by the total number of proteins predicted incorrectly by either or both algorithms, expressed as a percentage.

In terms of the results for the 17-predictor space, the Euclidean statistic as compared to the LVQ neural network depicts a 43% coincidence of incorrectly predicted proteins. Similarly, for the Euclidean statistic vs. the back-propagation network, there exists a 50% coincidence of incorrect predictions. For the two neural network algorithms, there exists a 45% coincidence of incorrect predictions.

Incorrectly classified proteins

Four proteins were incorrectly classified in all test sets for both the 26-input and the 17-input categories: sea lamprey hemoglobin V (an all-alpha protein), rubredoxin, and Cu,Zn superoxide dismutase (all-beta proteins), and staphylococcal nuclease (an alpha-beta protein).

Homology

To determine the effect of sequence homology on the results, we measured the amount of homology for each structural class. Table 6 depicts the homologous pairs of proteins within the 64-protein set used in the study. The criterion for a homologous pair is a residue identity of 25% or greater for an alignment of 80 or more residues (Sander & Schneider, 1991). The all-alpha class contains the most homology, with 26 homologous pairs. Thus, 15.2% of all possible unique pairs are homologous in the all-alpha class. (A protein paired to itself is not counted as a "unique pair.") The all-beta class shows nine homologous pairs, meaning that 8.6% of all possible unique pairs are homologous. The alpha-beta class shows the

Table 6. Homologous pairs within the three structural classes

All-alpha		All-beta		Alpha-beta	
Proteins	%ID ^a	Proteins	% ID	Proteins	% ID
2MHBA-HAHU	87	1MCG-3FABL	77	2CAB-1CAC	59
2MHBB-HBHU	83	3SGB-2SGA	58	2GD1-4GPD	51
1MBS-5MBN	83	1REI-1MCG	45	2ACT-9PAP	47
1HRB-1HMQ	75	1REI-3FABL	43	5CPA-1CPB	46
HGHUA-HBHU	72	4PTP-2CHA	40		
HGHUA-2MHBB	71	3EST-2CHA	37		
2MHR-1HRB	45	4PTP-3EST	34		
2MHR-1HMQ	44	1MCG-3FABH	28		
2MHBB-2MHBA	43	3FABL-3FABH	27		
2MHBB-HAHU	42				
HBHU-2MHBA	42				
HBHU-HAHU	42				
155C-2C2C	42				
HGHUA-2MHBA	41				
HGHUA-HAHU	40				
5CYT-2C2C	35				
5CYT-155C	30				
2LHB-2MHBA	30				
1MBS-HAHU	27				
1MBS-HBHU	26				
2LHB-HAHU	26				
2LHB-HGHUA	25				
1MBS-2MHBB	25				
5MBN-2MHBA	25				
5MBN-HAHU	25				
5CYT-05C1	25				

^a This column refers to the percent residue identity between the sequence pairs after alignment using the PAM-256 matrix within the program FASTDB (Brutlag et al., 1990).

least homology with four homologous pairs (0.9% of all possible pairs). No homologous sequences are found in aligned pairs in which the members of the pair belong to different structural classes.

Discussion

Comparison to previous work

Both the statistical and the neural network algorithms performed well their ability to learn training sets and in cross-validation runs. The algorithms used here are difficult to compare to previous studies for the reasons mentioned in the Introduction; however, the performance of the Euclidean statistic on the training sets (93.5% correct overall) compared favorably to the results of Chou (1989), where the prediction accuracy ranged from 84% (for the all-alpha proteins) to 75% (for the alpha-beta proteins), as well as to the newest study by Zhang and Chou (1992), which had an overall prediction accuracy of 83%. These previous studies utilized only amino acid frequencies in their structural class prediction paradigms. In addition, the comparisons are not exact due to the fact that these

previous studies divided the proteins into four structural classes, whereas in the present investigation there exists three classes. This may explain part of the wide differences between this study and previous ones.

