Automatic Derivation of Substructures Yields Novel Structural Building Blocks in Globular Proteins

Xiru Zhang*, Jacquelyn S. Fetrow†, William A. Rennie‡, David L. Waltz*
and George Berg‡

*Thinking Machines Corp.
245 First Street
Cambridge, MA

†Dept. of Biological Sciences
‡Computer Science Dept.
University at Albany – SUNY
Albany, NY 12222

Abstract

Because the general problem of predicting the tertiary structure of a globular protein from its sequence is so difficult, researchers have tried to predict regular substructures, known as *secondary structures*, of proteins. Knowledge of the position of these structures in the sequence can significantly constrain the possible conformations of the protein. Traditional protein secondary structures are α-helices, β-sheets, and coil. Secondary structure prediction programs have been developed, based upon several different algorithms. Such systems, despite their varied natures, are noted for their universal limit on prediction accuracy of about 65%. A possible cause for this limit is that traditional secondary structure classes are only a coarse characterization of local structure in proteins. This work presents the results of an alternative approach where local structure classes in proteins are derived using neural network and clustering techniques. These give a set of local structure categories, which we call Structural Building Blocks (SBBs), based upon the data itself, rather than *a priori* categories imposed upon the data. Analysis of SBBs shows that these categories are general classifications, and that they account for recognized helical and strand regions, as well as novel categories such as N- and C-caps of helices and strands.

Introduction

Traditionally, protein structure has been classified into continuous segments of amino acids called *secondary structures*. The existence of the regular secondary structures, α-helices and β-sheets, was hypothesized even before the first protein structure had been solved at atomic resolution [Pauling and Corey, 1951; Pauling *et al.*, 1951]. These structures have regular patterns of hydrogen bonding and repeating backbone dihedral angles and are easy to locate in protein crystal structures. Following the solution of a few protein structures, Venkatachalam suggested the existence of a third class of structure, the β-turn [Venkatachalam, 1968]. Often, the remainder of protein structure is called "coil" or "other"; however, attempts have been made to identify other structures such as Ω-loops [Leszczynski and Rose, 1986] or Ω, straps, and ζ-loops [Ring *et al.*, 1992] in these regions.

Because the classical secondary structures were predicted before any protein structures were solved and because these regular structures are so easy to identify by eye in visualized protein structures, these categories have traditionally been used in protein structure prediction routines. From the earliest prediction algorithms [Chou and Fasman, 1974], through artificial neural network models [Qian and Sejnowski, 1988], to current hybrid systems using multiple prediction algorithms [Zhang *et al.*, 1992], these systems consistently used the traditional secondary structures, usually the categories provided by the DSSP program [Kabsch and Sander, 1983]. Despite the variety of algorithms used, the best prediction rates for these programs consistently classify only about 65% of the residues' secondary structures correctly. This rate of accuracy is too low to be of practical use in constraining the conformation for tertiary structure prediction. Re-categorization of protein structure may be one way of increasing prediction accuracy

One indication that these classical secondary structures may not be suitable is that attempts to define secondary structures in proteins of known structure produce inconsistent results. Such programs may use the criteria of hydrogen bonding [Presta and Rose, 1988], alpha carbon dihedral angles [Richards and Kundrot, 1988], backbone dihedral angles or some combination of these criteria [Kabsch and Sander, 1983; Richardson and Richardson, 1988]. When comparing output from these programs which use proteins of known structure, there is a great deal of disagreement in their secondary structure assignments (Fetrow and Berg, unpublished observations). It thus seems reasonable to hypothesize that the classical categories of secondary structures are too coarse and attempts to predict such artificial categories will ultimately fail [Zhang *et al.*, 1992].

Zhang, X., J. Fetrow, W.A. Rennie, D.L. Waltz, & G. Berg. "Automatic Derivation of Substructures Yields Novel Structural Building Blocks in Globular Proteins," in L. Hunter, D. Searls, & J. Shavlik (eds.) Proceedings of the First International Conference on Intelligent Systems for Molecular Biology, 1993, 438-446.

