PROTEINS: Structure, Function, and Genetics 27:249–271 (1997)

Patterns, Structures, and Amino Acid Frequencies in Structural Building Blocks, a Protein Secondary Structure Classification Scheme

Jacquelyn S. Fetrow,[1]* Michael J. Palumbo,[1] and George Berg[2]
Departments of [1]Biological Sciences and [2]Computer Science, The University at Albany, SUNY, Albany, New York

ABSTRACT **To study local structures in proteins, we previously developed an autoassociative artificial neural network (autoANN) and clustering tool to discover intrinsic features of macromolecular structures. The hidden unit activations computed by the trained autoANN are a convenient low-dimensional encoding of the local protein backbone structure. Clustering these activation vectors results in a unique classification of protein local structural features called Structural Building Blocks (SBBs). Here we describe application of this method to a larger database of proteins, verification of the applicability of this method to structure classification, and subsequent analysis of amino acid frequencies and several commonly occurring patterns of SBBs. The SBB classification method has several interesting properties: 1) it identifies the regular secondary structures, α helix and β strand; 2) it consistently identifies other local structure features (e.g., helix caps and strand caps); 3) strong amino acid preferences are revealed at some positions in some SBBs; and 4) distinct patterns of SBBs occur in the "random coil" regions of proteins. Analysis of these patterns identifies interesting structural motifs in the protein backbone structure, indicating that SBBs can be used as "building blocks" in the analysis of protein structure. This type of pattern analysis should increase our understanding of the relationship between protein sequence and local structure, especially in the prediction of protein structures. Proteins 27:249–271** © 1997 Wiley-Liss, Inc.

Key words: protein structure; secondary structure; protein conformation; protein backbone structure; protein structure classification; helix capping; strand capping; neural networks; structural building blocks

INTRODUCTION

In the analysis of protein structure, continuous regions of local structures along the polypeptide chain are defined as secondary structures. The classical examples are the α helices and β sheets originally predicted by Pauling and coworkers.[1,2] Because of their regular hydrogen-bonding patterns and repeating backbone dihedral angles, these structures are called regular secondary structures. Later, the category of β turn or reverse turn was described,[3] and the definition of these structural elements has been refined by several researchers.[4–7]

The helices, sheets, and turns together only account for about 50–55% of all protein structure on average.[8] The remaining structure has been termed "random coil," and attempts to categorize these nonregular structures have resulted in the classification of several types of loops[8,9] (reviewed in reference 10). Specific turn and loop types between regular secondary structures, such as the βα and αβ loops, have also been identified.[11–14] Compared to α helix and β strand, the loop and turn secondary structural elements are more difficult to identify because they lack the regular hydrogen bonding and repeating backbone dihedral angle patterns of the regular secondary structures; however, even though they are difficult to classify, it is clear that recurring motifs do appear in the nonregular structures along the polypeptide backbone.

A rigorous and objective categorization of the secondary structural elements is an important step in understanding protein structure and function and in understanding those interactions that stabilize proteins. Several algorithms for quantitatively assigning helix, strand, and loop regions for proteins with known three-dimensional coordinates have been developed.[15–18] Although these algorithms often agree on the location of the regular secondary structures, they usually differ on the exact endpoints of these structures.[19] Furthermore, the algorithms frequently disagree on the locations of the more irregular helices and strands and can disagree on up to one third of these classifications.

The discrepancies in secondary structure assignment and the lack of an objective classification

Contract grant sponsor: NSF, contract grant number BIR9211256.
*Correspondence to: Dr. Jacquelyn S. Fetrow, Department of Biological Sciences, The University at Albany, SUNY, 1400 Washington Avenue, Albany, NY 12222. E-mail: jacque@isadora.albany.edu.
Received 9 September 1996; accepted 11 September 1996.

scheme for the "random coil" regions make secondary structure prediction more difficult. To date, the extensive efforts to predict protein secondary structures from amino acid sequence information have been only somewhat successful. Predictions of α helix, β strand, and turns by a variety of methods[20–26] have only attained accuracies of approximately 65%. More recent work has shown that an accuracy of about 72% may be attainable by using information on the evolutionary relatedness of proteins and by combining the results of several different prediction algorithms.[27–30] An improved representation of all local protein structures, especially one that classifies nonregular structures, may enable secondary structure prediction methods to exceed these levels.

Several groups have attempted to produce such representations by objectively reclassifying protein secondary structures based on clustering of residue three-dimensional coordinates or dihedral angle differences.[31–33] While these algorithms have been successful in identifying the classical helix and strand structures, and in some cases have identified new structural motifs, there are several problems with the algorithms that might limit their usefulness. Using the "raw geometric data" (distances and angles) directly in clustering algorithms is problematic because the standard clustering algorithms are ill-suited for handling high-dimensional data. These algorithms are also very sensitive to the input information used in clustering and to its properties, such as magnitude, value range, and correlation.[34] In addition, it is unclear whether the similarity criteria adopted, such as the root mean square (rms) errors in the three-dimensional coordinates, are appropriate measures of the similarity of protein local structures.[32]

To produce a more useful, objective representation of protein secondary structures, these limitations must be overcome. We have used an autoassociative neural network (autoANN)[35] to accomplish this. An autoANN is a machine learning algorithm for a network that learns to reproduce the activity of its input units at its output units, mediated by a smaller layer of hidden units[36] (Fig. 1). We have previously encoded the geometry of seven-residue protein segments from a small database of proteins as input for such a network.[37] Since all information from the input layer passes through the smaller, hidden layer in order to produce the output layer activations, the intrinsic features of the input data (in this case, the geometry of local protein structures) are encoded in the activation values of the smaller hidden layer as the network trains. This hidden layer vector has lower dimensionality than the raw data presented as input, can be computed from the raw data, and can be used to reconstruct the raw data through the nonlinear transformations of the trained autoANN. Classifications of local protein structures were generated by clustering the hidden unit vectors for all of the segments in the database. As previously described, the resulting structure categories, called

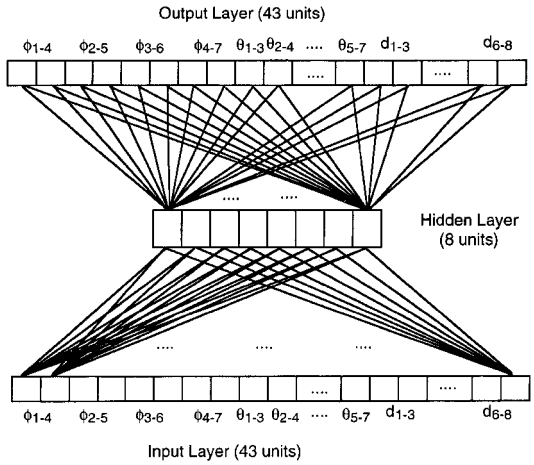

Fig. 1. Design of the autoassociative neural network used to compute the important structural features of the seven-residue protein segments. The α-carbon geometry of each seven-residue peptide is represented as a 43-unit vector; each unit of the vector is a real number between 0 and 1 and represents some aspect (atomic distances, *d*, virtual bond angles, τ, or virtual dihedral angles, φ) of the peptide backbone geometry.[37] During training, the network attempts to reproduce the input layer as the output layer; thus both the input and output layers of the network are 43-unit vectors. The hidden layer of the network is an eight-unit vector and is fully connected by weighted links to both input and output layers.[35] A standard backpropagation algorithm[36] is used to train the network by altering the weights' values and the network is fully trained when it can reproduce the input to an rms error of less than 0.08. After the network is trained, the weights are frozen and the geometry of each seven-residue peptide is again presented to the network. The activation values of the hidden layer calculated by the network for each segment are used to represent the geometry of that segment; thus the segment geometry is represented by an eight-unit vector rather than a 43-unit vector. These hidden unit activations are then clustered by a k-means clustering algorithm.

structural building blocks or SBBs, included the well-known regular secondary structures (helices and strands) as well as helix and strand capping structures.[37]

The data generated from this small database suggested that recurring patterns could be found in the random coil regions of proteins and that there were amino acid preferences in these structures; however, the database was too small to determine the statistical significance of these results. We have now extended this work to a larger database of well-resolved, nonhomologous proteins. As the original version of the network was written in LISP for the CM5 massively parallel computer, the network algorithm has been rewritten in the C programming language for use on Unix workstations. This new network was trained on the larger database and the previous results were confirmed. The results of the training and clustering algorithms were also more extensively verified. Using the SBB categories, amino acid frequencies were analyzed and statistically significant patterns in the nonregular secondary structure regions were identified. By examining backbone geometries in commonly occurring SBB sequences, several potential structural motifs have been identified. The results suggest that SBBs are useful for

understanding structural regularities in proteins, especially in the capping regions and so-called loop regions. These recurring motifs could be used in "spare parts"[38] or segment-based[23] approaches to protein structure prediction and can help to further our understanding of the relationship between amino acid sequence and local protein structure.

METHODS

Description of the Databases

Two databases were used in the work described here: 1) the original database described in reference 37; and 2) a new, larger, better-resolved database. The original database was built from an older version of the Brookhaven Protein Database[39] and contains 75 protein chains, with 13,114 residues and 12,664 seven-residue segments. The proteins are well-resolved, with a crystallographic resolution of less than 2.5 «, and have limited sequence homology, as verified by the BLAST sequence homology program.[40] In the current work, this older database was used to verify the results obtained from the new autoANN, that was rewritten in C, as described below.

For more complete statistical analysis of SBB patterns in globular proteins, an extended database of proteins with limited sequence similarity was selected from the PDB SELECT database at EMBL.[41] These proteins have 25% or less sequence identity with one another. These globular proteins have been solved by x-ray crystallography to a resolution of less than 2.2 Å and a refinement value of less than 0.2. The resulting database contains 116 different protein chains, for a total of 23,355 residues and 22,659 seven-residue segments. Most examples of tertiary structure architectures of globular proteins whose structures have been solved crystallographically are represented in this database. These proteins are listed in Table I.

From the old database 19 protein chains were found to be the same as or similar to protein chains in the new database, and these are marked by an asterisk in Table I. For some validation procedures, reduced versions of the two databases were used. The reduced databases were created by removing the 19 proteins identical or similar to the ones in the other database. The reduced version of the new database contained 97 protein chains and 19,438 segments. The reduced version of the old database was 56 protein chains with 9,471 segments. This allowed the validation tests described in the Results section to be done by using databases that had no protein chains identical or similar to those on which the networks were trained.

Encoding the Protein Segment Geometry as Input for the autoANN

The actual network and data encoding is similar to that previously described.[35,37] The geometry of contiguous seven-residue protein segments was used as input for the neural network. All segments that spanned gaps or chain breaks, and were therefore noncontiguous, were eliminated from the input databases. The input data were generated by computing the distances (d), virtual bond angles (τ), and virtual dihedral angles (ϕ) of the α-carbons in each seven-residue segment along the protein sequence, advancing one position in the sequence for each segment. Thus, most residues in the protein database appeared as the first residue in one SBB segment, the second residue in the next SBB segment, and so on; consequently, all residues except the three N- and C-terminal residues were classified in each of seven distinct, but overlapping, segments (Fig. 2). Since the three N-terminal and three C-terminal residues of each protein chain were not the centers of seven-residue segments, the number of SBB segments was less than the number of residues in the database.

The raw data for the seven-residue segments, representing distances, virtual bond angles, and virtual dihedral angles of the α-carbon conformations, were encoded as N 43-dimensional vectors for input to the autoANN, where N is the number of seven-residue segments in the database, as previously described.[35] Each of the 15 distances between nonneighboring α-carbons in a segment was encoded by two input units. Two units were used because the distribution of the data for each Cα(i)-Cα(j) distance was bimodal. The two-unit representation reflected these bimodal distributions. When a distance fell in the first mode of the distribution, the first unit was set to a value between [0,1] that was proportional to the relationship between the distance and the range of the first mode. The second unit was set to zero. When a distance was in the second mode of the distribution, the first unit was set to one and the second unit was set to a value between [0,1] that was proportional to the relationship between the distance and the range of the second mode.

Different representations were used to encode the virtual bond angles and the virtual dihedral angles. The five virtual bond angles (τ) between α-carbons were each encoded by one input unit. The angles range from 0° to 180° and were normalized to the range [0,1]. Each of the four virtual dihedral angles (ϕ) was encoded by two input units: one unit each for the sine and cosine of the angle, normalized to the range [0,1]. This representation reflected the continuity of the dihedral angles, that is, that a dihedral angle of 180° is the same as an angle of −180°.

Description of the Autoassociative Network

Originally the autoANN software was written in LISP and run on a Thinking Machines CM5 massively parallel computer. In this work, the autoassociative neural network (autoANN) software was rewritten in the C programming language to run on Unix workstations. The results obtained from training this network on the original database were compared to the results obtained from the previously published work. The autoANN is a feedforward net work[36] that has the same number of input units as

TABLE I. Database of Proteins and Summary of SBB and DSSP Secondary Structure Category Assignments