Nakashima et al. (1986), who also based their prediction scheme on amino acid frequencies and used a Euclidean statistic, obtained an overall score of 70% correct, dividing the proteins into five structural classes, using the same classes as Chou (1989) except with the addition of the "irregular" class. However, Nakashima et al. stated that the worst prediction accuracies were obtained between the alpha + beta and the alpha/beta classes; as a result, the two classes were unified into a single "ab complex" class, and the irregular class was excluded. The prediction accuracy for the three classes all-alpha, all-beta, and "ab complex" rose to 80%. It needs to be noted again that this figure was obtained on a different set of proteins than the Chou set and was not cross-validated. However, it seems a fair conclusion that given our results, and given the hydrophobic pattern frequencies displayed in Table 3 that appear to confirm Lim's theories on secondary structure, the addition of hydrophobic patterns adds a predictive advantage to the amino acid frequency values. Overall, important patterns and residue composition in-

formation do exist in the primary sequences of proteins that are useful in predicting structural class.

Cross-validation

The cross-validation results are also relatively high — well above 70% for all three methods. Although it is difficult to compare this result to previous studies due to differences in cross-validation methodologies (if cross-validation was used at all), the relatively simple primary sequence information used here is generalizable to the prediction of the structural class of unknown proteins, since all Matthews correlation coefficients obtained are statistically significant.

The change in results when the input space was decreased to the 17 statistically most meaningful predictors is quite interesting. For the Euclidean distance metric, the learning set performance showed a moderate decrease, yet the testing set performance improved slightly (although this improvement is not statistically significant). This decrease in the “generalization gap” — the difference in performance between the learning and test sets — illustrates that a decrease in some of the noisy sequence-specific (or example-specific) information occurs with this simplification but without a corresponding decrease in structural class-specific information. Because the class-specific information is what actually distinguishes the structural classes, this underscores the importance of cross-validation in studies of this type.

One may think of a large generalization gap as implying that the algorithm is “memorizing” the information in a rigid fashion rather than learning the underlying informational concepts behind a classification scheme. The neural networks show the largest generalization gaps, with the performance on the learning sets at or near 100% accuracy while the performance on the test sets came in 20–25% lower. This memorization is likely due to the high number of connections with respect to the size of each learning set (Kneller et al., 1990), which tends to cause the network to readily extract and retain sequence-specific information. The networks used in this study contained more weights than training examples. The fact that the networks performed well despite this limitation suggests that the three classes divide the input space in a rather simple fashion. This is supported by the back-propagation network's performance when no hidden units were used, which was only 4.7% below the best performance for both the 26-input and 17-input categories. This suggests that first-order features, rather than interactions between two or more predictors, describe the vast majority of the division between the structural classes in the input space.

The best solution to the memorization problem appears to be to increase the size of the training set (Qian & Sejnowski, 1988); investigations on larger sets will get underway in the near future. A very recent report indicates development of a representative protein set of over

150 sequences with a minimum of sequence homology (Hobohm et al., 1992).

An increased training set size will also likely decrease some of the “wobble” associated with small sets; that is, larger sets will contain more structural class information, decreasing the variation in prediction accuracy due to the random compositions of the training and test sets. The “wobble” was evaluated by noting the ranges between the average scores for each of the three sets of 8 test files out of the 24 total test files, since each protein occurs once in an 8-file set. These ranges vary from 3.1% (for the Euclidean distance metric, 17 inputs) to 11.0% (back-propagation network, 17 inputs). None of the average scores for a set of 8 test files fell below 70% for any algorithm, showing that the structural class information presented in the training sets appears stable to at least a 70% prediction accuracy. Although the information stability is reasonable considering the relatively small size of the protein sample, a larger set could likely provide noticeable improvement.

Independence of the algorithms

The percent coincidence scores for the algorithms supports the notion that some independence exists between the algorithms, with each algorithm responding to different aspects of the input predictors or their combinations. Thus, a combination of two methods may provide substantially better results than either one alone. A possibility for future research includes combining a statistical method with a neural network method. For example, the prediction resulting from a statistical method could be output to a neural network where it would then become one of the inputs to the network. In this way, the network could be used to “fine tune” the results from the statistical method and improve prediction accuracy.