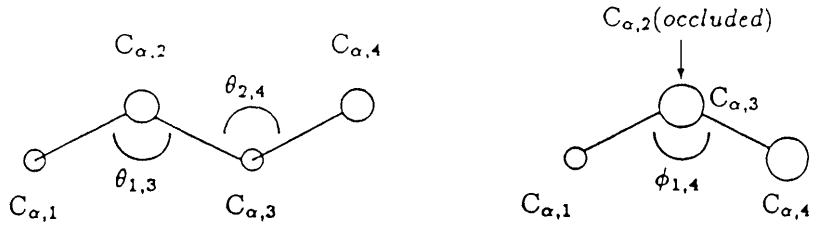

Figure 1: The bond and dihedral angles used for residue-feature-vector representations. For simplicity, a window size of four is displayed. A bond angle $\theta_{i-1,i+1}$ is the virtual angle formed by the three C_α atoms centered at residue i. The dihedral angle $\phi_{i,i+3}$ is defined as the angle between the virtual bond between $C_{\alpha,i}$ and $C_{\alpha,i+1}$ and the virtual bond between $C_{\alpha,i+2}$ and $C_{\alpha,i+3}$ in the plane perpendicular to the virtual bond formed between $C_{\alpha,i+1}$ and $C_{\alpha,i+2}$.

The purpose of this research, therefore, is to attempt an objective re-classification of protein secondary structure. Here we present the results of a categorization system combining artificial neural network and clustering techniques. The first part of the system is an auto-associative artificial neural network, called GENEREP (GENErator of REPresentations), which can generate structural representations for a protein given its three-dimensional residue coordinates. Clustering techniques are then used on these representations to produce a set of six categories which represent local structure in proteins. These categories, called *Structural Building Blocks* (SBBs), are general, as indicated by the fact that the categories produced using two disjoint sets of proteins are highly correlated. SBBs can account for helices and strands, acknowledged local structures such as N- and C-caps for helices, as well as novel structures such as N- and C-caps for strands.

Methods and Materials

The initial goal of this work was to find a low-level representation of local protein structure that could be used as the basis for finding general categories of local structure. These low-level representations of local regions were used as input to an auto-associative neural network. The hidden layer activations produced by this network for each local region were then fed to a clustering algorithm, which grouped the activation patterns into a specified number of categories, which was allowed to vary from three to ten. Patterns and category groupings were generated by networks trained on two disjoint sets of proteins. The correlations between the categories generated by the two networks were compared to test the generality of the categories and the relative quality of the categories found using different cluster sizes.

The structural categories found along protein sequences were then analyzed using pattern recognition software in order to find frequently occurring groupings of categories. Molecular modeling software was also used to characterize and visualize both the categories themselves and the groupings found by the pattern recognizer.

In contrast to earlier work on GENEREP [Zhang and Waltz, 1993], in which a measure of residue solvent-accessibility was used, a purely structural description of the protein was employed in this study, as well as a more general input/output encoding scheme for the neural network. Each protein was analyzed as a series of seven-residue "windows". The residues were represented by the seven α-carbon (C_α) atoms of the adjacent residues. The structure of the atoms in the window was represented by several geometric properties. For all except adjacent C_α atoms, the distances between each pair of C_α atoms in the window were measured. The distance between adjacent atoms was not utilized because it is relatively invariant. There were fifteen such distances per window. The four dihedral and five bond angles which specify the geometry of the seven C_α atoms in each window were used as well (Figure 1).

Because these measurements were used as input to an artificial neural network, they had to be represented in a form that was consistent with the values of the network's units, while also preserving information implicit in the measurements. The following encoding was used. Each dihedral angle was represented using two units, one each for the sine and cosine of the angle. These were normalized to the value range $[0, 1]$ of the input units. This representation preserved the continuity and similarity of similar angles, even across thresholds such as 360° to 0°. The distances were represented using two units. Analysis of the distances showed a rough bi-modal distribution of distance values. The units were arranged so that the activation level of the first unit represented a distance from the minimum distance value found to a point mid-way between the two "humps" of the distribution. If the distance was greater than the value of the mid-way point, the first unit was fully activated, and the second unit activated in proportion to how much the distance was between the mid-way point and the maximum distance value. The bond angles were each represented using

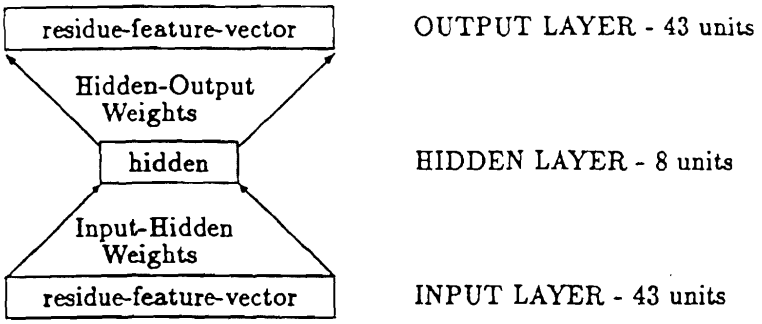

Figure 2: The auto-associative neural network used in this study to find the residue-state-vectors. This network was trained using the residue-feature-vectors described in Methods as both the input and output patterns. The learning method used was error backpropagation [Rumelhart *et al.*, 1986].

one unit, with the values representing the angles in the range $[0°, 180°]$ normalized to $[0, 1]$. The representations of these C_α distances, dihedral and bond angles in a window constituted the *residue-feature-vector* for a window.