		SBB counts (%)						DSSP counts (%)†								
Name-Ch	N_r	α	β	ζ	η	τ	ι	H	E	T	B	G	I	S	O	Description
1aap-A	56	12.0	36.0	14.0	8.0	18.0	12.0	14.3	25.0	7.1	3.6	8.9	0.0	16.0	25.0	Proteinase inhibitor
1aba	87	127.2	22.2	14.8	1.1	11.1	13.6	34.5	18.4	14.5	2.3	0.0	0.0	8.0	21.8	Glutaredoxin mutant
1abk	211	151.2	7.3	14.6	4.2	5.4	7.3	56.9	0.0	10.0	0.0	4.3	0.0	11.9	17.1	Endonuclease III
1ads	315	131.1	17.5	14.2	2.9	12.0	12.3	33.6	11.8	12.7	3.2	4.8	0.0	8.9	25.1	Aldose reductase
1arb	263	19.0	39.7	14.0	0.9	14.0	12.5	5.7	31.6	13.3	0.8	9.9	0.0	10.3	28.5	Ach. protease I
1ayh	214	1.4	52.4	10.1	9.1	14.4	12.5	1.9	47.7	8.9	3.3	4.2	0.0	10.8	23.4	Glucanohydrolase H
1bab-B	146	72.1	1.4	10.7	8.6	3.6	3.6	68.5	0.0	4.8	0.0	12.3	0.0	2.7	11.6	Hemoglobin Th-Ville
1bbh-A	131	70.4	2.4	9.6	8.0	4.0	5.6	65.7	0.0	6.1	3.1	8.4	0.0	3.1	13.7	Cyto c′ *(C. vinosum)*
1bbp-A	173	17.2	44.9	13.2	0.2	13.2	11.4	9.8	48.0	8.1	0.0	3.5	0.0	11.0	19.7	Bilin binding protein
1bgc	152	71.2	5.5	8.9	5.5	4.8	4.1	73.7	0.0	8.6	0.0	2.0	0.0	3.3	12.5	Granulocyte Col-Stim factor
1btc	491	130.9	20.2	13.6	4.2	11.1	9.9	33.0	11.4	12.2	1.0	7.3	0.0	9.4	25.7	β-Amylase
1caj	258	10.3	38.1	15.1	9.9	13.5	13.1	8.1	29.5	10.5	1.9	8.1	0.0	15.5	26.4	Carbonic anhydrase I
1cmb-A	104	132.7	22.5	13.3	1.2	10.2	10.2	38.5	12.5	8.7	1.9	3.9	0.0	9.6	25.0	Met apo-repressor
1cob-A	151	12.8	39.3	14.5	5.9	14.5	13.1	0.0	39.1	15.9	2.7	4.6	0.0	14.6	23.2	Superoxide dismutase
1cpc-A	162	72.4	1.3	10.3	7.7	4.5	3.9	77.8	0.0	7.4	0.0	0.0	0.0	5.6	9.3	C-phycocyanin (A)
1cpc-L	172	72.9	3.6	7.8	7.2	3.6	4.8	71.5	0.0	12.2	0.0	1.7	0.0	3.5	11.1	C-phycocyanin (L)
1cse-I	63	115.8	31.6	14.0	2.3	19.3	7.0	17.5	30.2	19.1	3.2	4.8	0.0	4.8	20.6	Eglin-C (complex)
1dri	271	39.3	23.4	10.2	9.8	7.2	10.2	45.0	22.5	10.7	0.0	0.0	0.0	6.3	15.5	Ribose-binding protein
1end	137	46.6	10.7	13.0	9.9	9.9	9.9	43.1	0.0	11.0	2.9	8.8	0.0	7.3	27.0	Endonuclease V
1ezm	298	34.3	15.1	13.0	13.0	13.0	11.6	37.3	11.7	11.7	3.4	2.4	0.0	12.1	21.5	Elastase
1fas	61	0.0	45.5	12.7	9.1	14.6	18.2	0.0	39.3	9.8	0.0	0.0	0.0	26.2	24.6	Fasciculin 1
1fba-A	360	39.3	22.3	10.2	9.6	11.0	7.6	41.4	14.4	11.7	0.6	3.1	0.0	5.3	23.6	F-1,6-Bisp aldolase
1fcs	154	73.7	0.7	10.1	10.1	2.0	3.4	70.8	0.0	7.1	0.0	7.8	0.0	2.0	12.3	Myoglobin mutant
1fdd	106	31.0	16.0	15.0	12.0	15.0	11.0	17.0	11.3	15.1	0.9	16.0	0.0	7.6	32.1	Ferredoxin mutant
1fia-B	74	67.7	2.9	7.4	11.8	4.4	5.9	70.3	0.0	12.2	1.4	0.0	0.0	2.7	13.5	Fis protein
1gky	186	38.9	21.7	13.3	9.4	8.3	8.3	43.0	22.0	10.8	0.5	1.6	0.0	5.9	16.1	Guanylate kinase
1glt	284	28.8	26.3	13.3	10.8	10.1	10.8	31.3	29.2	9.2	0.0	5.6	0.0	7.8	16.9	Glutathione synthase
1gmp-A	96	10.0	26.7	20.0	14.4	15.6	13.3	11.5	19.8	20.8	1.0	3.1	0.0	12.5	31.3	Guanyloribonuclease
1gox	344	35.5	17.8	12.4	11.0	12.1	11.2	39.8	12.8	11.3	1.2	4.6	0.0	8.7	21.5	Glycolate oxidase
1gpb	823	45.5	18.5	11.4	10.0	8.5	6.1	45.1	15.3	10.1	0.5	4.6	0.7	7.1	16.6	Glycogen phosphorylase B
1hil-A	217	2.8	48.3	12.8	10.9	13.7	11.4	4.1	50.7	13.8	0.9	1.4	0.0	7.4	21.7	Immunoglobulin Fab
1hiv-A	99	4.3	53.8	10.8	7.5	11.8	11.8	4.0	56.6	12.1	0.0	0.0	0.0	11.1	16.2	HIV-1 protease
1hsb-A	270	25.8	34.5	9.8	9.5	11.0	9.5	24.4	38.9	11.5	0.7	2.2	0.0	9.6	12.6	Histocompatibility Ag
1ifc	131	10.4	50.4	8.8	9.6	8.8	12.0	11.5	58.8	13.7	0.0	0.0	0.0	3.1	13.0	Fatty Acid binding protein
1isu-A	62	8.9	17.9	23.2	16.1	19.6	14.3	9.7	6.5	25.8	12.9	4.8	0.0	9.7	30.7	High-Pot Fe-S protein
1l92	162	51.9	5.1	12.8	12.8	7.7	9.6	64.2	8.6	7.4	0.6	1.9	0.0	7.4	9.9	Lysozyme mutant
1lga-A	343	37.4	16.6	13.4	11.3	11.9	9.5	34.7	3.5	12.0	2.9	6.4	0.0	15.2	25.4	Lignin peroxidase
1lts-A	185	16.2	19.6	17.9	15.1	16.8	14.5	21.1	22.2	13.0	1.1	10.8	0.0	7.6	24.3	Enterotoxin (A)
1lts-D	103	21.7	37.1	10.3	10.3	10.3	10.3	22.3	36.9	12.6	0.0	0.0	0.0	10.7	17.5	Enterotoxin (D)
1nxb*	62	0.0	53.6	10.7	8.9	10.7	16.1	0.0	41.9	16.1	0.0	0.0	0.0	9.7	32.3	Neurotoxin B
1ofv	169	34.4	19.0	14.1	12.3	10.4	9.8	29.0	21.9	17.8	0.0	8.9	0.0	5.9	16.6	Flavodoxin
1omp	370	39.8	19.2	11.8	11.8	8.8	8.5	41.9	17.8	11.9	2.2	2.4	0.0	7.6	16.2	Maltodextrin-binding protein
1osa	148	56.3	3.5	12.0	12.7	7.8	7.8	62.2	0.0	7.4	2.7	0.0	0.0	9.5	18.2	Calmodulin
1paz*	120	13.2	37.7	14.0	11.4	11.4	12.3	14.2	36.7	13.3	0.8	2.5	0.0	10.0	22.5	Pseudoazurin
1pda	290	31.0	26.4	10.9	10.9	12.7	8.1	32.8	24.5	14.1	0.0	4.1	0.0	6.6	17.9	Porphobilin deaminase
1phb	405	43.6	12.0	15.5	11.5	9.0	8.3	44.0	9.6	11.9	0.7	7.7	0.0	7.2	19.0	Cyto P450 *(P. putida)*
1poa	118	41.1	10.7	14.3	13.4	11.6	8.9	40.7	6.8	17.0	2.5	5.9	0.0	7.6	19.5	Phospholipase A2 snake
1poc	134	25.0	23.4	13.3	15.6	11.7	10.9	26.9	17.2	11.9	3.7	0.0	0.0	14.9	25.4	Phospholipase A2 bee
1ppf-E	218	6.1	37.3	14.6	11.8	15.1	15.1	3.7	34.9	18.4	3.7	4.6	0.0	11.5	23.4	Leukocyte elastase
1ppn	212	18.9	28.6	13.1	13.6	12.6	13.1	23.1	17.9	13.2	4.3	2.8	0.0	10.4	28.3	Papain
1rbp	174	7.1	45.8	11.9	10.7	11.9	12.5	7.5	47.1	13.2	0.0	1.7	0.0	9.2	21.3	Retinol binding protein
1rnd	124	18.6	38.1	12.7	11.0	11.0	8.5	17.7	33.1	14.5	2.4	3.2	0.0	9.7	19.4	Ribonuclease A
1rro	108	47.1	2.9	17.7	12.8	9.8	9.8	48.2	3.7	11.1	0.0	9.3	0.0	11.1	16.7	Oncomodulin
1s01	275	23.4	25.3	13.4	13.4	11.9	12.6	29.8	17.1	16.0	2.6	0.0	0.0	9.5	25.1	Subtilisin BPN
1sbp	309	38.3	18.8	13.5	12.5	7.9	8.9	45.0	17.5	11.3	0.7	4.9	0.0	7.8	13.0	Sulfate-binding protein
1sgt	223	7.8	38.3	14.8	11.5	14.3	13.4	9.4	34.5	15.7	1.8	2.7	0.0	13.5	22.4	Trypsin
1sha-A	103	13.4	29.9	13.4	17.5	13.4	12.4	15.5	31.1	19.4	1.0	0.0	0.0	13.6	19.4	Tyrosine kinase
1shf-A	59	0.0	37.7	17.0	13.2	17.0	15.1	0.0	40.7	8.5	5.1	5.1	0.0	17.0	23.7	Tyrosine kinase (SH3)
1smr-A	299	9.6	41.0	15.7	9.9	13.0	10.9	9.7	46.2	14.7	2.0	7.7	0.0	5.4	14.4	Renin
1snc	135	26.4	23.3	14.7	10.1	14.0	11.6	24.4	29.6	15.6	2.2	2.2	0.0	10.4	15.6	Staph nuclease
1ten	89	1.2	56.6	8.4	8.4	15.7	9.6	0.0	53.9	11.2	0.0	0.0	0.0	7.9	27.0	Tenascin
1tfg	112	13.2	34.9	11.3	13.2	16.0	11.3	18.8	38.4	4.5	2.7	2.7	0.0	8.9	24.1	Growth factor β-2
1tgs-I	56	14.0	38.0	10.0	10.0	12.0	16.0	16.1	19.6	5.4	0.0	0.0	0.0	17.9	41.1	Trypsin inhibitor
1trb	316	25.2	32.0	11.3	11.6	9.0	11.0	23.7	26.9	9.8	1.3	5.1	0.0	13.3	19.9	Thioredox reductase
1tro-A	104	71.4	0.0	11.2	6.1	7.1	4.1	77.9	0.0	7.7	0.0	0.0	0.0	3.9	10.6	Trp repressor
1ttb-A	127	5.0	48.8	9.9	11.6	12.4	12.4	5.5	48.0	13.4	0.0	0.0	0.0	10.2	22.8	Transthyretin
1utg	70	68.8	0.0	14.1	6.3	6.3	4.7	71.4	0.0	7.1	0.0	4.3	0.0	4.3	12.9	Uteroglobin
1ycc	108	34.3	16.7	14.7	12.8	10.8	10.8	40.7	0.0	14.8	1.9	0.0	0.0	5.6	37.0	Cytochrome c
256b-A	106	75.0	0.0	9.0	6.0	5.0	5.0	76.4	0.0	8.5	0.0	2.8	0.0	3.8	8.5	Cytochrome B562
2aaa	476	25.5	24.0	14.9	12.6	11.5	11.5	26.9	17.9	15.3	2.3	6.3	0.0	8.4	22.9	α-amylase
2aza-A*	129	10.6	36.6	12.2	13.8	14.6	12.2	11.6	33.3	17.1	2.3	4.7	0.0	8.5	22.5	Azurin
2bop-A	85	26.6	40.5	11.4	7.6	7.6	6.3	28.2	35.3	4.7	0.0	3.5	0.0	7.1	21.2	BPV-1 E2 Protein
2ccy-A*	127	73.6	2.5	9.1	7.4	3.3	4.1	70.9	0.0	10.2	1.6	3.9	0.0	1.6	11.8	Cyto c′ *(R. molisch.)*
2cdv*	107	15.8	11.9	23.8	18.8	13.9	15.8	25.2	9.4	13.1	1.9	2.8	0.0	21.5	26.2	Cytochrome c3
2cpl	164	13.9	29.1	15.8	14.6	13.9	12.7	12.2	29.3	17.7	3.1	1.8	0.0	11.0	25.0	Cyclophilin A
2ctc	307	34.9	20.6	13.0	11.3	10.6	9.6	36.8	16.3	12.1	1.3	1.0	0.0	13.0	19.5	Carboxypeptidase A
2cts*	437	52.9	7.2	12.3	9.7	10.4	7.4	58.8	1.4	10.5	1.1	2.3	0.0	6.2	19.7	Citrate synthase

(continued)

TABLE I. *(Continued)* Database of Proteins and Summary of SBB and DSSP Secondary Structure Category Assignments

		SBB counts (%)						DSSP counts (%)†								
Name-Ch	N_r	α	β	ζ	η	τ	ι	H	E	T	B	G	I	S	O	Description
2cyp*	293	40.8	14.3	14.6	10.8	10.8	8.7	45.7	5.5	12.3	2.1	4.4	0.0	8.5	21.5	Cyto c peroxidase
2er7-E	330	6.8	41.1	13.9	12.7	13.9	11.7	7.6	44.2	14.2	1.2	3.6	0.0	10.6	18.5	Endothiapepsin
2had	310	37.8	19.4	12.8	12.2	8.9	8.9	34.2	14.2	14.8	0.0	7.7	0.0	8.7	20.3	Dehalogenase
2hpd-A	457	45.2	13.5	14.2	10.6	8.9	7.5	48.1	10.9	9.9	0.4	5.3	0.0	8.1	17.3	Cyto P450 (*B. mega.*)
2ihl	129	32.5	5.7	17.9	17.9	12.2	13.8	29.5	6.2	23.3	4.7	10.9	0.0	8.5	17.1	Lysozyme
2lal-A	181	0.6	48.6	14.3	9.1	13.1	14.3	0.0	46.4	16.6	2.2	1.7	0.0	9.9	23.2	Lentil lectin (A)
2lal-B	47	2.4	63.4	9.8	7.3	9.8	7.3	8.5	63.8	0.0	0.0	0.0	0.0	8.5	19.2	Lentil lectin (B)
2mhr	118	67.0	2.7	10.7	7.1	5.4	7.1	64.4	0.0	7.6	0.0	5.9	0.0	5.1	17.0	Myohemerythrin
2mnr	357	35.6	24.8	12.0	8.3	9.1	10.3	40.1	18.8	11.2	3.1	1.7	0.0	8.4	16.8	Mandelate racemase
2msb-A	111	15.2	31.4	11.4	14.3	16.2	11.4	18.9	30.6	12.6	1.8	0.0	0.0	10.8	25.2	Mannose binding protein A
2pia	321	16.5	33.3	13.3	11.4	13.7	11.8	15.0	29.9	18.1	0.3	1.9	0.0	9.7	25.2	Phthalate Diox Reductase
2rn2*	155	33.6	29.5	9.4	9.4	10.1	8.1	34.8	28.4	12.3	1.9	0.0	0.0	7.1	15.5	Ribonuclease H
2scp-A	174	61.3	2.4	10.7	11.9	7.7	6.0	56.3	4.6	11.5	0.0	7.5	0.0	8.1	12.1	Sarcoplasmic Ca-binding Protein
2sga	181	5.7	44.0	12.6	10.3	15.4	12.0	6.6	54.1	13.3	1.1	3.3	0.0	4.4	17.1	Proteinase A
2sn3*	65	10.2	23.7	20.3	13.6	13.6	18.6	12.3	18.5	18.5	6.2	0.0	0.0	13.9	30.8	Scorpion neurotoxin
3adk*	194	50.5	16.5	8.0	8.5	7.5	9.0	54.6	12.9	11.9	0.0	0.0	0.0	5.2	15.5	Adenylate kinase
3b5c*	85	35.4	13.9	17.7	11.4	10.1	11.4	24.7	22.4	25.9	1.2	7.1	0.0	3.5	15.3	Cytochrome B5
3chy	128	38.5	21.3	13.9	8.2	8.2	9.8	45.3	17.2	8.6	0.0	0.0	0.0	9.4	19.5	CheY
3cla	213	28.0	31.9	11.6	8.7	11.1	8.7	28.2	28.6	13.2	0.0	1.4	0.0	8.5	20.2	Chloramph ace-transferase
3dfr*	162	19.9	33.3	12.2	10.3	12.2	12.2	19.1	31.5	9.3	1.2	5.6	0.0	12.4	21.0	DHF reductase
3grs*	461	27.7	29.2	11.2	11.2	9.5	11.2	28.6	24.1	12.2	1.1	5.6	0.0	9.1	19.3	Glutathione reductase
3il8	68	21.0	32.3	11.3	9.7	16.1	9.7	22.1	25.0	10.3	2.9	4.4	0.0	10.3	25.0	Interleukin 8
3rub-S	123	20.5	22.2	13.7	12.8	17.1	13.7	22.0	22.0	14.6	1.6	0.0	0.0	12.2	27.6	Rubisco
3sgb-I	50	20.5	34.1	9.1	9.1	13.6	13.6	20.0	22.0	14.0	2.0	0.0	0.0	10.0	32.0	Ovomucoid inhibitor
3sic-I	107	13.9	38.6	15.8	11.9	6.9	12.9	15.0	36.5	11.2	0.9	0.0	0.0	15.9	20.6	Strepto subtilisin inhibitor
4blm-A	256	34.8	21.2	14.0	12.8	7.6	9.6	35.6	18.8	13.3	0.0	7.8	0.0	7.0	17.6	β-Lactamase
4enl	436	35.8	18.1	13.3	11.9	10.7	10.2	39.2	16.3	10.8	0.7	6.0	0.0	7.6	19.5	Glycerate hydrolase
4fxn*	138	40.9	22.0	9.9	10.6	7.6	9.1	34.1	21.0	16.7	1.5	2.2	0.0	8.7	15.9	Flavodoxin
4gcr*	174	3.0	42.3	11.9	8.3	14.9	19.6	2.9	46.0	8.1	2.3	6.3	0.0	12.1	22.4	Gamma-B crystallin
4sgb-I	51	0.0	44.4	15.6	13.3	15.6	11.1	0.0	23.5	21.6	5.9	0.0	0.0	3.9	45.1	Potato inhibitor
5fbp-A	307	30.9	27.2	11.3	10.0	10.3	10.3	29.3	23.8	10.1	2.3	8.1	0.0	8.5	17.9	Fructose-1,6-bisphosphatase
5p21	166	31.9	33.1	11.3	8.8	8.1	6.9	34.3	26.5	12.1	0.0	1.8	0.0	7.8	17.5	H-Ras P21 protein
7aat-A	401	40.5	18.0	12.4	11.1	9.4	8.6	46.1	14.0	12.7	0.0	1.5	0.0	4.7	21.0	Asp aminotransferase
8abp*	305	40.1	21.1	11.4	9.0	9.4	9.0	42.6	21.3	8.9	0.7	4.9	0.0	7.5	14.1	Arabinose binding protein
8acn	753	28.3	26.2	13.4	11.2	9.9	11.0	29.6	17.8	13.2	2.8	5.4	0.0	8.6	22.6	Aconitase
8rxn-A*	52	0.0	13.0	30.4	19.6	19.6	17.4	0.0	15.4	26.9	7.7	17.3	0.0	1.9	30.8	Rubredoxin
9ldt-A	331	37.9	20.6	10.5	8.9	12.0	10.2	40.2	17.5	8.5	0.3	5.1	0.0	8.8	19.6	Lactate dehydrogenase
9rnt*	104	14.3	37.8	10.2	12.2	13.3	12.2	15.4	26.9	17.3	1.9	0.0	0.0	12.5	26.0	Ribonuclease T1
9wga-A*	171	8.5	16.4	18.8	25.5	11.5	19.4	9.4	9.4	13.5	7.0	14.6	0.0	17.0	29.2	Wheat germ agglutinin
Totals for dataset:		31.2	24.0	12.8	11.2	10.7	10.2	32.4	20.7	12.2	1.4	4.2	0.03	8.8	20.1	