Effect of sequence homology on results

The all-alpha class is clearly the most internally homologous of the three classes, with the proteins in general belonging to the globin or cytochrome (electron transport) families. The all-beta proteins have an intermediate level of homology, much of which is due to the presence of several members of the immunoglobulin family. The alpha-beta proteins are by far the most heterologous, with less than 1% of the possible sequence pairs having indications of homology. Although general trends in the Matthews correlation coefficients appeared to show better results for more homologous structural classes, the fact that these trends are not statistically significant as well as the fact that all Matthews correlation coefficients for the relatively nonhomologous alpha-beta class are significant to the $P < 0.01$ level underscores the notion that homology considerations alone are not responsible for the predic-

tive value of the primary sequence information presented here.

Incorrectly predicted proteins

Although amino acid composition and certain hydrophobic patterns appear to show a strong correlation to the type of fold, there do exist some clear exceptions, most notably the four proteins listed previously as incorrectly predicted for all test instances. Rubredoxin, an all-beta protein predicted to be in the alpha-beta class, has only 54 residues. Out of that number, only about 26% of its structure is in beta-sheet form (the remaining structure appears to be irregular or random coil). Thus, the low number of residues in beta-sheet form may have made the structural class difficult to ascertain. The other three "problem proteins," consisting of Cu,Zn superoxide dismutase, sea lamprey hemoglobin V, and staphylococcal nuclease appear to fit well within their classes in terms of secondary structure, and no simple explanation exists as to why these proteins systematically failed to be classified correctly. Other markers of structural class besides these predictors may need to be included.

Statistical distribution assumptions

A possible advantage of the neural network methods over the modified Euclidean distance metric is the lack of a normality assumption for neural network input data. As noted in Table 3, the standard deviation of the $(i, i + 2, i + 4)$ pattern for the all-alpha class is greater than the mean, implying nonnormality in some of the frequency distributions despite our attempt to "normalize" the data into number of occurrences per 100 residues.

Conclusions

Overall, given that studies in the use of local information to predict local conformation (also usually using three categories of structure—helix, sheet, or coil) can at best obtain a prediction accuracy of 60–65%, our results are indeed encouraging. It appears that the mapping of non-local information (such as frequencies of primary sequence elements) to nonlocal structure designation (i.e., structural class) likely generates more accurate information about protein structure than the mapping of local information to local structure, given the results of this study. Although it is arguable that the present study is difficult to compare to the previous local conformation prediction studies due to their differing natures, the evidence is strong that nonlocal information such as structural class and frequencies of primary sequence elements is an important part of the overall protein structure prediction problem. Because it has been shown that the addition of structural class information can improve the results of local secondary structure prediction (Deleage & Roux, 1987;

Kneller et al., 1990), further studies on the relationship of structural class information to secondary structure prediction are warranted.

Materials and methods

Protein sets

The protein sequences used in this study were the set used by Chou (1989) in order to permit comparison to those results. This set consists of 64 proteins that can be partitioned into 19 all-alpha proteins, 15 all-beta, and 30 alpha-beta mixed proteins (Table 1). All proteins in the all-alpha class contain at least 35% of their amino acids in a helical conformation and less than 7% of their amino acids designated as beta-strand. The all-beta class contains proteins with at least 25% of their amino acids in a beta-strand conformation and less than 11% in a helical-type structure. The alpha-beta class contains proteins that do not meet the criteria for the other two classes and contain at least 10% helix and 10% beta-strand. The protein sequences were obtained from the Protein Data Bank (Bernstein et al., 1977) and the Protein Information Resource (Barker et al., 1990). The alpha + beta and alpha/beta classes were combined into a single alpha-beta mixed class. All protein sequences were converted from three-letter amino acid codes into the single-letter code.