The residue-feature-vectors were calculated for every window for each of the proteins in Table 1. The protein list, consisting of 74 globular proteins of known structure, with 75 distinct sequence chains and a total of 13,114 residues, was chosen such that all protein structures had a resolution of 2.5Å or better and a refinement R-factor of 0.3 or less. These limits excluded proteins which were not resolved well enough to determine their backbone structure satisfactorily. Using a standard sequence alignment algorithm [Smith and Waterman, 1981], the list was also tested to ensure that the amount of sequence similarity between proteins was below 50%. This list of proteins was then divided into two disjoint sets, Data Set One and Data Set Two (Table 1). Subsequent work was done using the proteins in one data set and verified against the other. Data Set One consisted of 38 sequences containing a total of 6650 residues. As defined by DSSP [Kabsch and Sander, 1983], 30.1% of the residues in this set were defined to be in α-helices and 18.9% in β-strands. Data Set Two consisted of 37 sequences with a total 6464 residues. For this set, 30.8% of the residues were in α-helices and 18.2% were in β-strands.

For each protein, a residue-feature-vector was calculated at each position along the sequence for which there was an amino acid present in all slots of the window. Since they do not have residues in all of the window locations, the first three positions at both the N- and C-termini did not have residue-feature-vectors associated with them. Thus, a protein sequence with n residues will provide $n - 6$ residue-feature-vectors. Data Set One provided 6422 residue-feature-vectors and Data Set Two provided 6242 residue-feature-vectors.

The residue-feature-vectors for a given data set were used as both input and output patterns for an auto-associative backpropagation neural network [Rumelhart *et al.*, 1986]. Using the representation for the residue-feature-vectors described above, both the input and output layers of the network contained 43 units. The hidden layer contained eight units (Figure 2). The hidden layer size was determined empirically as the smallest size which produced the network most likely to succeed at the auto-association task (where the root mean squared error of the network eventually went below 0.01).

The goal of an auto-associative network is to learn to reproduce the pattern of each input to the network at the output layer of the network. Each residue-feature-vector pattern is presented individually to the network as a set of activations to the input layer. By multiplying each input unit's activation by the value of the weight connecting it to the hidden units, summing them at the hidden units, and then scaling them into $[0, 1]$ with an exponentiation function, the hidden units' activation values are calculated. This same process is then used to calculate the output units' activations from the hidden units' activation values. The output units' activation values are then compared to those of the corresponding input units. These differences (the *errors*) are used to change the value of the weights between the layers, using error-backpropagation [Rumelhart *et al.*, 1986], a gradient descent technique. This process is repeated for each pattern in the data set, which constitutes an *epoch* of training.

In this study, the auto-associative networks were trained for some number of epochs (approximately 1500) on a Connection Machine CM5 until the RMS error at the output layer was at most 0.01. At this point, the networks were run one additional epoch on the residue-feature-vector patterns, without changing the weights. For each pattern, the values of the hid-