*denotes same or similar protein in original database
†DSSP Legend: H, α Helix; E, β strand; T, 3,4,5 turn; B, β bridge; G, 3-helix; I, 5-helix; S, 5-residue bend; O, not classified by DSSP.

output units and a smaller number of hidden units (Fig. 1). It trains to reproduce the activation values of its input units as its output. If the network learns to do this, then it must have developed a concise representation of the input data in the activations of the small hidden layer; thus, we hypothesized that a large input vector encoding a peptide conformation could be reduced to a smaller vector in the hidden layer and that the activation values of the hidden units would still contain the relevant local structural information. Given the limitations of clustering algorithms,[34] this reduced representation of the raw data would be more suitable for clustering than the raw data themselves.

The autoANN was trained on the input data (protein segment geometry) using a standard backpropagation algorithm.[36] All of the segments from the database were presented to the network, and the accumulated differences of the outputs from the associated inputs were then used by backpropagation to modify the weights in the network; one complete cycle of data presentation and weights modification is called an epoch. This process was repeated until the rms difference between the actual output and the input was *less than* 0.08, which usually took about 1000–2000 epochs. Earlier work stated that the networks trained to an rms difference of 0.01.[37] The network implemented on Unix workstations could not be trained to this level. We investigated this difference and were unable to explain it. Additional training to as many as 8000 epochs did not significantly lower the rms difference between the input and the output, nor did it improve the ultimate biological relevance of the clusters. The same data trained on a public domain backpropagation network[42] produced rms values similar to our current results. Furthermore, we compared the hidden unit activation values from a network created in the original study to one of our networks trained on the same database. The correlation between the activation values was better than 0.99, thus we assumed that the new networks were returning data quite similar to those previously reported. The time to train a network using the new autoANN software is about 10 hours on a Silicon Graphics (SGI) Indigo2

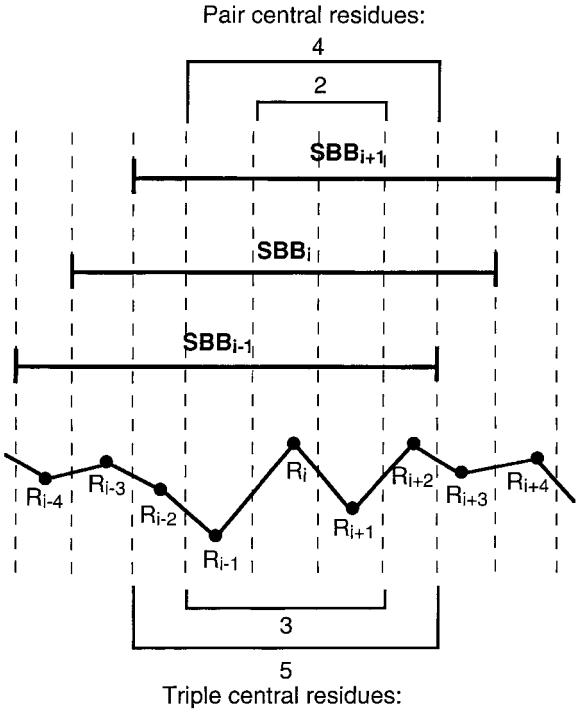

Fig. 2. The overlapping nature of SBB categories. Because SBBs are classifications of the local structure of a seven-residue protein segment, two successive SBBs in a protein overlap six of their seven residues. The SBB classification for a segment is associated with the residue in the middle of the segment, thus the SBB classification for residue R_i is actually the local structure category of the seven-residue segment centered at residue R_i. For the SBB pair SBB_i and SBB_{i+1} the central two residues are R_i and R_{i+1}, and the central four residues are R_{i-1}, R_i, R_{i+1} and R_{i+2}. For the SBB triple SBB_{i-1}, SBB_i and SBB_{i+1} the central three residues are R_{i-1}, R_i and R_{i+1}, and the central five residues are R_{i-2}, R_{i-1}, R_i, R_{i+1} and R_{i+2}.

R8000 workstation. The network results were validated as described in the Results section.

Clustering of the Hidden Unit Activations

After training was complete, the autoANN weights and biases were frozen and the activation values of the hidden unit vectors for each seven-residue segment were computed. These eight-unit vectors were clustered using a k-means clustering algorithm,[43] implemented within the Splus statistical software package (StatSci) on an SGI workstation. In a k-means clustering algorithm, the number of clusters, *k*, is chosen and *k* vectors are arbitrarily assigned as initial cluster centers or "seeds." Each eight-unit vector is then assigned to one of the clusters by calculating the euclidean distance between the vector and each center and then assigning the vector to its "closest" center. After all eight-unit vectors are assigned to a cluster, cluster centers are recalculated. This process of assigning segments and recomputing the cluster centers is repeated until it converges, that is, the centers and the segment assignments do not change. To validate the clustering process, each clustering analysis was repeated a minimum of 70 times starting with different random choices of cluster centers.

K-means clustering does not provide a method for automatically determining the optimum number of clusters, so various numbers of clusters were tried, with *k* varying from 2 through 12. Six clusters gave the most consistent results (see Results), so the six clusters were designated as six structural building block (SBB) categories. The six SBBs were given the designations alpha (α), beta (β), zeta (ζ), eta (η), tau (τ), and iota (ι), based on their structures (see Results). A seven-residue protein segment is said to belong to a particular SBB category when the hidden unit vector calculated by the autoANN from that segment's geometry is assigned to the cluster corresponding to that SBB category.

Analysis of Amino Acid Frequencies in the SBBs

For each position in the seven-residue segment, 1–7, of each SBB category, the frequency of occurrence of each amino acid type was calculated. The normalized frequency of occurrence of amino acid type X at position i in each SBB category s was calculated as

$$f_{x,s,i} = (X_{s,i}/X_{tot})/(N_{s,i}/N_{tot}) \quad (1)$$

where $X_{s,i}$ is the total number of type X residues at position i in SBB category s, X_{tot} is the total number of type X residues in the database, $N_{s,i}$ is the total number of all residues at position i in SBB category s, and N_{tot} is the total number of all seven-residue segments in the database.

The statistical significance of the amino acid frequency data was determined by applying a chi square test for each amino acid type at each position of each SBB category. A log linear model was used to determine the expected distribution of data. Frequencies are reported as significant at a 0.95 confidence level.

Analysis of Patterns in the SBBs

In the database of protein chains classified into SBBs, some SBB pairs and triples occur more often than their individual, independent occurrences would warrant, leading us to analyze pairwise and triplet occurrences of the SBBs in the database. (Occurrence of some quadruples was significant, but quintuples were not done because there were not enough data to determine statistical significance.) Pattern analysis was done using a program, RelFreq, written in the C programming language. This program takes as its input the SBB assignment for each seven-residue segment in the database. The program analyzes the occurrences of the SBBs, individually, in pairs, and in triples. These data prepare queries for the SPlus statistical software package (StatSci) proportions test. Based on the occurrence data for the individual SBB categories, SPlus determines if the occurrence of pairs or triples of SBB categories significantly differs from that expected of independently occurring pairs or triples at a 0.99 confidence level.

To evaluate whether the SBB pairs and triples represent structural motifs, the sequences of ϕ and ψ

angle values at each position of the instances of the pairs and triples were analyzed. To recognize similar backbone conformations despite variations in ϕ and ψ values, the Ramachandran (ϕ, ψ) map was divided into six regions using a simplified version of the regions proposed by Zimmerman and colleagues[44](Fig. 3). The sequences of the Ramachandran regions for the two and four central residues in pairs and the three and five central residues in the triple instances (Fig. 2) were examined to see which Ramachandran region sequences occurred most often. The frequency of occurrence of the most common sequences for each pair and triple were computed.

RESULTS

Overview

The autoANN software used in this work was rewritten to run on Unix workstations. It was verified by correlating its results to those produced by the previously written software trained on the original protein database. The original database and new network software were then used to empirically optimize the network and clustering parameters. However, the small size of the original database prevented an analysis of the statistical significance of the pair and triple patterns in the SBBs. Thus, a new, larger database was developed. This new database was used to train the autoANN with the same network parameters that were optimized for the original database. The generality of the trained networks and of the clustering results was validated by comparison of the results from the network trained on the larger database to results from the network trained on the original database. The new database of SBBs was then analyzed for amino acid composition and patterns of SBBs.

AutoANN Training and Parameter Selection

The autoANN was trained with a momentum constant of 0.9 and an initial learning rate of 0.00001.[36,45] The learning rate was incremented 0.000001 every 100 epochs until it reached a maximum value of 0.000043. These training parameters were experimentally determined to yield the smoothest reduction in rms error during training, until the rms error cutoff of less than 0.08 was reached. Momentum constants ranging from 0.0 to 0.75 and initial learning rates from 0.0000001 to 0.01 were tested, but resulted in either slower reduction in rms error or networks that became unstable. In the unstable networks, the rms error increased dramatically, varied erratically, and did not subsequently decrease significantly during the course of the run. A wide variety of learning rate change schedules were also tried, including relatively large changes to the learning rate over the course of training and large initial learning rates that were reduced over the course of the training. Of all values tried, the learning rate schedule described above produced the most consistent results for those networks that trained to a low rms; these empirically determined values were subsequently used to train all networks described here.

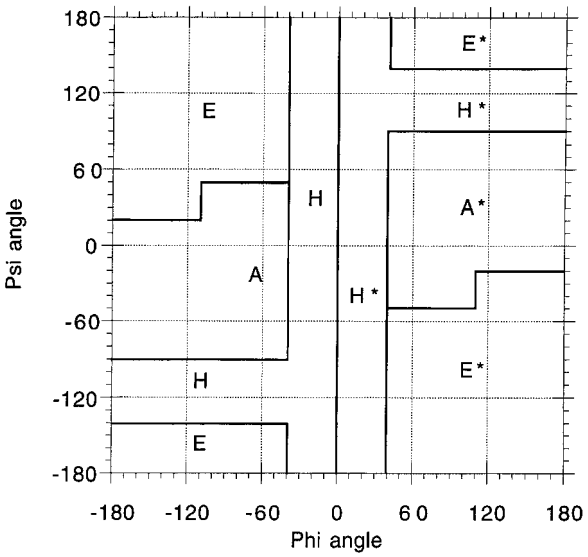

Fig. 3. Ramachandran map regions used for analyzing SBB triples. The Ramachandran map (ϕ and ψ angles) was divided into six regions using a simplified version of the regions described by Zimmerman et al.[44] Region A contains conformations where $-180° < \phi \leq -40°$ and $-90° < \psi \leq 20°$, or $-110° < \phi \leq -40°$ and $20° < \psi \leq 50°$. Region E includes conformations in the range $-180° < \phi \leq -40°$ and $50° < \psi \leq 180°$, or $-180° < \phi \leq -40°$ and $-180° < \psi \leq -140°$, or $-180° < \phi \leq -110°$ and $20° < \psi \leq 50°$. Region H classifies all other conformations where $\phi \leq 0°$. Region A represents helical conformations, region E extended conformations, and region H is a rarely occupied, energetically unfavorable conformation. Regions A*, E*, and H* are the symmetric equivalents of A, E, and H created by rotating 180° around the origin.[44] Region A* represents left-handed helical conformations.