Structural class predictors

Twenty-six different predictors are used in this study. The first 20 are the frequencies of occurrence of the 20 amino acid types. The next six are hydrophobic patterns $(i, i + 2; i, i + 3; i, i + 2, i + 4; i, i + 5; i, i + 3, i + 4; i, i + 1, i + 4)$ based on the theory of Lim (1974). All frequencies are normalized to the number of occurrences per 100 residues. The selection of hydrophobic amino acids is based on Rose et al. (1985) and includes amino acids that have a mean fractional area loss (on transfer from the standard state to the folded protein) of greater than 80%. These amino acids tend to be buried in the hydrophobic core of a protein. The seven amino acids that meet this criterion are Val (V), Leu (L), Ile (I), Met (M), Cys (C), Phe (F), and Trp (W). The remaining 12 amino acids (excluding proline) are included in the "nonhydrophobic" group for the purposes of the above patterns. Proline (P), which tends to kink, bend, or break elements of secondary structure, is not included in either group.

The program, DOMAIN

All predictor frequency calculations, sample randomization, Euclidean clustering algorithm computations, and preparation of input for the neural network was performed using DOMAIN, a Lisp-based software toolkit for the analysis of protein sequences. A brief description

of the program can be found in Saurugger and Metfessel (1991). Neural network performance analysis was carried out by a separate program written in C.

Randomization and cross-validation

The 64 proteins were randomly divided in an iterative fashion into 24 pairs of training and test sets, using a methodology similar to that used by Zhang et al. (1992). Pairs of training-test sets were created using the Allegro Common Lisp pseudorandom number generator, with eight proteins selected for each test set. Selection of test set proteins was sampled without replacement. Each test set was matched with a training set that consisted of the 56 proteins not in the test set. This procedure was iterated eight times so that each protein in the 64-protein set occurred in a test set exactly once. The proteins were then "shuffled," and the selection process was repeated. Iteration of the process resulted in 24 unique training set-test set pairs, with each protein appearing in 3 test and 21 training sets.

This "multiple-pass k -way" cross-validation methodology ensures that each training set is large enough for adequate representation of all structural classes, yet not so large compared to the test set that homology within a structural class becomes overly important, which might be the case using the "leave-one-out" method of testing in which each test set contains only one protein. Having each protein represented in three test sets decreases the influence of information that may result from the specific combination of proteins in the training and test sets rather than from differences between the structural classes.

Modified Euclidean statistical clustering algorithm

This modification of a Euclidean distance metric can be used to classify a protein into a structural class (Widman & Pierce, 1990). Briefly, for each of the 26 predictors, the mean and standard deviation of the normalized frequency of occurrence is calculated for each structural class. For each individual protein, frequencies are calculated for all 26 predictors. These values are compared to the respective means and standard deviations for the proposed structural class and a distance value is calculated (Equation 1). The Euclidean distance between a protein x and a particular structural class i ($i = \alpha, \beta, \alpha\beta$) is given by the following equation:

$$D_i(x) = \{\sum_j [(v_j(x) - \mu_j(i))^2 / \sigma_j(i)]\}^{0.5} \quad (1)$$

with v_j referring to the individual protein's frequency value (per 100 residues) for predictor j . The means $\mu_j(i)$ and standard deviations $\sigma_j(i)$ of the frequency for each of the 26 predictors were calculated from the respective training set. When the matching test set was run (contain-

ing the remaining eight proteins not in the training set), the distances of the test proteins from the structural class spaces (frequency means and standard deviations) created from the training set were calculated. Each test protein is predicted as being in the class with the lowest distance (Fig. 1).

Neural network-based structural class prediction

Back-propagation networks

A back-propagation neural network (Rumelhart et al., 1986) was created using the NeuralWorks Professional II/Plus software package, version 4.05 (NeuralWare, Inc., Pittsburgh, Pennsylvania). The output layer consisted of three neuronal units, each one representing a single structural class. The network was built with a maximum of 26 input units, with each input unit representing the normalized frequency of 1 of the 20 amino acids or 1 of the 6 hydrophobic patterns and a variable number of hidden units (Fig. 2A). The network is a fully connected network such that each unit is connected to every unit in the next layer in a feed-forward fashion.