Name	Chains	Residues	Set	Resolution	Refinement	Description
155C		134	1	2.5Å	–	P. Denitrificans Cytochrome C550
1ACX		107	1	2.0Å	–	Actinoxanthin
1BP2		123	1	1.7Å	0.171	Bovine phospholipase A2
1CCR		111	1	1.5Å	0.19	Rice Cytochrome C
1CRN		46	1	1.5Å	–	Crambin
1CTF		68	1	1.7Å	0.174	Ribosomal Protein (C terminal fragment)
1ECD		136	1	1.4Å	–	Deoxy hemoglobin (erythrocruorin)
1FX1		147	1	2.0Å	–	Flavodoxin (D. Vulgaris)
1HIP		85	1	2.0Å	0.24	Oxidized High Potential Iron Protein
1HMQ	A	113	1	2.0Å	0.173	Hemerythrin
1LH1		153	1	2.0Å	–	Leghemoglobin (Acetate, Met)
1MLT	A	26	1	2.5Å	–	Melittin
1NXB		62	1	1.38Å	0.24	Neurotoxin
1PAZ		120	1	1.55Å	0.18	A. faecalis Pseudoazurin
1PCY		99	1	1.6Å	0.17	Plastoryanin
1RNT		104	1	1.9Å	0.18	Ribonuclease T1 complex
1UBQ		76	1	1.8Å	0.176	Human Ubiquitin
2ACT		218	1	1.7Å	0.171	Actinidin
2APP		323	1	1.8Å	0.136	Acid proteinase
2AZA	B	128	1	1.8Å	0.188	Azurin (oxidized)
2CAB		256	1	2.0Å	0.193	Carbonic anhydrase
2CNA		237	1	2.0Å	–	Jack Bean Concanavalin
2CPP		405	1	1.63Å	0.19	Cytochrome P450
2CYP		287	1	1.7Å	0.202	Yeast Cytochrome C peroxidase
2HHB	A	141	1	1.74Å	0.16	Human deoxyhemoglobin
2LZM		164	1	1.7Å	0.193	T4 Lysozyme
2PRK		279	1	1.5Å	0.167	Fungus Proteinase K
2SOD	B	151	1	2.0Å	0.256	Cu Zn Superoxide dismutase (bovine)
3ADK		194	1	2.1Å	0.193	Porcine adenylate kinase
3ICB		75	1	2.3Å	0.178	Bovine Calcium-binding protein
3PGK		415	1	2.5Å	–	Yeast Phosphoglycerate kinase
3RXN		52	1	1.5Å	–	Rubredoxin
4ADH		374	1	2.4Å	0.26	Equine Apo-liver alcohol dehydrogenase
4DFR	B	159	1	1.7Å	0.155	Dihydrofolate reductase complex
4PTI		58	1	1.5Å	0.162	Trypsin inhibitor
5CPA		307	1	1.54Å	–	Bovine carboxypeptidase
7CAT	A	498	1	2.5Å	0.212	Beef catalase
9PAP		212	1	1.65Å	0.161	Papain CYS-25 (oxidized)
1ABP		306	2	2.4Å	–	L-arabinose binding protein E.Coli
1CPV		108	2	1.85Å	0.4	Ca-binding Parvalbumin
1FB4	H,L	445	2	1.9Å	0.189	Human Immunoglobulin FAB
1FDX		54	2	2.0Å	–	Ferredoxin
1GCR		174	2	1.6Å	0.23	Calf γ-crystallin
1LZ1		130	2	1.5Å	0.177	Human Lysozyme
1MBD		153	2	1.4Å	–	Deoxymyoglobin (Sperm Whale)
1PHH		394	2	2.3Å	0.193	hydroxybenzoate hydroxylase
1PPT		36	2	1.37Å	–	Avian Pancreatic Polypeptide
1RHD		293	2	2.5Å	–	Bovine rhodanese
1RN3		124	2	1.45Å	0.26	Bovine Ribonuclease A
1SBT		275	2	2.5Å	–	Subtilisin
1SN3		65	2	1.8Å	1.3	Scorpion Neurotoxin

Table 1: The protein structures used in this work.

Name	Chains	Residues	Set	Resolution	Refinement	Description
2ABX	A	74	2	2.5Å	0.24	α bungarotoxin
2APR		325	2	1.8Å	0.143	Acid Proteinase (R. chinensis)
2B5C		85	2	2.0Å	-	Bovine Cytochrome B5 (oxidized)
2CCY	A	127	2	1.67Å	0.188	R. Milischianum Cytochrome C'
2CDV		107	2	1.8Å	0.176	Cytochrome C3 (D. Vulgaris)
2CGA	A	245	2	1.8Å	0.173	Bovine Chymotrypsinogen
2CI2	I	65	2	2.0Å	0.198	Chymotrypsin inhibitor
2CTS		437	2	2.0Å	0.161	Pig citrate synthase
2GN5		87	2	2.3Å	0.217	Viral DNA Binding Protein
2HHB	B	141	2	1.74Å	0.16	Human deoxyhemoglobin
2LHB		149	2	2.0A	0.142	Hemoglobin V (Cyanomet, lamprey)
2OVO		56	2	1.5Å	0.199	Ovomucoid third domain (protease inh.)
2PAB	A	114	2	1.8Å	0.29	Human prealbumin
2SNS		141	2	1.5Å	-	S. Nuclease complex
351C		82	2	1.6Å	0.195	Cytochrome C 551 (oxidized)
3C2C		112	2	1.68Å	0.175	R. Rubrum Cytochrome C
3GAP	A	208	2	2.5Å	0.25	E. Coli catabolite gene activator protein
3GRS		461	2	2.0Å	0.161	Human glutathione reductase
3WGA	B	171	2	1.8Å	0.179	Wheat Germ Agglutinin
3WRP		101	2	1.8Å	0.204	TRP aporepressor
4FXN		138	2	1.8Å	0.2	Flavodoxin (Semiquinone form)
4TLN		316	2	2.3Å	0.169	Thermolysin (B. thermoproteolyticus)
4TNC		160	2	2.0Å	0.172	Chicken Troponin C

Table 1: The Protein Structures used in this work. The columns contain the following information: Name: The name of the protein as assigned by the Brookhaven database. Chains: If the protein contains multiple chains, the chain used is indicated. Residues: The number of residues in the sequence, as indicated by DSSP. Set: 1 corresponds to Data Set One and 2 to Data Set Two in this study. Resolution: The resolution of the structure, as given in the Brookhaven entry. Refinement: when available, the refinement as given in the Brookhaven entry. Description: A short description of the protein, based upon the information in the Brookhaven entry.

den layer units were recorded. This pattern of activation was the *residue-state-vector* associated with each residue-feature-vector pattern.