Validation of the AutoANN Training

The generality of the trained autoANNs was validated. Ideally, the performance of the trained network should reflect the general geometric regularities of the virtual bond angles, dihedral angles and distance data in globular proteins. But, an autoANN with large numbers of weights and biases can overfit to, or partially memorize, a particular database.[45] In this case, the results of the network are too closely based on the values found in the dataset on which it was trained and the network will perform poorly when tested on other proteins. To determine if the autoANNs were generalizing from the database or if they were overfitting to the database on which they were trained, a simple cross-validation technique was used. Networks trained on each of the two databases, the original database[37] and the new database, were evaluated using those proteins from the reduced version of the other database on which they were not trained. The network rms values produced by these training and test sets were compared to evaluate whether the nets were generalizing or overfitting. The network trained on the original database was trained to an rms value of 0.071. When this network was tested on the 97 nonhomologous

proteins from the new database that were not in the original database (Table I), the resulting rms was 0.072. Another network trained on the new database was also trained to the cutoff rms value of 0.078. When this network was tested on the 56 disjoint proteins from the original database, the rms was 0.082. If the networks were memorizing or overfitting to the training data, significantly lower rms values for the training than for the testing databases would have been observed; however, each network had similar rms values for both its training and test sets (0.071 compared to 0.072, and 0.078 compared to 0.082), even though the two sets were disjoint. This result supports the contention that both networks are generalizing to the regularities of the α-carbon angles and distances in globular proteins, rather than overfitting to the particular values presented in their respective training databases.

Validation of the Clustering Analysis

According to our hypothesis, each eight-unit vector (the hidden unit activations of the autoANN) should contain all of the relevant structural information encoded in its associated 43-unit input vector and should, therefore, be a concise representation of the most important structural features of the protein segment. Thus, clustering on these vectors should produce relevant classes of local protein structure. Previously, different cluster sizes were tested and six categories (given the letter names A through F) were found to give the most consistent results by cross-validation.[37] For the rewritten autoANN, the clustering validations were done more thoroughly. Three separate issues needed to be addressed. First, the meaningfulness of six clusters or categories (rather than three, five, seven or some other number) needed to be demonstrated. Second, to show that the clustering was meaningful, and not arbitrary, clustering the hidden units from a single network must be reproducible. Third, to demonstrate that the classifications produced from distinct networks were general, clustering hidden units from separate networks must produce comparable results.

The activation values of the eight-unit hidden vectors computed by the trained autoANN on the original database were clustered using a k-means clustering algorithm. K-means clustering algorithms require that the number of clusters or groups be specified, so separate tests were done on vectors from a single network with numbers of clusters ranging from two to twelve. For each test, clustering was performed seventy times. The number of distinct results obtained from the seventy attempts ranged from one (for two clusters) to seventy (for ten, eleven, and twelve clusters), with a distinct plateau for five, six, and seven clusters (data not shown). The increase in the variability of the clustering beyond seven suggests that clusters of five through seven gave the most consistent results among sizes large enough to represent a set of plausible local structural categories. Furthermore, of the 70 runs for each of five, six, and seven clusters, one result always predominated, demonstrating that clustering runs on data from a single network produced one clear winner, at least for five, six, and seven clusters.

To determine if multiple net runs produced similar groupings and to make a decision on the cluster set that produced the most general, reproducible categories, a second analysis was done using five networks trained from different starting weights and biases. For each network, the hidden unit activations were clustered by the k-means algorithm into five, six, or seven groups. For each cluster set, the five classifications from the different networks were correlated. These ten pairwise correlations of categories for each cluster set were examined. None of the classifications for a cluster set of seven were closely correlated. For a cluster set of five, four of the ten pairwise comparisons were highly correlated. For a cluster set of six, six of the ten comparisons resulted in highly correlated classifications. On the basis of this analysis, a cluster set of six was chosen as that giving the most general and reproducible classifications. Further, this experiment demonstrated that clustering results from different networks produce comparable (but not perfect) results.

To further confirm that the clusterings underlying the SBB categories reflect general geometric regularities in local protein structure and are not arbitrary, a "negative control" database of seven-residue segments was developed. The virtual dihedral angles of a seven-residue segment were rotated through 30° increments and the α-carbon coordinates were saved at each increment. The segments were *not* tested for physically overlapping conformations. The resulting database of 20,736 segments differed from the SBB database in two respects. First, the segments in the database were a sampling of the continuum of the virtual dihedral angle combinations of the seven residues. This is in contrast to the SBB database, where the geometries in the segments are only those physically realizable geometries that actually occurred in the SBB database of proteins. Second, the control database only contained one example of each sampled conformation, whereas the SBB database contains many examples of similar segments (e.g., helical segments). If clustering on the control database produced categories comparable to those of the SBBs, the hypothesis that SBBs represent local structural regularities would be undermined.

The control database was then used to train an autoANN with the same training parameters as the other networks described here. This net trained to an rms difference between input and output values of 0.056. The hidden unit activation values obtained from the autoANN trained on the control database were then clustered into six groups 70 different times. These clusterings produced inconsistent results; 70 clustering runs produced 70 distinct clusterings. Similar results were obtained for other cluster

TABLE II. Separation of Six Clusters that Correspond to the Six SBB Categories

	Within-category distances		Distance from other category centers				
	Mean	(SD)	SBB-β	SBB-ζ	SBB-η	SBB-τ	SBB-ι
α	0.179	(0.170)	1.361	0.708	0.705	0.998	0.998
β	0.402	(0.128)		0.932	0.886	0.587	0.632
ζ	0.528	(0.102)			0.642	0.858	0.629*
η	0.559	(0.083)				0.567*	0.783
τ	0.499	(0.096)					0.658
ι	0.485	(0.122)					

*Indicates that the mean within category distance plus one standard deviation is greater than the center-to-center distance between the two categories.

set sizes. Although the inability to find consistent clusters in a uniform sampling of possible geometries was not surprising, it supported our hypothesis that the clusters from the SBB database reflect classes of local protein structure. Clusterings of hidden unit activation values computed by the network from actual protein geometries resulted in consistent, reproducible results, whereas data uniformly sampled from all virtual dihedral angle combinations did not, suggesting that the SBB classifications, that is, the clusters, are based on the allowed conformations in the local structure of globular proteins.

To measure cluster separation, the euclidean distances from each segment in a given cluster to the cluster center were computed. The average of these "within-cluster" distances was compared to the "center-to-center" distances between clusters (Table II). In all but two cases, the center-to-center distance between any two clusters is greater than the mean distance of all vectors in the cluster from its center plus one standard deviation. In these two cases, ζ to ι and η to τ, the center-to-center distance is close to the mean distance plus one standard deviation. This result indicates that the individual clusters are fairly well separated from one another.

Description of the Structural Building Block Categories in the New Database

Following complete validation and optimization of the new autoANN on the original database, the network was trained on the new, extended database. Following training, the hidden unit vectors for each of the 22,659 segments were computed and clustered into six categories to produce the six SBB categories, α, β, ζ, η, τ, and ι, described here. We found 7062 (31.2%) SBB-α segments, 5431 (24%) SBB-β segments, 2908 (12.8%) SBB-ζ segments, 2534 (11.2%) SBB-η segments, 2424 (10.7%) SBB-τ segments, and 2300 (10.2%) SBB-ι segments. The percentage of each SBB type found in every protein are presented in Table I. For comparison, the DSSP program[15] applied to the 116 proteins our database found 7577 (32.4%) helix (H) residues, 4841 (20.7%) strand (E) residues, 2859 (12.2%) turn (T) residues, 2048 (8.8%) 5-residue bends (S), 329 (1.4%) β-bridge (B) residues, 992 (4.2%) 3-helix (G) residues, and 6 (0.03%) 5-helix (I) residues (Table I). 4703 residues or 20.1% of this database were not assigned to any category by the DSSP algorithm.

Examination of globular protein structures supports the hypothesis that the hidden unit activations encode biologically relevant structural information. Sulfate binding protein from *Salmonella typhimurium* (1sbp)[46] and β-lactamase, chain A, from *Bacillus licheniformis* (4blm)[47] were chosen to display the results, although similar patterns are found in all proteins that were examined. Sulfate binding protein, which is composed of 309 residues and was solved to a resolution of 1.7 Å, is a two-domain protein consisting of two α/β domains and contains parallel β sheets. β-Lactamase chain A is 256 residues in length, has been solved to a resolution of 1.0 Å, and is also a two-domain protein. One domain is an α + β domain with an antiparallel β sheet; the other domain is largely α-helical, with one small antiparallel β sheet. Thus, a variety of common globular protein domain types are represented in these two proteins.

The structures of sulfate binding protein and β-lactamase, chain A, are displayed in Figure 4, colored by their DSSP classification or by the SBB category of the central residue in each seven-residue segment. In Figure 5, the SBB classification and DSSP assignment are compared for each residue in these proteins. As found with the original software and database,[37] two SBB categories closely correspond to the helix and sheet structures. Residues assigned to SBB-α and β correlate with DSSP categories H (α helix) and E (β sheet), respectively. The central residue of the SBB-α category is classified by the DSSP program as H 87.9%, as T 4.9%, as G 5.2%, and as unclassified (other) 1.2% of the time. The central residue of SBB-β is classified by the DSSP program as E 65.6%, S 6.1%, B 1.7%, and T 0.1% of the time; 26.4% of the central residues in SBB-β segments are not classified by DSSP. These correlations are found in all proteins in the database (Table I and Fig. 6).

Our method did not distinguish between the conformations of β strands that participate in parallel and antiparallel β sheets; both are found as runs of SBB-β residues (see the parallel (1sbp) and antiparallel (4blm) β sheets in Figures 4 and 5). This result is expected because strands that participate in parallel and antiparallel β sheets have similar local

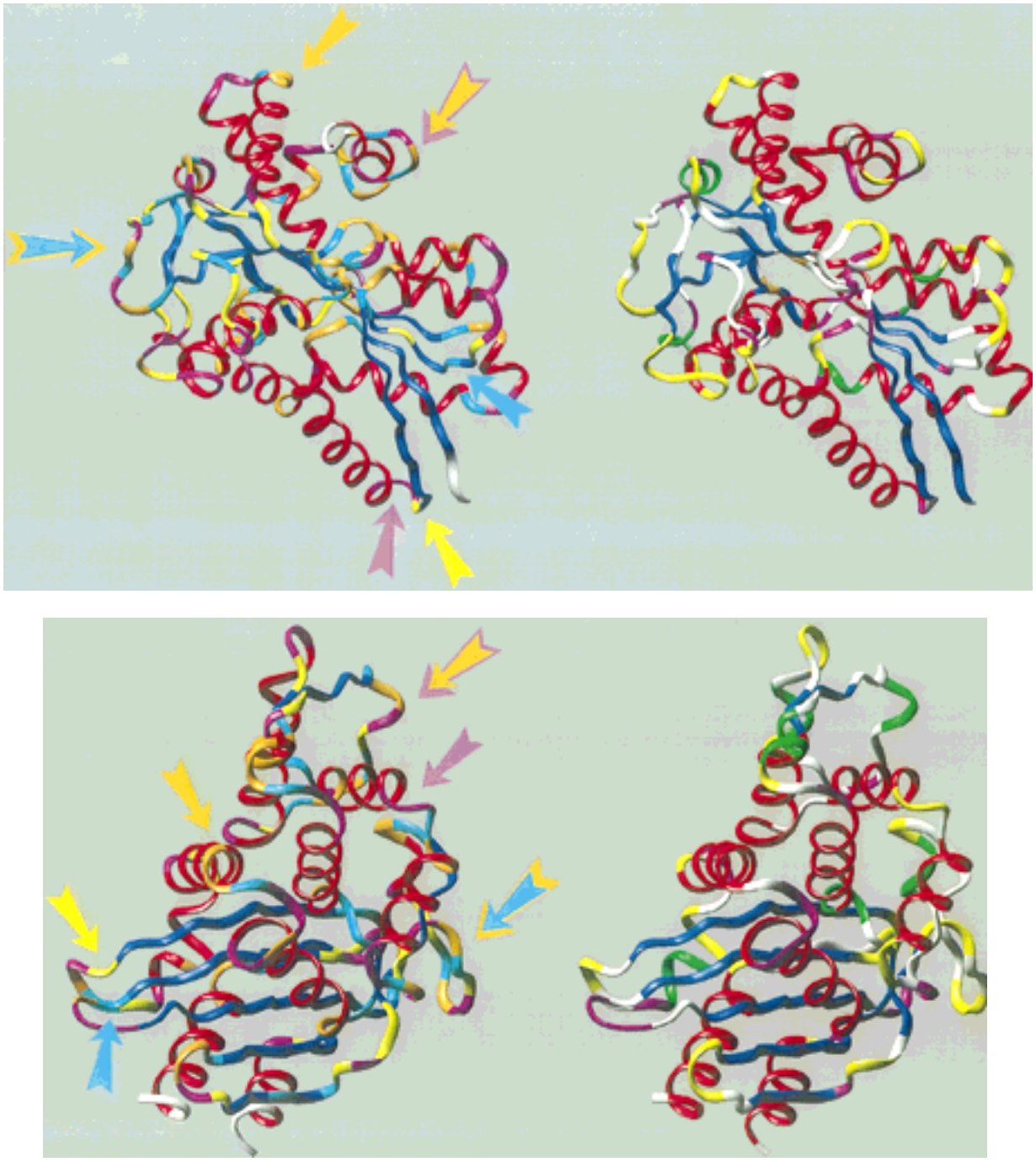

Fig. 4. Positions of SBB (A and C) and DSSP (B and D) classes in the tertiary structure of sulfate binding protein (1sbp, A and B) and β-lactamase (4blm, C and D). **A, C:** Each residue is colored by the SBB category of that segment of which it is the central residue: SBB-α, red; SBB-β, blue; SBB-ζ, orange; SBB-η, magenta; SBB-τ, yellow; SBB-ι, cyan. The three N-terminal and three C-terminal residues are delineated as white because they are not the central residue of any SBB segment. SBB-α and β largely correspond to the regular secondary structures, helix and strand. SBB-ζ and η are often found at the N- and C-termini of helices (orange and magenta arrows, respectively), while τ and ι are found at the N- and C-termini of strands (yellow and cyan arrows, respectively). SBB-ζ, η, τ, ι, and, rarely, α and β, categories are also found in the nonregular secondary structure regions. Examples of short patterns in the coil regions are indicated by bicolored arrows: SBB-ζ followed by SBB-η (orange-magenta), a "tight" turn and SBB-ι followed by SBB-ζ (cyan-orange), an "S-type" turn. **B, D:** Residues are colored by their DSSP category assignment: helix (H), red; strand (E), blue; turn (T), yellow; 3-helix (G), green; bend (S), magenta; and β bridge (B), orange; residues unclassified by DSSP are white. (Pictures were printed from a Silicon Graphics R8000 Extreme graphics workstation using the InsightII modeling package from Biosym Technologies, Inc.)

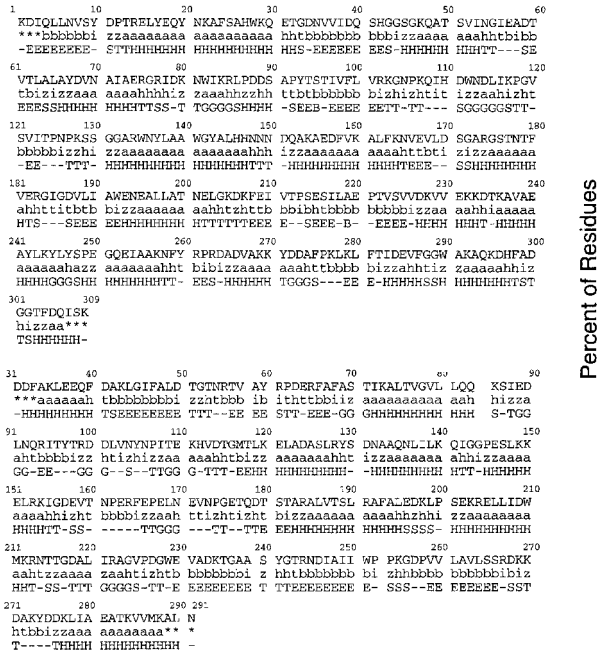

Fig. 5. Comparison of SBBs (this work) and DSSP secondary structure assignments in the sequences of sulfate binding protein **(A)** and β-lactamase **(B)**. The first row is the amino acid sequence, given as the one-letter amino acid code. The second row is the SBB category (α, β, ζ, η, τ, and ι) for the central residue of each seven-residue segment, and the third row is the DSSP category (H, helix; E, sheet; T, turn; S, bend; G, 3/10 helix; and B, isolated β bridge). A dash indicates that a residue is unclassified by DSSP. To improve readability, the SBB categories α, β, ζ, η, τ, and ι are labeled as their English equivalent, a, b, z, h, t, and i, respectively.