Prior to the start of a run, the weights of the connections in our network were randomly set to small values (between 0.2 and -0.2). When the network is in the learning phase, the product of an input value V_i from a unit

Fig. 1. Simplified example of the modified Euclidean statistical clustering algorithm, using only two patterns (predictors) and two structural classes. Initially, the "spaces" are created for each structural class based on the means and standard deviations for the frequencies of patterns A and B for the all-alpha and all-beta proteins in the learning set. The centers of each space are where the means for patterns A and B (for the relevant structural class) intersect. The boundaries of each space are an approximate measure of spread in frequency values. The pattern frequency values for an unknown test protein X lie at the tip of the arrow. Because its frequency values place that sequence closer to the all-beta space, protein X is predicted to be an all-beta protein (adapted from Metfessel & Saurugger [1993]).

Fig. 2. **A:** Basic architecture for a 26-input, three-layer back-propagation network used in the study. The squares represent the neuronal units, and the dark lines represent the weighted connections. Each of the input layer units contains the value of a protein's frequency (per 100 residues) of an amino acid type or a hydrophobic residue pattern. Each output unit denotes a single structural class. The bias unit, which has a value of 1 and connects to the hidden and output layers, is not shown. **B:** Basic learning vector quantization network architecture. The main difference from the back-propagation network is in the hidden (Kohonen) layer, where each hidden unit is connected to only a single output unit. The hidden units compete with each other in a winner-take-all fashion.

i in the previous layer multiplied by its respective connection weight W_{ij} is obtained for each neuron in the previous layer that is connected to neuron j , and the products are summed. A bias term b_i , which has a constant activation of 1, is connected to the hidden and output layers of the network used here. This relationship is given in Equation 2:

$$A_j = \sum_i W_{ij} V_i + b_i. \quad (2)$$

Once the activation level of a neuronal unit is calculated, a transfer function is used to convert it into an output value, which can then be transmitted to units connected to it in the next layer. We achieved efficient learning with the hyperbolic tangent (tanh) function (Equation 3).

$$V_j = f(A_j) = (e^{A_j} - e^{-A_j}) / (e^{A_j} + e^{-A_j}). \quad (3)$$

The value V_j that results from this nonlinear function can vary from between -1.0 to 1.0 and shows the steepest slope at $A_j = 0$.

Once the output layer becomes active, its result is compared to the desired result provided by the experimenter, and an error term is calculated based on the difference between the actual outputs of the network (a_o) and desired

outputs (d_o). The error term (E_o) for a particular neuron o in the output layer is calculated according to Equation 4:

$$E_o = f'(A_o)(d_o - a_o), \quad (4)$$

where $f'(A_o)$ is the derivative of the transfer function.

Based on the error term and the learning rule used (often called the "delta rule"), the connection weights are updated for each connection in order to decrease the error in subsequent runs. If there exists a hidden layer that is connected to the output layer, the error term for a particular neuron h in the hidden layer is calculated based on the back-propagated error terms of the output layer neurons to which it is connected.

$$E_h = f'(A_h) \sum_o (E_o * W_{ho}), \quad (5)$$

with W_{ho} representing the connection weight from neuron h (in the hidden layer) to the output neuron o . The error for the neurons in the layer previous to the hidden layer (i.e., another hidden layer or input layer) is calculated similarly, but the error from the hidden layer neurons is back-propagated instead of the error from the output layer neurons. The decrease in error as learning progresses is known as gradient descent. For a connection between any neuron i in layer $L - 1$ to neuron j in layer L , the descent is implemented by the delta rule equation (Equation 6):