One auto-associative network was trained on the protein sequences in Data Set One and one on the protein sequences in Data Set Two. After training, the residue-state-vectors for Data Set Two were calculated by both the network trained on Data Set One and the network trained on Data Set Two. The residue-state-vectors produced by each of the two networks were then separately grouped using a k-means clustering algorithm [Hartigan and Wong, 1975]. Cluster sizes of three through ten were tested. Each residue-feature-vector was then assigned the category found for it in the residue-state-vector clustering for each network. The category assignments assigned by the clustering algorithm are the Structural Building Blocks (SBBs), and are the categories of local structure which form the basis for this study.

To facilitate the location of interesting structural regions along the protein sequence, the patterns of SBBs along the protein sequences were analyzed using simple pattern recognition software. For pattern sizes of three through seven, all of the patterns of SBBs occurring along the protein sequence which occurred in the protein Data Set Two were tabulated. Frequency counts for these patterns were also calculated. For each SBB category, the most frequently occurring patterns were examined using molecular modeling and visualization software (from Biosym Technologies, Inc.). The regions in proteins exhibiting the frequently occurring patterns of SBBs were displayed in order to analyze what structural properties they exhibited.

Results

In a network which masters the auto-association task of reproducing its input at its output layer, the activation levels of the hidden layer units must be an encoding of the input pattern, because all information from the input to the output layers passes through the hidden layer in this architecture. Since the hidden layer constitutes a "narrow channel", the encoding the network develops must be an efficient one, where each unit corresponds to important properties necessary to reproduce the input and where there are minimal activation value correlations among the units. We thus hypothesize that the encoding provided by the hidden layer activations provides the basis for general categorization of the local structure of a protein.

The most appropriate cluster size for producing meaningful SBBs was determined empirically. For each cluster size used in k-means clustering (i.e. three

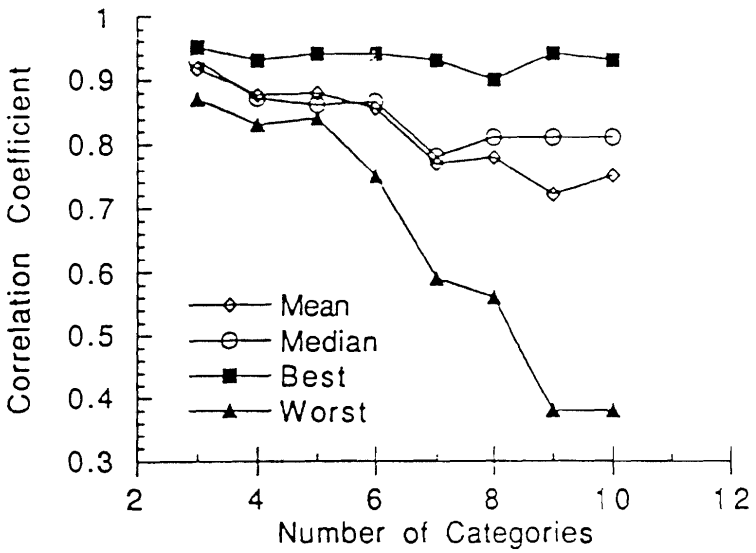

Figure 3: A comparison of the categorization results for different cluster sizes. For each cluster size used in k-means clustering (i.e. three through ten), the best correlations between the categories found in Data Set Two by the two networks were compared, separately trained on Data Set One and Data Set Two. The mean, median, best (highest) and worst (lowest) of these category correlations were then determined.

through ten), the best correlations between the categories found in Data Set Two by the two networks trained separately on Data Set One and Data Set Two were compared. The mean, median, best and worst of these category correlations were then calculated. There exists a steep relative dropoff in the mean and median correlations from clusterings using a category size of six to those using a category size of seven (Figure 3), indicating that for these data sets category selection becomes much less reproducible at a category size of seven, and further suggesting that the network is able to generalize at a category size of six. Thus, a clustering of the data into six structural categories was used throughout the remainder of this work.