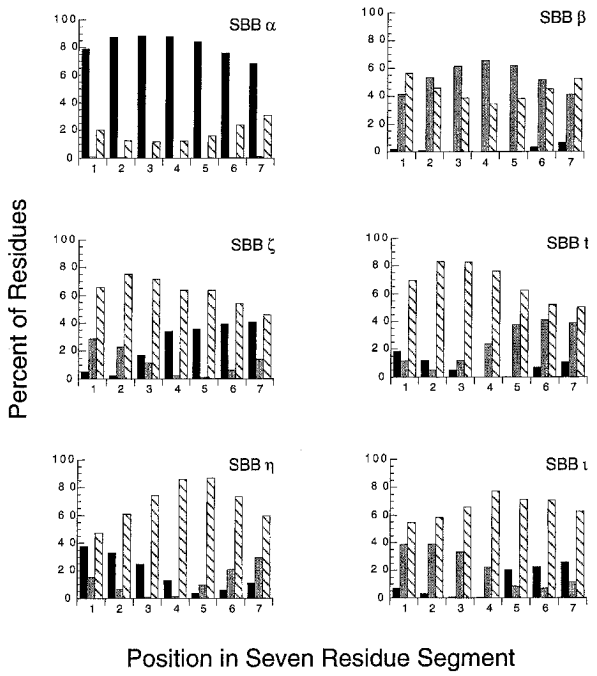

Fig. 6. Comparison of the six SBB categories to the classical secondary structures in the database of 116 proteins. Each seven-residue segment was assigned the SBB category of its central residue, then each residue in the seven-residue segment was assigned to helix (black bars), strand (gray bars), or coil (striped bars) by the DSSP program.[15] The total percentage of each secondary structure type was calculated for each position and these data are illustrated as a histogram. The correlation of SBB-α with helix and SBB-β with strand is clear. The concentration of helix at the C-terminus of SBB-ζ segments and at the N-terminus of SBB-η segments suggests that these categories act as helix-capping structures. Likewise, the concentration of strand at the C-terminus of SBB-τ segments and at the N-terminus of SBB-ι segments suggests that these categories act as strand-capping structures.

backbone conformations at the level of α-carbon geometry, but differ in their relationship to other β strands; as it is currently encoded the autoANN is only given local α-carbon geometries and would not be expected to recognize tertiary interactions. Runs of SBB-β residues are found in extended regions that are not part of β sheets (not shown), further demonstrating that the SBB categories represent local geometries that are not necessarily part of larger hydrogen-bonded structures and thus are not expected to fully correlate with the DSSP-assigned structures.

Further patterns can be found in the two proteins shown in Figures 4 and 5. Stretches of SBB-α residues (helices) often begin with one or two SBB-ζ residues and end with one or two SBB-η residues. Similarly, runs of SBB-β residues (strands) frequently begin and end with SBB-τ and SBB-ι residues, respectively. Thus, SBBs classify separate structural categories for the N- and C-terminal caps of both helices and strands.

The patterns seen in sulfate binding protein and β-lactamase are consistent with those observed throughout the database. Figure 6 shows that SBB-α and -β contain the largest fraction of residues that are classified as helix and strand by the DSSP algorithm, respectively, while all positions of all SBBs contain residues classified as "coil." Evidence for the capping structures is also apparent in this figure. SBB-η contains a substantial amount of DSSP-classified helix in the first four (N-terminal) residues of the seven-residue strand, while SBB-ζ contains a substantial amount of helix in the last four (C-terminal) residues. Furthermore, SBB-ι and SBB-τ contain a significant amount of strand at the first and last four residues of the seven-residue segment, respectively. However, other than the correlations between helix and SBB-α and strand and SBB-β, no strong correlations were found between any of the six SBB patterns and any of the DSSP categories (data not shown). Similar patterns were also observed in the smaller dataset.[37]

These results demonstrate that both helix and strand capping structures can be objectively recognized by their local α-carbon geometry. Our classification system of an autoANN and clustering algorithm is a unique example of a method that objectively recognizes these structures. However, all SBB structures, α, β, η, τ, ζ, and ι, are also found in the nonregular structures or the "random coil" regions (Figs. 5 and 6), demonstrating that segments with these local geometries are not solely found in the regular secondary or capping structures. Analysis of the patterns in the random coil regions is presented below.

TABLE IIIa. Virtual Bond Angle Means and Standard Deviations

	Central α-carbon of angle					
SBB	2	3	4	5	6	All positions
α	92.5(5.0)	92.3(4.2)	92.4(4.9)	92.7(4.9)	94.4(8.3)	92.8(5.7)
β	119.3(14.4)	120.3(13.4)	121.0(13.4)	121.4(13.0)	116.9(15.7)	119.8(14.1)
ζ	116.2(15.0)	110.3(16.5)	95.4(9.7)	96.2(10.6)	102.0(15.0)	104.0(15.9)
η	101.5(15.0)	96.4(10.4)	100.3(12.3)	112.9(15.8)	112.2(15.8)	104.6(15.5)
τ	98.0(10.9)	107.8(17.0)	116.8(15.1)	118.4(14.0)	114.7(15.9)	111.1(16.5)
ι	115.1(16.1)	115.8(15.4)	118.2(13.9)	100.0(13.5)	102.0(16.7)	110.2(17.0)

TABLE IIIb. Virtual Dihedral Angle Means and Standard Deviations

	Central α-carbon pair of dihedral angle				
SBB	2–3	3–4	4–5	5–6	All positions
α	52.4(17.5)	52.8(24.3)	50.9(22.9)	45.7(37.5)	50.5(26.8)
β	−63.2(129.4)	−72.7(120.2)	−71.2(121.9)	−74.5(121.4)	−70.4(123.3)
ζ	−57.8(112.9)	−32.9(97.6)	44.3(32.1)	39.3(75.3)	−1.8(96.0)
η	15.0(81.6)	17.7(56.5)	21.7(110.7)	−32.8(124.6)	5.4(99.5)
τ	23.8(63.5)	15.0(129.0)	−50.9(119.5)	−67.7(115.7)	−20.0(117.0)
ι	−41.3(125.6)	−42.8(120.6)	−71.3(104.1)	37.6(39.0)	−29.8(110.9)

Structure of Each Class of SBBs

The distinct geometry of each SBB category is demonstrated by comparison of the virtual bond angles and virtual dihedral angles (Table III), but is more easily represented by superposition of all segments classified in each SBB category in sulfate binding protein (Fig. 7). Note that only α-carbon geometry was encoded as input into the autoANN, but the complete backbone, N, Cα, C, and O, are shown in this figure; despite this, the general structural cohesiveness of each SBB category is still quite evident. SBB-α is clearly helical, while SBB-β is extended. SBB-η can be described as a fiddlehead, while SBB-ι resembles a shrimp and SBB-ζ resembles a nose. (Except for α and β, each has some similarity to the shape of the Greek letter chosen to represent that category.) These observations demonstrate that the neural network can meaningfully reduce the number of parameters needed to represent the geometry of each seven-residue protein segment and that these vectors can be clustered.

As expected, the structures do not superimpose perfectly (Fig. 7) and the standard deviations of the virtual bond angles and dihedral angles can be rather large (Table III). There are two reasons for this. First, only α-carbons were used to describe segment geometry. More coherent categories might be achievable if all backbone atoms (including the carbonyl oxygen and the β-carbon of each residue) are encoded as input to the autoANN. Second, protein structure is not definable as six precise structural categories. As these are general classifications, there will always be a "fuzziness" to them.

Amino Acid Residue Frequencies in the SBBs

The normalized frequency of occurrence, *f*, of each amino acid at each position in the six SBBs was calculated (Fig. 8). Clear amino acid preferences at specific positions exist. The preferences for the helical SBB-α and helix caps, SBB-ζ and SBB-η, are consistent with the known amino acid preferences.[20,21,48–53] For instance, proline and glycine are known as helix breakers and have a very low frequency of occurrence in most positions in SBB-α. Serine and threonine are also found infrequently in SBB-α. Consistent with helix dipole preferences, lysine and arginine are favored at the C-termini of SBB-α, while glutamic acid (but not aspartic acid) is found in the N-terminus of these segments. The hydrophobic amino acids alanine, leucine, and methionine are preferred in SBB-α segments.

In SBB-ζ, the helix N-capping structure, proline is strongly preferred at positions 3 and 4 and disfavored at positions 5 and 6. Consistent with the previously characterized helix capping box,[54,55] serine is preferred at positions 2, 3, and 4, while glutamic acid is preferred at positions 4 and 5 of SBB-ζ. Valine, leucine, and isoleucine are all disfavored at positions 3, 4, and 5 in this structure. In SBB-η, the helix C-capping structure, glycine is strongly favored in positions 4 and 5, consistent with previously proposed capping structures.[56,57] Again, isoleucine, leucine, and valine are disfavored in positions 3 and 4 of this capping structure.

SBB-β, the extended or β strand structure, and SBB-τ and SBB-ι, often found at the N-termini and C-termini of these strands, respectively, also exhibit amino acid preferences (Fig. 8). The β-branched residues isoleucine, threonine and valine are preferred at the central positions of SBB-β segments. The negatively charged residues, aspartic acid and glutamic acid, are not commonly found in the central positions of SBB-β. Proline and serine are found at the C-termini and glycine is found at the N-termini of SBB-β segments. In SBB-τ, the strand N-cap structure, valine, leucine, and isoleucine are disfavored at positions 2 and 3. Proline is favored in

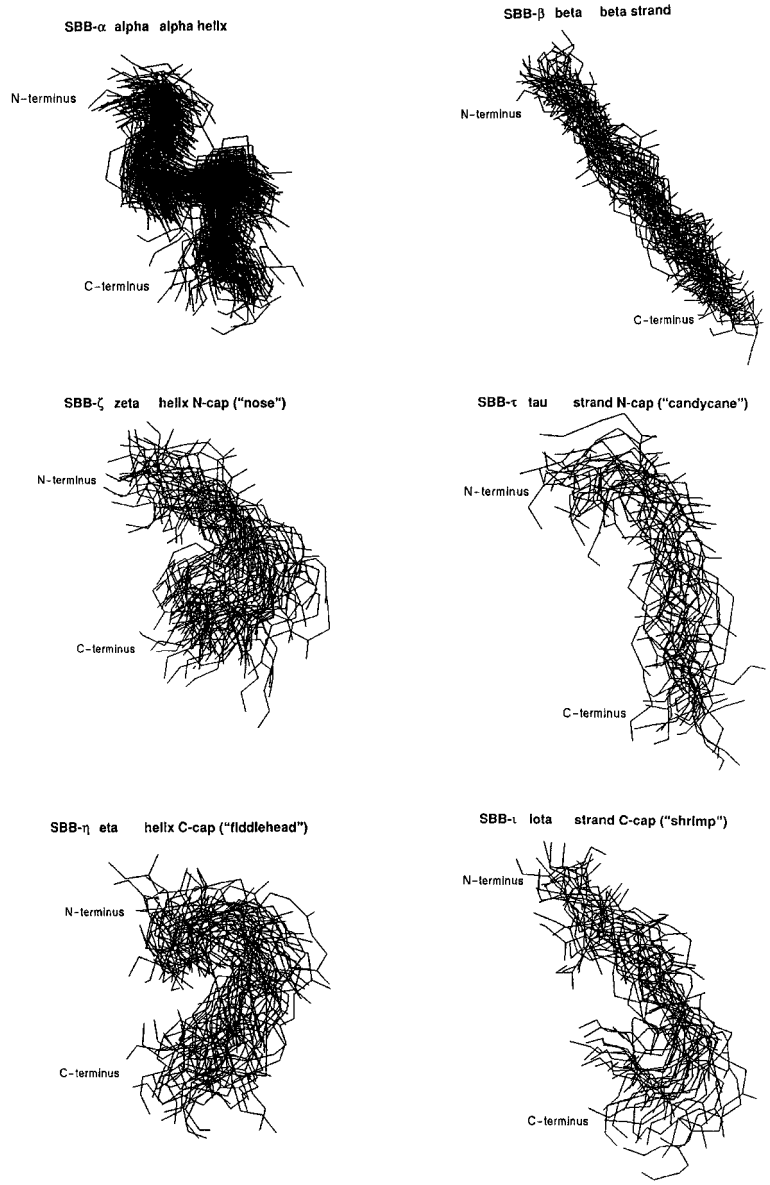

Fig. 7. The structures of each of the six SBBs. All segments belonging to each SBB from sulfate binding protein were superimposed. In sulfate binding protein, 116, 57, 41, 38, 24, and 27 segments belonging to SBB categories α, β, ζ, η, τ, and ι, respectively, were found. As described, only the α-carbons of each protein segment were used to describe the geometry to the neural network, however, the complete backbone (N, Cα, C, O) is shown in these superpositions.

positions 1, 2, 5, 6, and 7, but strongly disfavored in position 3. Glycine and asparagine are strongly favored in positions 2 and 3. Colloc'h and Cohen[58] found proline to be disfavored in their strand N-capping structures. These researchers also found a strong preference for charged residues at the N-termini of parallel β strands that is not observed in any position of SBB-τ, except for aspartic acid at position 2; however, the data calculated by Colloc'h and Cohen were for β strand ends in a small set of parallel β sheet proteins,[58] while our data were collected on a large general database of proteins. In SBB-ι, the strand C-cap structure, proline is strongly preferred at position 5 and somewhat less so at position 4. Interestingly, asparagine and aspartic acid are preferred at positions 4 and 6 and serine is preferred at positions 4, 5, and 6, suggesting the possibility of specific capping structures, similar to the helix capping box previously described.[54,55]

These data show that the SBBs exhibit both amino acid preferences consistent with previously published data, as well as novel amino acid preferences not previously recognized. The amino acid preferences are not as strong as those found for specific structures, but our database is not limited to a specific type of protein structure, nor have we imposed visual examination or researcher bias in the selection of structures. The distinct positional preferences of some residues suggest that interactions important for protein folding and structure can be found in these structures.