$$\Delta W_{ij} = (R)(E_j)(V_i) + M(\Delta W_{ij}(\text{prev})), \quad (6)$$

where R refers to the learning coefficient. The weight of the connection is changed by the delta-weight (ΔW_{ij}). The momentum term (M) is multiplied by the previous delta-weight on the same connection. The momentum term helps to filter out the more random, oscillatory signals and allows a faster learning rate. The network was trained for 200 epochs (11,200 examples) per run. For the hidden layer in our networks, $R = 0.3$ and $M = 0.4$ for the first 10,000 examples. For the last 1,200 examples, $R = 0.15$ and $M = 0.2$. The output layer had $R = 0.15$ for the first 10,000 examples and $R = 0.075$ for the remaining 1,200 examples. The momentum term was the same as that for the hidden layer.

Learning vector quantization networks

The second form of neural network used in the study is the LVQ network, first mentioned by Kohonen et al. (1988) and implemented in this study using the Neural-Works Professional II/Plus package (version 4.05) using the default "lvq" learning and recall schedule. The hidden (Kohonen) layer is fully connected to the input layer but not to the output layer. Instead, each output neuron has "dedicated" hidden layer neurons that connect only

to that output neuron and become feature detectors as learning proceeds (Fig. 2B). Each unit in the hidden layer competes with the other hidden units until one unit clearly shows the highest activation. The other units shut down, and the remaining (winning) hidden unit activates the output unit to which it is connected.

To accomplish learning, the Euclidean distance between the data for a learning example and the weights for the connections coming into a hidden unit are calculated from Equation 7:

$$D = [\sum_i (W_{ij} - V_i)^2]^{0.5}, \quad (7)$$

where W_{ij} is the weighted connection between an input and hidden unit, and V_i is the value of the input datum at input unit i . The squared differences between W_{ij} and V_i were summed over all 26 (or fewer) inputs in the fully connected network used.

If the hidden layer unit with the lowest Euclidean distance is connected to the output neuron in the correct (desired) class, the weight vector (containing the weights of all the connections from the input layer to that hidden layer neuron) is increased by an amount that depends on a learning coefficient (R_1) and the difference between the input vector (V) and the old weight vector (W_{old}):

$$W_{new} = W_{old} + R_1(V - W_{old}). \quad (8)$$

If the winning hidden unit is connected to the output unit representing the incorrect class, its weight vector is adjusted (Equation 9) such that:

$$W_{new} = W_{old} - R_2(V - W_{old}). \quad (9)$$

This implementation will cause the Euclidean distance to increase the next time the example is presented, discouraging the selection of that incorrect class.

Several enhancements to the basic LVQ algorithm are possible. The LVQ1 algorithm (DeSieno, 1988) has a "conscience" mechanism in which a hidden layer unit that wins too often is penalized by the addition of a distance bias. This bias helps ensure that the feature detection is spread out among most or all of the hidden units. LVQ2 is another enhancement that may be used during a later stage of training. In examples where a hidden unit connected to the correct output unit does not win but comes in a close second, the correct second place hidden neuron is moved "closer" in Euclidean distance to the training data vector, while the winning (incorrect) neuron is moved farther away. In the implementation here, LVQ1 learning took place during the first 3,000 examples, after which the learning mode was switched to LVQ2. A total of 5,600 examples were provided to the network per training session. More examples did not improve learning.

Statistical analysis and results evaluation

The "prediction accuracy" may be defined simply as the total percentage of correct predictions, given by Equation 10:

$$P = (P_\alpha + P_\beta + P_{\alpha\beta})/N, \quad (10)$$

where P_i refers to the percentage of correctly predicted proteins in each structural class i , with N referring to total sample size.