For a clustering using a category set of six, the categories are general, rather than reflecting properties specific to the data set on which a network was trained. The categories found by the two networks were highly correlated, even though the two networks were trained on disjoint sets of proteins (Table 2).

To compare the SBBs and the traditional secondary structure classifications, the overlap between the classification and standard secondary structure was calculated. For each SBB category, the number of times the central residue in an SBB was specified as α-helix, β-strand or one of the other secondary structure categories by the DSSP program [Kabsch and Sander, 1983] was calculated (Figure 4). For the network trained on Data Set One, SBB category 0 clearly accounts for most of the α-helix secondary structure in

	A	B	C	D	E	F
0	-.21	-.35	-.23	-.26	-.21	**0.94**
1	-.09	-.15	**0.87**	-.10	-.13	-.22
2	-.19	**0.84**	-.10	-.04	-.20	-.36
3	**0.87**	-.16	-.11	-.12	-.09	-.23
4	-.12	-.19	-.12	-.11	**0.86**	-.22
5	-.11	-.07	-.09	**0.75**	-.13	-.25

Table 2: A comparison of the categories found in Data Set Two by a network trained on the protein sequences in that data set and a network trained on the protein sequences in Data Set One. Results shown are for the categories found with a cluster set of six. The columns are the categories (A through F) found in Data Set Two by the network trained on Data Set One. The rows are the categories (0 through 5) found by the network trained on Data Set Two. For each pair of categories the correlation between the category found by the network trained on Data Set One and the network trained on Data Set Two is given for their categorization of the sequences in Data Set Two. The best matches are indicated in bold type.

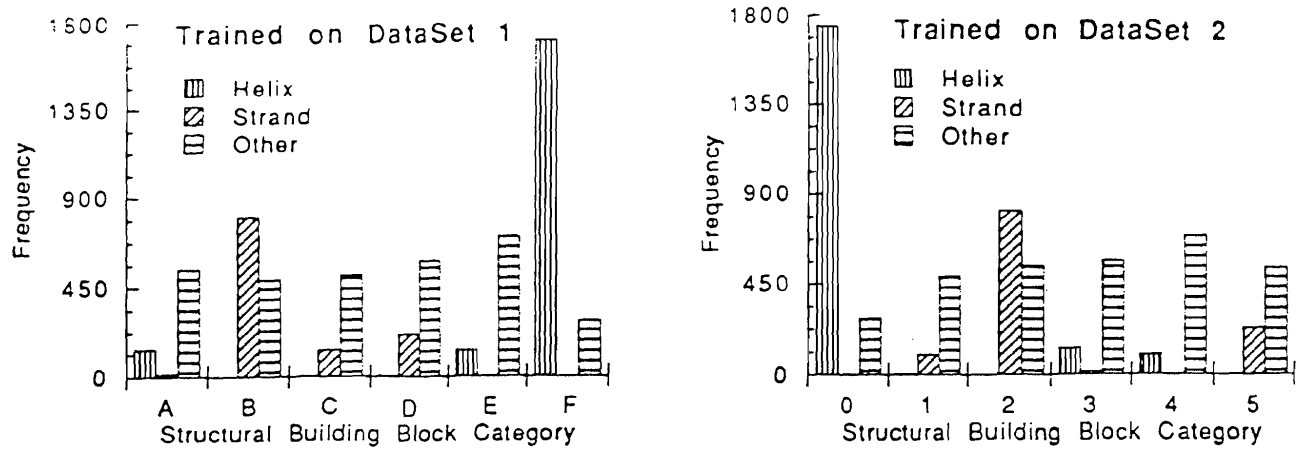

Figure 4: An analysis of the overlap between Structural Building Block categories and secondary structure classifications. For each occurrence of a SBB in the proteins Data Set Two, the DSSP [Kabsch and Sander, 1983] classifications of the central residue in the SBBs are tabulated. Frequencies are given for the SBB categories of both the network trained on Data Set One and the network trained on Data Set Two.

the sequences. SBB category 2 accounts for most of the β-strand, although it is almost as often identified with regions of "coil". Other SBB categories all have clearly delimited distributions with respect to the three secondary structure types. The generality of the categories is also shown. The SBBs found by each of the two networks which are most closely correlated (Table 2) show essentially identical frequency distributions for the related categories. (Figure 4).

In addition, there are strong amino acid preferences for the central residue in the SBBs (Table 3). For each amino acid in each SBB, the relative frequency, F_a was calculated by

$$F_a = \frac{X_a/X_t}{N_a/N_t}$$

where X_a is the number of residues of amino acid type X in SBB category a. N_a is the total number of residues in SBB category a in the protein database used in this project. X_t is the total number of residues of amino acid type X in the protein database. N_t is the total number of residues of all amino acid types in the entire database. For this calculation, the central residue in each window was the residue considered. Amino acid preferences found for these six SBBs are stronger than the preferences for traditional secondary structures in these data sets (data not shown).