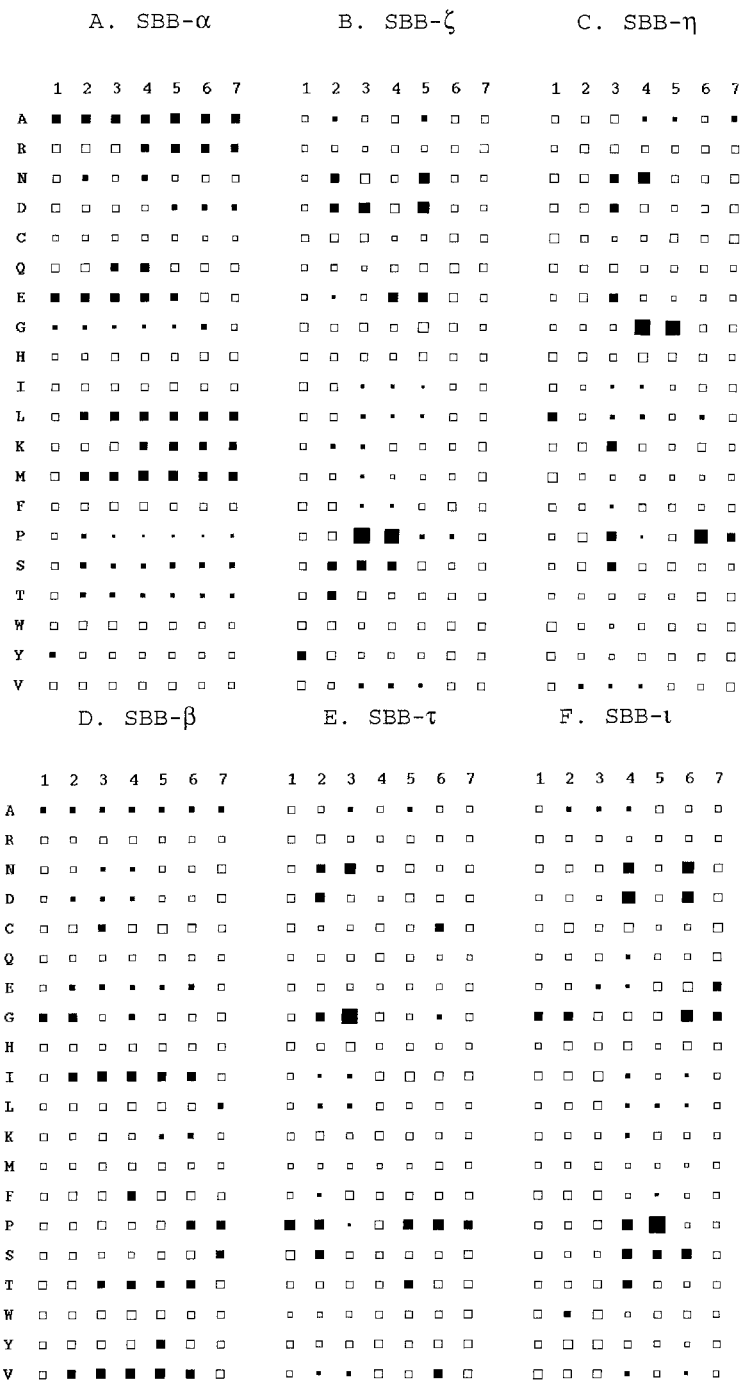

Fig. 8. Hinton diagrams representing the normalized frequency of amino acid occurrence and its statistical significance at each of the seven positions in the six SBBs. **A:** SBB-α. **B:** SBB-ζ. **C:** SBB-η. **D:** SBB-β. **E:** SBB-τ. **F:** SBB-ι. The amino acids (represented by their one-letter codes) are plotted on the *y*-axis and the position within each segment (1–7) is plotted on the *x*-axis. The size of the square is proportional to the normalized frequency of occurrence for each amino acid at each position. Black squares are those frequencies that were determined to be statistically significant at the 0.95 confidence level, as described in the Methods section.

Recurring Patterns in the "Random Coil" Regions

Closer examination of Figures 4 and 5 suggests that patterns of SBBs are also found in the coil regions of proteins. For example, two easily observable patterns are SBB-ι followed by SBB-ζ, together forming an "S-type" turn, and SBB-ζ followed by SBB-η, comprising a "tight" turn. Although these structures are found frequently in sulfate binding protein and β-lactamase (Fig. 4) and in other proteins, the DSSP algorithm does not consistently classify them (data not shown).

This observation suggests that other significant patterns in the nonregular secondary structure re-

TABLE IV. Statistically Significant Occurrence and Structural Characterization of SBB Adjacent Pairs

Pattern	Actual	Expected	Signif-icance[1]	Ramachandran region sequence[2]		f_4[3]	f_2[4]
αα	27.13%	9.71%	+	A A A A	(A A A E)	0.95 (0.01)	0.98
ββ	16.90	5.74	+	E E E E	(E A E E)	0.73 (0.05)	0.84
ιζ	7.39	1.30	+	E E A A	(E E E A*)	0.43 (0.09)	0.66
ητ	6.43	1.20	+	A A E E	(A A*E E)	0.22 (0.11)	0.32
τβ	5.51	2.56	+	E E E E	(A*E E E)	0.35 (0.22)	0.72
βι	5.32	2.43	+	E E E A	(E E E E)	0.48 (0.18)	0.78
ζη	4.52	1.44	+	E A A E[5]	(A A A E)	0.17 (0.12)	0.50
ζζ	4.26	1.65	+	E A A A	(E E A*A*)	0.71 (0.05)	0.80
ζα	3.42	4.00	−	A A A A	(E*A A A)	0.78 (0.03)	0.92
αη	3.20	3.48		A A A E[5]	(A A A A*)	0.40 (0.34)	0.91
ηη	2.87	1.25	+	A A A*E	(A A E A)	0.29 (0.13)	0.40
τι	2.66	1.09	+	E E E A[5]	(A E E A)	0.23 (0.16)	0.68
ττ	1.95	1.14	+	A E E E	(A E A E)	0.41 (0.11)	0.59
βτ	1.68	2.56	−	A E E E	(E A*E E)	0.59 (0.09)	0.73
ιβ	1.24	2.43	−	E E A E	(E E E A*)	0.54 (0.12)	0.64
ηι	0.94	1.14	−	A E E A[5]	(A A E A)	0.17 (0.15)	0.25
ιι	0.78	1.03	−	E E E A*[5]	(E E E E*)	0.13 (0.12)	0.43
αζ	0.67	4.00	−	E A A A[5]	(A E A A)	0.33 (0.28)	0.39
ηα	0.55	3.49	−	A A E A	(A E A A)	0.42 (0.22)	0.46
ζι	0.44	1.30	−	E A E A	(E A E E)	0.66 (0.14)	0.85
τη	0.40	1.20	−	A E A*E[5]	(E A A E)[7]	0.20 (0.20)	0.34
ιτ	0.39	1.09	−	A E A E	(E E A E)	0.33 (0.10)	0.65
ηζ	0.30	1.44	−	A E A A	(A A*A A)	0.31 (0.13)	0.41
ια	0.19	3.16	−	A E A A	(E A*A A)	0.45 (0.21)	0.52
ιη	0.18	1.14	−	A*A A*E[5,6]	(A E A E)	0.22 (0.12)	0.29
τζ	0.18	1.37	−	A E A A[5,6]	(A E A E)	0.17 (0.15)	0.47
ατ	0.14	3.33	−	A A E A*	(A*A E A*)	0.42 (0.16)	0.71
ζτ	0.11	1.37	−	E A*A E[6]	(A A E E)	0.40 (0.12)	0.48
ζβ	0.11	3.08	−	E A E E	(E E A*E)	0.62 (0.17)	0.67
ηβ	0.08	2.68	−	E E*E E[5,6]	(E A E E)[7]	0.24 (0.24)	0.35
βη	0.04	2.68	−	E A E E[6]	(E A E E*)	0.62 (0.12)	0.75
βζ	0.01	3.08	−	E E A E[5,6]	(E E E E)[7]	0.50 (0.50)	0.50
αι	0.01	3.16	−	A A A E[6,8]		1.00	1.00
αβ	0.00	7.47	−				
βα	0.00	7.47	−				
τα	0.00	3.33	−				

[1]A "+" indicates a pair that occurred significantly more often than expected by the combination of the independent occurrences of the individual SBBs, at a 0.99 confidence level. A "−" indicates a significantly underoccurring pair, using the same criterion.
[2]The most commonly occurring Ramachandran region sequence (Fig. 3) for four residues is listed first, followed by the next most common sequence in parentheses.
[3]For the most frequently occurring four-residue Ramachandran region sequence, the frequency of its occurrence among the pair's instances is listed. The frequency of the next most common sequence is given in parentheses.
[4]For the two middle residues of the pair, the frequency of the Ramachandran region sequence among the pair's instances is listed.
[5]For this pair, the difference between the frequencies of occurrence of the first and second most common four-residue sequences is 0.10 or less.
[6]There were ten or fewer instances of the most commonly occurring sequence in the database for this pair.
[7]There were the same number of instances of the two Ramachandran region sequences in the database. Where there was a difference, the sequence with the middle two regions that matched the most frequently occurring two-residue Ramachandran region sequence is listed first.
[8]There was only one instance of this pair in the database.

gions might be found and that these patterns might result from specific residue-residue interactions that could be classified. Thus, patterns of all possible consecutive SBB pairs, triples, and quadruples were analyzed for frequency and statistical significance in this database of 116 proteins. Table IV presents the percentage of occurrence of all consecutive SBB pairs. As indicated, all but one pair was found significantly either more often or less often than expected based on chance alone, at a confidence level of 0.99. As expected, the pairs αα and ββ, corresponding to successive segments (Fig. 2) of helix or strand, are found most often, 6116 times (27.13%) and 3810 times (16.9%), respectively (Table IV). The two patterns mentioned above that are commonly found in coil regions, ιζ (an "S-type" turn) and ζη (a "tight" turn), occur more often than expected, with pattern ιζ occurring more often than the capping patterns τβ and βι (Table IV). Further, two of the capping pairs, τβ, and βι occur more often than expected, at a

TABLE V. Structural Characterization of Statistically Significant Over Occurring SBB Triples

Pattern	Actual	Expected	Ramachandran region sequence[1]	f_5[2]	f_3[3]
ααα	23.98%	3.03%	A A A A A (A A A A E)	0.94 (0.01)	0.97
βββ	11.96	1.38	E E E E E (E E E E A)	0.72 (0.04)	0.84
τββ	4.01	0.61	E E E E E (A*E E E E)	0.36 (0.25)	0.77
ββι	3.97	0.58	E E E E A (E E E E E)	0.51 (0.19)	0.78
ιζζ	3.87	0.17	E E A A A (A E A A A)	0.55 (0.13)	0.76
βιζ	3.75	0.31	E E E A A (E E E E A*)	0.44 (0.13)	0.61
ζαα	3.09	1.25	A A A A A (A*A A A A)	0.79 (0.03)	0.93
ζητ	3.07	0.15	E A A E E[4] (A A A E E)	0.14 (0.11)	0.33
ητβ	2.92	0.29	A A*E E E[4] (E A*E E E)	0.16 (0.13)	0.38
ααη	2.89	1.09	A A A A E[4] (A A A A A*)	0.39 (0.36)	0.89
ζζα	2.83	0.51	E A A A A (E*A A A A)	0.81 (0.04)	0.88
ιζη	2.75	0.15	E E A A E[4] (E E E A*E)	0.18 (0.12)	0.40
τιζ	2.00	0.14	E E E A A[4] (A E E A A)	0.21 (0.18)	0.62
ηητ	1.86	0.13	A A A*E E (A A E E E)	0.34 (0.08)	0.39
αηη	1.70	0.39	A A A A*E (A A A E A)	0.31 (0.15)	0.45
ηττ	1.61	0.13	A A E E E (A A E A E)	0.35 (0.12)	0.47
ητι	1.52	0.12	A A E E A (A E E E A)	0.23 (0.09)	0.31
ζζη	1.31	0.18	E A A A E (E E A*A*E)	0.34 (0.15)	0.56
βτβ	1.21	0.61	A E E E E (E A*E E E)	0.60 (0.08)	0.74
ττβ	1.16	0.27	A E E E E (A E A E E)	0.48 (0.14)	0.64
αητ	1.11	0.37	A A A E E (A A A A*E)	0.41 (0.24)	0.56
ζηη	0.98	0.16	E A A A*E (A A A E E)	0.27 (0.06)	0.39
τβι	0.97	0.26	E E E E A (A*E E E A)	0.25 (0.13)	0.63
ββτ	0.95	0.61	E A E E E (E E A*E E)	0.60 (0.10)	0.64
βιβ	0.84	0.58	E E E A E (E E E E A*)	0.48 (0.14)	0.58
ηιζ	0.78	0.15	A E E A A[4] (A A E A A)	0.18 (0.15)	0.26
ιββ	0.78	0.58	E E A E E (E E E A*E)	0.56 (0.12)	0.66
ττι	0.61	0.12	A E E E A[4] (A*E E E A)	0.25 (0.15)	0.49
ιιζ	0.60	0.13	E E E A*A*[4] (E E E*A A)	0.09 (0.08)	0.16[5]
βιι	0.57	0.25	E E E E E*[4] (E E E E A*)	0.16 (0.13)	0.40
τβτ	0.52	0.27	E A E E E (E E A*E E)	0.58 (0.06)	0.67
ηηι	0.41	0.13	A A A*E A[4] (A A E E A)	0.25 (0.16)	0.27
ιζι	0.40	0.13	E E A E A (E E A E E)	0.58 (0.13)	0.84
βτι	0.40	0.26	A E E E A (E A*E E A)	0.36 (0.09)	0.60
ιβι	0.36	0.25	E E A E A (E E E A*A)	0.46 (0.11)	0.59
ζηι	0.30	0.15	E A A E A[4] (E E A*E A)	0.18 (0.10)	0.31
τητ	0.29	0.13	A E A*E E[4] (E A A E E)	0.25 (0.18)	0.40
ζιζ	0.26	0.17	E A E A A (E A E E A*)	0.58 (0.15)	0.69
ηηζ	0.25	0.16	A A E A A (A A A*A A)	0.34 (0.12)	0.43
ητη	0.22	0.13	A E A A E (A E E A*E)	0.24 (0.10)	0.27

[1]The most commonly occurring Ramachandran region sequence (Fig. 3) for five residues is listed first, followed by the next most common sequence in parentheses.
[2]For the most frequently occurring five-residue Ramachandran region sequence, the frequency of its occurrence among the triple's instances is listed. The frequency of the next most common sequence is given in parentheses.
[3]For the three middle residues of the triple, the frequency of the Ramachandran region sequence among the triple's instances is listed.
[4]For this triple, the difference between the frequencies of occurrence of the first and second most common five-residue sequences is 0.10 or less.
[5]For the SBB triple ιιζ the most frequently occurring Ramachandran region sequence of the three middle residues is EEE*, which does not match the middle three residues of the most common five-residue Ramachandran region sequence. That sequence, EEA*, is the second most frequently occurring, with $f_3 = 0.15$.

confidence level of 0.99. The helix capping pattern αη is found in about the expected numbers. The final capping pattern, ζα, is found less often than expected (at a confidence level of 0.99), but just barely (Table IV). In addition, the pairs ητ, ζζ, ηη, τι, and ττ occur more frequently than expected. Strikingly, the patterns αβ, βα, and τα never occur in the database. Given the number of occurrences of these SBB categories and the number of patterns in the database, 1684 occurrences each of αβ and βα and 752 occurrences of τα are expected. The patterns βζ, βη, and αι occur rarely.