One difficulty with this measure is that it does not take into account disparities in the number of sequences per class. The Matthews correlation coefficient (Matthews, 1975) remedies this problem. For example, if all proteins were predicted as alpha-beta, the prediction accuracy would be about 50% but the Matthews coefficient would be zero. The coefficient (C_i) for structural class i is depicted in Equation 11:

$$C_i = \frac{p_i n_i - o_i u_i}{[(n_i + u_i)(n_i + o_i)(p_i + u_i)(p_i + o_i)]^{0.5}}. \quad (11)$$

If the Matthews correlation coefficient was being calculated for the all-alpha class, for example, then p_i would be the number of all-alpha proteins correctly predicted (true positives), n_i would stand for the number of proteins not belonging to the all-alpha class that were correctly rejected from that class (true negatives), whether or not the prediction itself was correct (e.g., if an all-beta protein were incorrectly predicted as an alpha-beta protein, it would still count as a true negative for the all-alpha class), o_i would represent the number of non-all-alpha proteins incorrectly predicted as all-alpha (false positives or "overpredicted" cases), and u_i would represent all-alpha proteins that were incorrectly predicted into other classes (false negatives or "underpredicted" cases).

For comparing the performance of one algorithm to that of another, McNemar's test (McNemar, 1947; Armitage, 1971) is used, which requires computation of two values: (1) the number of proteins (P_1) that algorithm 1 predicts correctly but algorithm 2 predicts incorrectly, and (2) the number of proteins (P_2) that algorithm 2 predicts correctly but algorithm 1 predicts incorrectly. The calculation of a standardized normal deviate (u) is performed according to Equation 12:

$$u = (P_1 - P_2 - 1)/(P_1 + P_2)^{0.5}. \quad (12)$$

To compute P_1 and P_2 , the prediction obtained from each algorithm is compared for every protein in all 24 test sets, making a total of $24 \times 8 = 192$ comparisons. Because each protein is found a total of three times in the 24 test sets, the final values of P_1 and P_2 used in the above equation consist of the raw values (for all 192 cases) divided by 3.

For comparing a prediction algorithm's performance for different classes of proteins, the two-sample test of proportions (unpaired cases) is used (e.g., Armitage, 1971).

Sequence homology determinations

Pairwise sequence alignments were performed to assist in determining sequence homology within or between classes. The program FASTDB (Brutlag et al., 1990) from the IntelliGenetics Suite (IntelliGenetics, Inc., Mountain View, California) was used to perform the alignments and determine sequence similarity. A PAM-256 (point accepted mutation—256) table was used to determine scores and criteria for conservative substitutions (Dayhoff et al., 1978).

Acknowledgments

We thank Dr. Dennis Hawkins for his valuable advice in the area of statistical analysis and Dr. Lynda Ellis for her suggestions in the preparation of this manuscript. This work was in part supported by grant LM007041 (to B.A.M.) from the National Library of Medicine.