To illustrate that the SBBs are significant structural elements, and not an artifact of the clustering technique, various classes of SBBs were visualized. One example is shown in Figure 5, where the 40 instances of SBB 4 along the sequence of the protein thermolysin (4tln) found with the network trained on Data Set 2 are superimposed. SBB 4 is clearly a cohesive structure, which can be characterized as having a "fish-hook" shape. Upon visualization, this structure occurs most

	A	C	D	E	F	G	H
0	1.47	0.67	0.90	1.47	1.10	0.53	0.99
1	0.69	1.32	2.07	0.48	0.39	1.25	1.62
2	0.83	1.33	0.45	0.71	1.11	0.69	0.68
3	0.92	0.46	1.43	1.27	0.73	0.80	0.43
4	0.59	0.94	1.32	0.87	1.32	2.62	1.38
5	0.81	1.62	0.69	0.55	0.92	1.12	1.27

	I	K	L	M	N	P	Q
0	1.00	1.18	1.44	1.42	0.82	0.34	1.17
1	0.60	0.57	0.51	0.73	1.91	1.50	0.77
2	1.67	0.77	1.13	1.03	0.59	1.11	0.80
3	0.50	1.24	0.40	0.53	1.08	2.95	1.03
4	0.23	1.03	0.56	0.49	2.00	0.37	0.83
5	1.38	1.01	1.02	1.01	0.40	0.99	1.24

	R	S	T	V	W	Y
0	1.18	0.67	0.71	0.91	1.33	0.84
1	0.47	1.73	1.65	0.45	0.55	0.73
2	1.01	0.80	1.28	1.89	1.06	1.35
3	0.99	1.54	1.01	0.57	0.64	0.79
4	0.87	0.99	0.66	0.46	0.75	1.02
5	1.04	1.15	1.11	1.04	0.95	1.21

Table 3: The relative frequency of each of the amino acids for the central residue position in each of the Structural Building Block classes, found by the network trained on Data Set One. The frequency counts are for that network's categorizations of the proteins in Data Set Two. Standard one-letter codes are used to represent the amino acids.

Figure 5: The Structural Building Block 4 for thermolysin. For the 40 instances of SBB 4 in thermolysin, the renderings of the backbone structural were aligned to minimize the RMS difference in the backbone atom displacement (Insight II, Biosym Technologies, Inc.). Only the backbone conformations are shown.

often at the C-terminal ends of α-helices (Figure 4) and in some loop regions.

By using molecular modeling and visualization software, several clear correlations between SBBs and protein structure were found. One class of SBB corresponds to the internal residues in helices and and one to the internal residues in strands. Also, different SBBs which correspond in many instances to N-terminal and C-terminal "caps" of helices were found [Richardson and Richardson, 1988; Presta and Rose, 1988]. In addition, SBBs which correspond to cap structures for strands were identified in many cases, a structural pattern which has not yet been described, to the authors' knowledge. Comparing these results to the frequency counts for the corresponding SBB sequence patterns confirms that the various cap-structure and structure-cap patterns are frequently occurring ones in the protein database.

Discussion

Based upon simple structural measurements, autoassociative networks are able to induce a representation scheme, the major classifications of which prove to be Structural Building Blocks: general local structures of protein residue conformation. SBBs can be used to identify regions traditionally identified as helical or strand. Other SBBs are strongly associated with the N- and C-termini of helical regions. Perhaps most interesting is that there are also SBBs clearly associated with the N- and C-termini of *strand* regions. Further, it is interesting to note that all structure, even that in the "random coil" parts of the protein, are well classified by these six SBBs. All of these results have been found both visually, using molecular modeling software and in the frequency results of the pattern generation software for the patterns of SBBs associated with these structures. Further quantification of these results is underway.

On the basis of these results, it is possible that SBBs are a useful way of representing local structure, one that is much more objective than the "traditional" model based upon α-helix and β-strand. The value of these more flexible structural representations may well be that they provide the basis for prediction and modeling algorithms which surpass the performance and usefulness of current ones.