SBB triples, three overlapping seven-residue segments (Fig. 2), were also analyzed. Forty SBB triples occur more often than expected at a 0.99 confidence level (Table V). Examples of their structures can be seen on close inspection of Figure 4. Again, the most common triples are ααα and βββ, corresponding to runs of helix and strand, respectively. All eight helix

TABLE VI. SBB and DSSP Comparison of Helix Ncaps

Protein	Helix bounds[1]	Sequence[2](N'''' to N6)	SBB/DSSP[3]		
Observed helix Ncaps[4]					
3grs	383–392	TVGL TEDE AIH	bbbi	izza	aaa
			EEE-	-HHH	HHH
2cts	37–43	VGQI TVDM MYG	bbbb	izza	aah
			----	-HHH	HHT
2lhb[5]	12–29	VAPL SAAE KTK	bbbb	izza	aaa
			----	-HHH	HHH
2ca2[5]	219–227	PISV SSEQ VLK	tbbb	izza	aaa
			-EEE	-HHH	HHH
3rnt[6]	12–30	SNCY SSSD VST	htbb	izza	aaa
			TEEE	-HHH	HHH
4cpa[5,7]	14–29	ATYH TLDE IYD	httb	izza	aaa
			SS--	-HHH	HHH
1bp2[5]	89–108	SSEN NACE AFI	izht	izza	aaa
			-TT-	-HHH	HHH
2cts	70–78	FRGY SIPE CQK	zzht	izza	aaa
			SSS-	BHHH	HHH
3grs	456–462	KMGA TKAD FDN	ahtt	izza	aah
			HTT-	BHHH	HHT

[1]First and last residues of complete helix. To be consistent with Seale et al.,[55] helix numbering begins with the Ncap residue.
[2]Amino acid sequences are given in their standard one-letter code and follow the nomenclature of Presta and Rose.[49] Here we show residues N''''-N'''-N''-N' Ncap-N1-N2-N3 N4-N5-N6. To facilitate alignment, a space precedes the Ncap residue and follows the N3 residue; the Ncap residue is shown in bold type.
[3]The top row is the SBB designation: SBB-α, a; SBB-β, b; SBB-ζ, z, SBB-η, h; SBB-τ, t; SBB-ι, i. The bottom row is the DSSP [15] designation: H, helix; E, sheet; T, turn; S, bend; B, β-bridge.
[4]Capping boxes contain reciprocal hydrogen bonds from the backbone of N3 to the side chain of Ncap and from the backbone of Ncap to the side chain of N3 as described.[54,55]
[5]Protein not in the training dataset described here; SBB assignment made based on categorizations described here.
[6]Compared to 9rnt from Seale et al.[55]
[7]Compared to 5cpa from Seale et al.[55]

and strand capping triples, ζαα, ζζα, ααη, αηη, τββ, ττβ, ββι, and βιι, are found more often than statistically expected. As with pairs of SBB categories, given the number of occurrences of the SBB categories and the number of patterns in the database some triples are expected to be present in the dataset, but are not. Seventy such triples are expected, but did not occur (data not shown), again showing that, as expected, successive SBB segments are highly correlated.

Some consecutive SBB quadruples are present in statistically significant quantities and a few of these will be discussed further below. Most consecutive SBB quadruples and quintuples could not be analyzed for statistical significance because of the size of the database.

These data demonstrate that consecutive SBB occurrences are highly correlated. This result is expected because consecutive residue conformations in proteins should be correlated and because consecutive SBB segments are seven residues long and overlap (Fig. 2). However, coil regions have recurring motifs that are difficult to classify, and SBB patterns and SBBs themselves provide a convenient, objective method for classification of the random coil regions in proteins. These patterns could potentially provide a wealth of information about structurally important interactions within proteins, similar to the hydrogen bonds and hydrophobic interactions recently observed in helix capping structures.[54–56]

Comparison of SBBs to Previously Observed Helix Caps and Other Local Protein Structures

To demonstrate the utility of using the SBB classifications for discovering new structure motifs in proteins, a comparison of the SBB helix capping structures to previously published capping structures is presented. The N-terminal helix capping box is a structure defined by two reciprocal hydrogen bonds, one from the backbone of the first helical residue, Ncap to the side chain of the third helical residue, N3, and one from the side chain of Ncap to the backbone of N3. In this structure, the most common amino acids at Ncap are serine and threonine, while the most common amino acid at N3 is glutamic acid.[54,55]

The SBB assignments (this work) and the DSSP secondary structure assignments[15] for the helix N-capping box described by Rose and coworkers[54,55] are reported in Table VI. In a remarkably consistent manner, the Ncap-N1-N2-N3 residues involved in the capping box are always classified as ιζζα. While DSSP consistently determines the ends of these

helices at the N1 position, the Ncap residue is inconsistently classified as B (bridge) or left unclassified by DSSP.

Two distinct structures at the C-termini of α helices that contain a glycine at the C′ residue have also been described.[14,56,57,59,60] The Schellman motif contains two backbone-backbone hydrogen bonds, one from C‘ to C3 and one from C′ to C2 . The α_L structure consists of one backbone-backbone hydrogen bond from C′ to C3. In Table VII, SBB and DSSP assignments are compared to the Schellman and α_L motifs previously described.[56] For the Schellman motif, the C1-Ccap-C′ residues are defined as αηη 10 times and αητ three times. Of the 10 αηη patterns, seven are αηητ, two are αηηι, and one is αηηη. Nine of the ten αηη patterns and all three of the αητ patterns are classified by DSSP as HTT. For two of the Schellman motifs (2cts., 276–292 and 3grs., 383–391), the SBB classification of the helices differs from that of Aurora and colleagues.[56] For these two motif instances, the Aurora group's classification of the helices extends one residue further at its C-terminal end than the SBB helix classifications of the motifs (Table VII). Thus, the SBB classification for these two helix C-capping motif instances is still αηη, but it is shifted by one residue compared to the classification scheme of Aurora and colleagues. These two Schellman motifs were identified by Aurora and colleagues as having high temperature factors, which may account for the offset.

The α_L motif is defined as αητ four times and αηη one time. Of the four αητ patterns, three are αηττ and one is αητβ. These α_L motifs are not consistently classified by DSSP. Thus, while our pattern recognition algorithm does not perfectly discriminate between these Schellman and α_L motifs, it does recognize motifs at the C-termini of helices. Given that the auto-ANN was only presented with the α-carbon geometry of the protein segments, it is astonishing that SBBs can discriminate at this level of structure.

Can we discern the previously described amino acid preferences from the SBB patterns? Further analysis of the 116 protein chains in our training database shows that there are 576 instances of the pattern ιζζα, which corresponds to the helix N-capping box. Of these 576 occurrences, 480 (83%) are followed by at least two additional α (α-helical) residues. Analysis of these 480 instances of ιζζα at the N-terminus of a helical segment shows that serine is found at the Ncap (corresponding to the central residue of SBB-ι in the ιζζα quadruple) position 81 times (16.9%) and threonine is found at this position 75 times (15.7%). Similarly, glutamic acid is found at the N3 (α) position 78 times (16.2%). Interestingly, at the Ncap position, aspartic acid was found 74 times (15.4%), asparagine was found 49 times (9.6%), and proline was found 45 times (9.4%). Besides glutamic acid at the N3 position, aspartic acid was found 46 times (9.6%) and glutamine was found 44 times (9.2%). These are strong amino acid preferences, and they are consistent with the amino acid patterns previously seen in helix N-capping boxes.[54,55]

Analysis of the C-cap structures shows that the pattern αηη occurs 381 times in the 116 protein chains. Of these occurrences, 329 (86%) of them are preceded by at least two SBB-α segments. Glycine is found at the third position of the αηη pattern 130 out of 329 times (39.5%). This site corresponds to the C′ position of the helix and is consistent with the previous identification of glycine at this site.[56] Preferences for alanine and lysine are found at the second position of αηη (the Ccap residue). Lysine, alanine, and glutamic acid are found at the first position of αηη pattern 14.6, 14.3, and 11.9% of the time, respectively, which corresponds to the C1 helix position. Alanine, leucine, lysine, and arginine are preferred at position C2, the residue just before the αηη pattern, consistent with the data previously reported.[56]

It appears that we can easily discern the helix capping structures from the SBB patterns, but that several types of capping structures are described by the patterns. This is expected because only α-carbon coordinates are used in the description of SBBs. In fact, it is amazing that α-carbon patterns alone can provide such useful structural information.

Use of SBB Patterns To Discern Unique Structures and Structural Motifs

To assess whether patterns of successive SBBs represent unique structural motifs, and whether they can be used as "building blocks" for representing protein structure, the backbone geometry of all SBB pairs and those SBB triples that occurred significantly more often than expected in the database was analyzed. The backbone ϕ and ψ angle values of the pairs and triples' instances were used to identify structural regularities. The two most commonly occurring Ramachandran region sequences and their frequencies are given for each pair (Table IV) and triple (Table V). The central two or four residues were used in analyzing the pairs, and the central three or five residues for the triples (Fig. 2).

The analysis showed evidence of structural motifs in the central four residues of at least some SBB pairs. Of the 36 possible SBB pairs, 33 occur in the database. For a majority of the pairs (21 of 33) one or two Ramachandran region sequences comprise over half of the instances of the pair. There is a strong trend toward a single, dominant Ramachandran region sequence-in 22 of the 33 pairs, the most common Ramachandran region sequence occurs over 10% more often than the next most common one. As might be expected, the variability in the Ramachandran region sequences is predominantly in the residues at the ends of the overlapping SBBs making up the pairs. For all of the SBB pairs, the center two residues are the most common Ramachandran region sequence of length two (Table IV). For the pairs the frequency of occurrence of the most common two-residue Ramachandran region sequence is an

TABLE VII. SBB and DSSP Comparison of Helix C-Capping Motifs

Protein	Helix bounds[1]	Sequence[2](C7 to C′′′′)	SBB/DSSP[3]		
Schellman motifs[4]					
1snc	98–106	VNEA LVR**Q** GLAK	zzaa	aaah	httt
			HHHH	HHH**T**	TSSE
4fxn	93–106	FEER MNG**Y** GCVV	aaaa	aaah	htbi
			HHHH	HHH**T**	T-EE
2cts	297–311	YIWN TLN**S** GRVV	aaaa	aaah	htbi
			HHHH	HHH**T**	T---
4fxn	10–26	IAKG IIE**S** GKDV	aaaa	aaah	htbb
			HHHH	HHH**T**	T---
3grs	29–42	SARR AAE**L** GARA	aaaa	aaah	htbb
			HHHH	HHH**T**	T--E
3grs	227–241	CTEE LEN**A** GVEV	aaaa	aaah	htbb
			HHHH	HHH**T**	T-EE
2cts	344–364	VPNV LLE**Q** GKAK	aaaa	aaah	htiz
			HHHH	HHH**T**	T--S
4cpa[5,6]	253–261	SIDW SYN**Q** GIKY	zzaa	aaah	hizh
			HHHH	HHH**H**	T--E
1gd1[6]	101–111	DAAK HLE**A** GAKK	aaaa	aaah	hibb
			HHTH	HHH**T**	T-SE
2cts	393–415	QLIW SRA**L** GFPL	aaaa	aaah	ttbb
			HHHH	HHH**T**	T---
3grs	444–453	GFAV AVK**M** GATK	aaaa	aaah	ttiz
			HHHH	HHH**T**	T-BH
2cts	152–160	NFAR AYA**E** GIHR	zzaa	aaah	ttiz
			HHHH	HHH**T**	T--G
2cts[7]	276–292	LTQL QKE**V** GKDV	aaaa	aahh	tzht
			HHHH	HHH**T**	-SS-
3grs[7]	383–391	EDEA IHK**Y** GIEN	zzaa	aahh	izzh
			HHHH	HHH**H**	-GGG
2cts[8]	37–43	ITVD MMY**G** GMRG	bizz	aaah	hhth
			--HH	HHH**T**	TSTT
α_L motifs[9]					
4cpa[5]	173–186	IVDF VKN**H** GNFK	aaaa	aaah	tttt
			HHHH	HHH**H**	--EE
2cts	89–98	GLFW LLV**T** GQIP	aaaa	aaah	tttb
			HHHH	HHH**H**	SS--
2cts[10]	208–217	SHNF TNM**L** GYTD	zaaa	aaah	ttbi
			HHHH	HHH**H**	T---
3grs	196–209	MAGI LSA**L** GSKT	aaaa	aaah	htbb
			HHHH	HHH**T**	T-EE
3rnt[11]	12–29	AGYK LHE**D** GETV	aaaa	aaah	tbbb
			HHHH	HHH**T**	T--B

[1]Residue number of first residue in the helix and the Ccap residue of the helix.
[2]Amino acid sequences are given in their standard one-letter code and follow the nomenclature of Presta and Rose.[49] Here we show residues C7-C6-C5-C4 C3-C2-C1-Ccap C′-C′′-C′′′-C′′′′. To improve readability, there are spaces preceding C3 and C′ and the Ccap residue is shown in bold type.
[3]The top row is the SBB designation: SBB-α, a; SBB-β, b; SBB-ζ, z, SBB-η, h; SBB-τ, t; SBB-ι, i. The bottom row is the DSSP[15] designation: H, helix; E, sheet; T, turn; S, bend; B, β-bridge.
[4]The Schellman motif is defined by backbone-backbone hydrogen bonds between residues C2 and C′ and residues C3 and C′′, as described.[56]
[5]Compared to 5cpa.
[6]Protein not in the training dataset described here; SBB assignment made based on categorizations described here.
[7]High temperature factors (B factor > 50) for some atoms in the motif (marked by Aurora et al.[56]).
[8]The residue numbering and amino acid sequence between Table 1 of Aurora et al.[56] and the Brookhaven File are inconsistent. Our analysis of hydrogen bonds suggests that the numbering of Aurora et al. (37–43)[56] is correct and we have adjusted the amino acid sequence accordingly.
[9]α_L motifs are defined by a hydrogen bond between the NH at C′ and the C=0 at C3, as described byAurora, Srinivasan and Rose.[56]
[10]Extra H bond from C′′ to C3 acceptor (marked by Aurora, et al.[56]).
[11]Compared to 9rnt.

average of 0.17 higher than that of the highest frequency four-residue one, strongly indicating that a coherent structure underlies many of the pairs, especially among the central residues.

The regular secondary structures and their caps are easily recognized in the analysis of the pairs. All four positions in the most common Ramachandran region sequence for the pair αα occupy the Ramachandran helix region, A (Fig. 3), as do those of the helix N-cap ζα. The helix C-cap αη occupies the helical region in all but the final (C-terminal) position. Analogous results are found for the strand conformation pair ββ and its N- and C-caps, τβ and βι. The structure for the ιζ "S-type" turn suggests a transition from helical to extended conformation, and that of ζη, the "tight" turn, a turn connecting extendedconformations. The repeating pairs ηη and ιι both have a residue in a left-handed helical conformation (region A*). The conformations of the remaining pairs that occur at significantly higher than expected frequencies, ζζ, ττ, ητ and τι, suggest transitions from one allowed region of the plot to another.

Eight Ramachandran region sequences are shared by two pairs each (αα and ζα, ββ and τβ, βι and τι, ζζ and αζ, ττ and βτ, αη and αι, ιβ and βζ, ζβ and βη) and one is shared by three pairs (ηζ, ια and τζ). In six of the cases where a Ramachandran region sequence is shared by two pairs, the two pairs have a common SBB (e.g. SBB ι in βι and τι).