References

- Armitage, P. (1971). Statistical inference. In *Statistical Methods in Medical Research*, pp. 99–146. John Wiley & Sons, New York.
- Barker, W.C., George, D.G., & Hunt, L.T. (1990). Protein sequence database. *Methods Enzymol.* 183, 31–49.
- Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., & Tasumi, M. (1977). The Protein Data Bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol.* 112, 535–542.
- Brutlag, D.L., Dautricourt, J.-P., Maulik, S., & Relph, J. (1990). Sensitive similarity searches of biological sequence databases. *Comput. Appl. Biosci.* 3, 237–245.
- Chou, P.Y. (1989). Prediction of protein structural classes from amino acid compositions. In *Prediction of Protein Structure and the Principles of Protein Conformation* (Fasman, G.D., Ed.), pp. 549–586. Plenum Press, New York.
- Dayhoff, M.O., Schwartz, R.M., & Orcutt, B.C. (1978). A model for evolutionary change in proteins. *Atlas Protein Sequence Struct.* 5(Suppl. 3), 345–352.
- Deleage, G. & Roux, B. (1987). An algorithm for protein secondary structure prediction based on class prediction. *Protein Eng.* 1, 289–294.
- DeSieno, D. (1988). Adding a conscience to competitive learning. *Proc. Second Annu. IEEE Intl. Conference on Neural Networks* 1, 117–124.
- Dill, K.A. (1990). Dominant forces in protein folding. *Biochemistry* 29, 7133–7155.
- Hobohm, U., Scharf, M., Schneider, R., & Sander, C. (1992). Selection of representative protein data sets. *Protein Sci.* 1, 409–417.
- Kinoshita, J. (1990). Net result: Folded protein. A neural network decipher the structure of protein. *Sci. Am.* 262, 24–26.
- Klein, P. (1986). Prediction of protein structural class by discriminant analysis. *Biochim. Biophys. Acta* 874, 205–215.
- Klein, P. & Delisi, C. (1986). Prediction of protein structural class from the amino acid sequence. *Biopolymers* 25, 1659–1672.
- Kneller, D.G., Cohen, F.E., & Langridge, R. (1990). Improvements in protein secondary structure prediction by an enhanced neural network. *J. Mol. Biol.* 214, 171–182.
- Kohonen, T., Barna, G., & Chrisley, R. (1988). Statistical pattern recognition with neural networks: Benchmarking studies. *Proc. Second Annu. IEEE Intl. Conference on Neural Networks* 1, 61–68.
- Levitt, M. & Chothia, C. (1976). Structural patterns in globular proteins. *Nature* 261, 552–557.
- Lim, V.I. (1974). Algorithms for the prediction of α -helical and β -structural regions in globular proteins. *J. Mol. Biol.* 88, 873–894.
- Matthews, B.W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* 405, 442–451.
- McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 12, 153–157.
- Metfessel, B. & Saurugger, P.N. (1993). Pattern-recognition in the prediction of protein structural class. *Proc. 26th Annu. Hawaii Intl. Conference on System Sciences*, pp. 679–688.
- Nakashima, H., Nishikawa, K., & Ooi, T. (1986). The folding type of a protein is relevant to its amino acid composition. *J. Biochem. (Tokyo)* 99, 153–162.
- Qian, N. & Sejnowski, J. (1988). Predicting the secondary structure of globular proteins using neural network models. *J. Mol. Biol.* 202, 865–884.
- Rose, G.D., Geselowitz, A.R., Lesser, G.J., Lee, R.H., & Zehfus, M.H. (1985). Hydrophobicity of amino acid residues in globular proteins. *Science* 229, 834–838.
- Rumelhart, D.E., Hinton, G.E., & Williams, R.J. (1986). Learning internal representations by error propagation. In *Parallel Distributed Processing*, Vol. 1 (Feldman, J.A., Hayes, P.J., & Rumelhart, D.E., Eds.), pp. 318–362. MIT Press, Cambridge, Massachusetts.
- Saurugger, P.N. & Metfessel, B.A. (1991). Patterns in protein primary sequences: Classification, display and analysis. *Proc. 15th Annu. Symposium on Computer Applications in Medical Care*, pp. 299–303.
- Sander, C. & Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins Struct. Funct. Genet.* 9, 56–68.
- Schiffer, M. & Edmundson, A. (1967). Use of helical wheels to represent the structures of proteins and to identify segments with helical potential. *Biophys. J.* 7, 121–135.
- Sheridan, R.P., Dixon, J.S., Venkataghavan, R., Kuntz, I.D., & Scott, K.P. (1985). Amino acid composition and hydrophobicity patterns of protein domains correlate with their structures. *Biopolymers* 24, 1995–2023.
- Widman, L.E. & Pierce, B.S., III (1990). Comparison of a knowledge-based pattern classification algorithm with the Mahalanobis statistical clustering algorithm. *Proc. 14th Annu. Symposium on Computer Applications in Medical Care*, pp. 529–533.
- Zhang, C.-T. & Chou, K.-C. (1992). An optimization approach to predicting protein structural class from amino acid composition. *Protein Sci.* 1, 401–408.
- Zhang, X., Mesirov, J.P., & Waltz, D.L. (1992). Hybrid system for protein secondary structure prediction. *J. Mol. Biol.* 225, 1049–1063.