Previous researchers have attempted novel recategorizations of local protein structure [Rooman *et al.*, 1990; Unger *et al.*, 1989]. However, the work described here differs from theirs in at least one important respect. They cluster directly on their one-dimensional structural criteria (e.g. C_α distances) and then subsequently do other processing (e.g. examination of Ramachandran plots) to refine their categories. SBBs are created by clustering on the hidden unit activation vectors created when our more extensive structural criteria (C_α distances, dihedral and bond angles) are presented to the neural network. By using the tendency of autoassociative networks to learn similar hidden unit activation vectors for similar patterns, SBBs are derived directly from multidimensional criteria without worrying about disparate dimensional extents distort-

ing the clustering, and without post-processing to refine the classifications. We hypothesize that the representations for the hidden unit vectors developed by the network also reduce the effect of spatial distortion and other "noise" in the data. This would yield cleaner data for the clustering algorithm, and more meaningful classifications. Analyses are underway to test this hypothesis, and to compare the SBB classifications to those derived from these different methods.

The results of the project described here can be readily extended. Pattern recognition techniques can be used to provide more sophisticated induction mechanisms to recognize the groupings of categories into regular expressions, and of the regular expressions into even higher-level groupings. Using molecular modeling software, the correspondence between the current categories, any higher-level structures found and the actual protein structures can be further investigated. The categories found in this research can be used as the basis for predictive algorithms. If successful, the results of such a predictive algorithm could be more easily used for full tertiary structure prediction than predictions of secondary structure. Because SBBs can be predicted for entire protein sequences, each SBB overlaps with neighboring SBBs and each SBB is a full description of the local backbone structure of that region of protein, SBB based predictions contain enough information that they can be used as input to standard distance geometry programs to predict the complete backbone structure of globular proteins.

References

Chou, P. Y. and Fasman, G. D. 1974. Prediction of protein conformation. *Biochemistry* 13:222–245.

Hartigan, J. A. and Wong, M. A. 1975. A k-means clustering algorithm. *Applied Statistics* 28:100–108.

Kabsch, W. and Sander, C. 1983. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometric features. *Biopolymers* 22:2577–2637.

Leszczynski, J. S. (Fetrow) and Rose, G. D. 1986. Loops in globular proteins: A novel category of secondary structure. *Science* 234:849–855.

Pauling, L. and Corey, R. B. 1951. Configurations of polypeptide chains with favored orientations around single bonds: Two new pleated sheets. *Proceedings of the National Academy of Science* 37:729–740.

Pauling, L.; Corey, R. B.; and Branson, H. R. 1951. The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain. *Proceedings of the National Academy of Science* 37:205–211.

Presta, L. G. and Rose, G. D. 1988. Helix signals in proteins. *Science* 240:1632–1641.

Qian, N. and Sejnowski, T. J. 1988. Predicting the secondary structure of globular proteins using neural network models. *Journal of Molecular Biology* 202:865–884.

Richards, F. M. and Kundrot, C. E. 1988. Identification of structural motifs from protein coordinate data: Secondary structure and first-level supersecondary structure. *Proteins: Structure, Function, and Genetics* 3:71–84.

Richardson, J. S. and Richardson, D. C. 1988. Amino acid preferences for specific locations at the ends of α helices. *Science* 240:1648–1652.

Ring, C. S.; Kneller, D. G.; Langridge, R.; and Cohen, F. E. 1992. Taxonomy and conformational analysis of loops in proteins. *Journal of Molecular Biology* 224:685–699.

Rooman, M. J.; Rodriguez, J.; and Wodak, S. J. 1990. Automatic definition of recurrent local structure motifs in proteins. *Journal of Molecular Biology* 213:327–336.

Rumelhart, D. E.; Hinton, G.; and Williams, R. J. 1986. Learning internal representations by error propagation. In McClelland, J. L.; Rumelhart, D. E.; and the PDP Research Group, , editors 1986, *Parallel Distributed Processing: Volume 1: Foundations.* MIT Press, Cambridge, MA.

Smith, T. F. and Waterman, M. S. 1981. Identification of common molecular subsequences. *Journal of Molecular Biology* 147:195–197.

Unger, R.; Harel, D.; Wherland, S.; and Sussman, J. L. 1989. A 3D building blocks approach to analyzing and predicting structure of proteins. *PROTEINS: Structure, Function and Genetics* 5:355–373.

Venkatachalam, C. M. 1968. Stereochemical criteria for polypeptides adn proteins. conformation of a system of three linked peptide units. *Biopolymers* 6:1425–1436.

Zhang, X. and Waltz, D. L. 1993. Developing hierarchical representations for protein structures: An incremental approach. In Hunter, L., editor 1993, *Artificial Intelligence and Molecular Biology.* MIT Press, Cambridge, MA.

Zhang, X.; Mesirov, J. P.; and Waltz, D. L. 1992. Hybrid system for protein secondary structure prediction. *Journal of Molecular Biology* 225:1049–1063.