The Ramachandran region sequence analysis indicates that many of the significantly over occurring SBB triples also represent unique structures (Table V). Pairs and triples share some common characteristics. The most common one or two Ramachandran region sequences comprise over half of the instances of that triple for 19 of 40 patterns. In 28 of the 40 triples, the most common Ramachandran region sequence occurs over 10% more often than the next most common one. The variability in the triples' Ramachandran region sequences is concentrated at the ends of the triples; for 39 of the 40 triples, the Ramachandran region sequence of the center three residues of the five-residue sequence is also the most common three-residue Ramachandran region sequence (Table V). For these 39 triples the frequency of occurrence of the most common three-residue Ramachandran region sequence is an average of 0.17 higher than the highest frequency five-residue Ramachandran region sequence. Just as for the pairs, this is evidence that many of the triples represent structural motifs.

The analysis of Ramachandran region sequences suggests that in many cases the triples represent unique structures. As expected, the regular secondary structure triples ααα and βββ occupy the helix and strand regions, respectively, at very high frequencies. The Ramachandran region sequences for triples marking the ends of regular secondary structures (ααη, ζαα, ββι, and τββ) are in the conformations of the corresponding secondary structure at very high frequency in all positions. Two of the triples representing C-caps (ααη, ββι) vary in the C-terminal position of the Ramachandran region sequence. Other patterns suggest interruptions in extended conformation regions (ββτ, βιβ, ιββ, and τβτ). Of particular interest are those triples, especially αηη, ηητ, and ζηη with their highly preferred Ramachandran region sequences, that contain regions of left handed helix. As described earlier, αηη is the SBB classification for most of the Schellman motif instances and some of the α_L motif instances. It may be the case that the other two triples, ηητ and ζηη, represent previously unrecognized motifs in coil regions. Overall, SBB triples appear to represent unique structural motifs, more so than the SBB pairs, but it must be kept in mind that *all* SBB pairs were analyzed, whereas only those SBB triples that occur significantly more often than expected were analyzed here.

Eight Ramachandran region sequences are shared by two triples each (ααα and ζαα, βββ and τββ, ββι and τβι, τιζ and βιζ, βτβ and ττβ, ββτ and τβτ, ττι and βτι, ιζι and ιβι). In each case the two triples that have the same most common Ramachandran region sequence also share two common SBBs. That different but related pairs and triples occupy the same Ramachandran region sequences is not surprising. The SBB categories are based solely on α-carbon geometries. The Ramachandran region sequence uses information about the full backbone. Also, the division of the Ramachandran plot used (Fig. 3) is a very coarse-grained classification of backbone conformations. Thus, it is not the occurrence of shared Ramachandran sequences that is surprising, but rather that the Ramachandran region sequence analysis distinguishes so many unique structures, despite the α-carbon-only representation used to create the SBBs and the coarse Ramachandran regions used.

Can the SBB patterns be used to find new motifs in protein structures? Once an SBB pattern is observed, can the amino acid pairwise correlations be used to discriminate among the various local structures that are described by this pattern? While such analysis is beyond the scope of this paper, the various results shown here suggest that unique structures and important residue-residue interactions might be uncovered by analyzing the local structures found in the statistically significant SBB patterns and demonstrate the biological relevance of the SBB local structure assignments.

Dissemination of Results

In order to promote the widest possible dissemination of this work, much of the data reported here has been made available on the World Wide Web. The URL http://barbara.bio.albany.edu/compbio links to a site containing the SBB classification data for all 116 protein chains in the database used for this paper. For each SBB in each protein, the SBB category and the sequence identification and amino acid for each of the seven residues in the SBB are given. In addition, the SBB cluster analysis data for

all pairs and significantly over occurring triples is available from this site.

DISCUSSION

The combination of an autoANN and a clustering algorithm is a novel method to automatically, consistently, and objectively classify *all* residues in a protein into local structural categories without visual examination or researcher bias in parameter selection. By using an artificial neural network to reduce the dimensionality of the input geometry and to represent the geometry descriptions consistently, we can then use a simple clustering algorithm to automatically identify structural categories in the "coil" regions of the protein, as well as in the regular structures. Two of the SBB categories, α and β, correspond closely to the classical α helix and β strand secondary structures and exhibit amino acid preferences consistent with these structures. The remaining four categories, SBB-ζ, η, τ, and ι, are consistently found at N- and C-termini of helices and strands, respectively, and are also found in regions typically identified as "loops" or "random coil." Within the "coil" regions, patterns of SBBs are found. This algorithm identifies the SBBs from atomic coordinates, so that SBBs can be identified for any globular protein whose structure has been solved.

The differences between the SBB local structure classifications described here and the classical helix, strand, and coil structures should be emphasized. One major difference is that helix, strand, and coil assignments are usually made on a residue-by-residue basis—each residue is given a single secondary structure classification. SBB classifications, on the other hand, are applied to seven-residue segments. Therefore, each residue in the protein (except for the first three and the last three) is a member of seven different SBB categories (Fig. 2). Five- and six-residue segments with the same amino acid sequence can have different conformations in the context of the protein tertiary structure,[61,62] and we hoped the conformation of the longer segments would be more uniquely determined by the amino acid sequence.

The second difference between SBBs and classical secondary structures is the structure assignment method. For classification into classical secondary structures, the definition of each category is provided (e.g., as hydrogen-bonding patterns or backbone dihedral angles) and each residue is checked to determine whether or not it fits into any category. Thus, some residues remain unclassified. In the SBB classification scheme, all seven-residue segments are necessarily assigned to a cluster, corresponding to an SBB category. No segments are unclassified. Six categories were chosen because the data clustered into six groups most consistently.

The third difference between classical secondary structures (especially those assigned by the DSSP algorithm[15]) and SBBs is in the tertiary, hydrogen-bonding information used in the assignments. The DSSP algorithm uses hydrogen bond information in secondary structure assignment, thus it does not find single stretches of extended strand. On the other hand, SBB assignments are done purely by local α-carbon geometry. Thus, local stretches of strand are classified as SBB-β, even though they are not necessarily part of a larger hydrogen-bonded sheet. Because of these differences between SBB and classical secondary structure classification systems, quantitative, one-to-one comparisons of the structures cannot be made. Qualitative comparisons (as presented in Figs. 5 and 6 and Tables VI and VII) can be done, but it should be kept in mind that SBB assignments apply to seven-residue segments, not to single residues.

General Applicability of the Method

Is this general strategy of autoANN and clustering algorithm useful for finding other structural motifs in proteins? The data presented on helix capping motifs and backbone angle sequences of SBB pairs and triples would indeed suggest that the six SBBs presented here will lead to the identification of some interesting structural patterns, and we are in the process of analyzing some of those patterns. However, this approach is not limited to seven-residue segments and patterns in longer segments could easily be selected in a similar fashion. The approach is likewise not limited to six clusters or classifications. The data can be analyzed so that the data itself suggests the proper number of clusters. Finally, the approach is not limited to α-carbon geometry. If complete backbone or side chain geometries are encoded as input into the neural network, this same general strategy could be used to search for patterns in these structures, as well.

Comparison of Our Algorithm to Other Local Structure Libraries

Other researchers have used smaller segments and tried clustering on the raw conformation data or on rms differences.[31,32] Clustering algorithms do not work well on the high-dimensional data needed to describe protein segment geometry. Simplified descriptions of the protein backbone geometry have been used to overcome this limitation[31–33]; however, these simplified descriptions are still high-dimensional. One result of clustering on high-dimensional data is a lack of generality, which can produce over 100 building block units.[32] It is advantageous to use an autoANN to compute a suitable representation before clustering. It not only reduces the dimensionality of the data, but also transforms the data (nonlinearly) to make the important information explicit. Furthermore, in our representation, the relative weighting of distance and angle information does not require manual optimization. Other attempts at secondary structure reclassification have required manual optimization of parameters[33] or subjective division of the Ramachandran plot.[31]

Comparison of SBB Structures to Previously Observed Local Structures

Because the hidden layer of the autoANN can be used to distinguish between helices and strands, our original hypotheses that the hidden layer contains a concise encoding of structural features and that clustering these hidden unit vectors yields biologically relevant information are strongly supported. Further, it is significant that the network's hidden layers can distinguish between the structures of the N- and C-terminal caps in helices and strands. The existence of helix capping structures has previously been proposed,[49,50] and in some cases the importance of these structures for protein stabilization has been verified by experiment.[63–69] Structures that act as strand caps or β breakers have also been proposed for a very small set of parallel β sheets.[58,70] Our algorithm is the first to use neural network techniques to automatically and consistently extract this local structural information in proteins. Further, all residues in a protein are assigned to a specific SBB category so that no residues are lumped into a "random coil" or leftover category.

Future Work

Clearly, patterns in SBB structures will be useful for identifying local structural motifs in proteins. For this work to be useful in protein structure prediction, SBB categories must be predictable from the primary amino acid sequence and local protein structure must be reconstructable from the categories. Because SBBs do not represent global interactions, the reconstruction will not be perfect; furthermore, the six SBBs represent an average segment conformation. The distinct amino acid preferences at specific positions of the SBBs, however, suggest that the SBBs might be predictable. Further, each amino acid in the protein (except the first and last three) is a member of seven unique SBBs and this additional positional information for each residue can be used in reconstruction.

The architecture of an autoANN itself suggests a method of reconstruction, in that the second half of the network, from the hidden layer to the output, can act as a "decoder" to regenerate the 43 parameters that specify the geometry of each seven-residue segment, thus we may not have to generalize segment structure into just six SBB categories, but could attempt to predict the hidden unit vector for each segment. Because all parts of the protein (not just the regular secondary structures) are categorized into SBBs, we should be able to reconstruct all parts of the protein, including the loop and nonregular structures. Even if reconstruction is too inexact to produce an entire protein, these structures can be used to produce probable loop structures during model building.

Other patterns in protein structure might be found using the combination of autoANN and clustering algorithm. This approach might be improved by including all backbone atoms and Cβ, rather than just α-carbons, in the geometric representation of each seven-residue segment and using the network and clustering to recompute other structural categories. Significant patterns that are important for understanding protein structure might be found with this more detailed representation.

ACKNOWLEDGMENTS

We thank Trevor Creamer and Reggie Aurora for stimulating discussions and useful comments, Chip Lawrence for help with statistics, Alan Grossfield for suggesting the use of Hinton diagrams in Figure 8, Brian Kell for help with graphical displays and critical reading of the manuscript, and a reviewer for helpful suggestions. This work was supported by NSF grant BIR9211256 to J.S.F and G.B.

REFERENCES

1. Pauling, L., Corey, R.B., Branson, H. The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain. Proc. Natl. Acad. Sci. U.S.A. 37:205–211, 1951.
2. Pauling, L., Corey, R.B. Configurations of polypeptide chains with favored orientations around single bonds: Two new pleated sheets. Proc. Natl. Acad. Sci. U.S.A. 37:729–740, 1951.
3. Venkatachalam, C.M. Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units. Biopolymers 6:1425–1436, 1968.
4. Richardson, J.S. The anatomy and taxonomy of protein structures. Adv. Prot. Chem. 34:167–339, 1981.
5. Rose, G.D., Gierasch, L.M., Smith, J.A. Turns in peptides and proteins. Adv. Prot. Chem. 37:1–109, 1985.
6. Milner-White, E.J., Poet, R. Loops, bulges, turns and hairpins in proteins. Trends Biochem. Sci. 12:189–192, 1987.
7. Sibanda, B.L., Blundell, T.L., Thornton, J.M. Conformation of β hairpins in protein structures: A systematic classification with applications to modeling by homology, electron density fitting and protein engineering. J. Mol. Biol. 206: 759–778, 1989.
8. Leszczynski, J.F., Rose, G.D. Loops in globular proteins: A novel category of protein secondary structure. Science 234:849–855, 1986.
9. Ring, C.S., Kneller, D.G., Langridge, R., Cohen, F.E. Taxonomy and conformational analysis of loops in proteins. J. Mol. Biol. 224:685–699, 1992.
10. Fetrow, J.S. Omega loops: Nonregular secondary structures significant in protein function and stability. FASEB J. 9:708–717, 1995.
11. Efimov, A.V. A novel super-secondary structure of proteins and the relation between the structure and the amino acid sequence. FEBS Lett. 166:33–38, 1984.
12. Edwards, M., Sternberg, M., Thornton, J. Structural and sequence patterns in the loops of β-α-β units. Prot. Eng. 1:173–181, 1987.
13. Efimov, A.V. Structure of coiled beta-beta-hairpins and beta-beta-corners. FEBS Lett. 284:288–292, 1991.
14. Milner-White, E.J. Recurring loop motif in proteins that occurs in right-handed and left-handed forms: its relationship with alpha-helices and beta-bulge loops. J. Mol. Biol. 199:503–511, 1988.
15. Kabsch, W., Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637, 1983.
16. Richards, F.M., Kundrot, C.E. Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure. Proteins 3:71–84, 1988.
17. Levitt, M., Greer, J. Automatic identification of secondary structure in globular proteins. J. Mol. Biol. 114:181–293, 1977.
18. Sklenar, H., Etchebest, C., Lavery, R. Describing protein structure: A general algorithm yielding complete helicoidal parameters and a unique overall axis. Proteins 6:46–60, 1989.
19. Colloc'h, N., Etchebest, C., Thoreau, E., Henrissat, B., Mornon, J.-P. Comparison of three algorithms for the assignment of secondary structure in proteins: The advantages of a consensus assignment. Prot. Eng. 6:377–382, 1993.

20. Chou, P.Y., Fasman, G.D. Prediction of protein conformation. Biochemistry 13:222–245, 1974.
21. Garnier, J., Osguthorpe, D.J., Robson, B. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. J. Mol. Biol. 120:97–120, 1978.
22. Cohen, F.E., Abarbanel, R.M., Kuntz, I.D., Fletterick, R.J. Turn prediction in proteins using a pattern-matching approach. Biochemistry 25:266–275, 1986.
23. Presnell, S.R., Cohen, B.I., Cohen, F.E. A segment-based approach to protein secondary structure prediction. Biochemistry 31:983–993, 1992.
24. Qian, N., Sejnowski, T.J. Predicting the secondary structure of globular proteins using neural network models. J. Mol. Biol. 202:865–884, 1988.
25. Holley, L.H., Karplus, M. Protein secondary structure prediction with a neural network. Proc. Natl. Acad. Sci. U.S.A. 86:152–156, 1989.
26. Kneller, D.G., Cohen, F.E., Langridge, R. Improvements in protein secondary structure prediction by an enhanced neural network. J. Mol. Biol. 214:171–182, 1990.
27. Zhang, X., Mesirov, J.P., Waltz, D.L. Hybrid system for protein secondary structure prediction. J. Mol. Biol. 225: 1049–1063, 1992.
28. Rost, B., Sander, C. Prediction of protein secondary structure at better than 70% accuracy. J. Mol. Biol. 232:584–599, 1993.
29. Yi, T., Lander, E.S. Protein secondary structure prediction using nearest-neighbor methods. J. Mol. Biol. 232:1117–1129, 1993.
30. Salamov, A.A., Solovyev, V.V. Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments. J. Mol. Biol. 247:11–15, 1995.
31. Rooman, M.J., Rodriguez, J., Wodak, S.J. Automatic definition of recurrent local structure motifs in proteins. J. Mol. Biol. 213:327–336, 1990.
32. Unger, R., Harel, D., Wherland, S., Sussman, J. A 3D building blocks approach to analyzing and predicting structure of proteins. Proteins 5:355–373, 1989.
33. Pestrelski, S.J., Williams, A.L., Jr., Liebman, M.N. Generation of a substructure library for the description and classification of protein secondary structure. Proteins 14: 430–439, 1992.
34. Shenkin, P.S., McDonald, D.Q. Cluster analysis of molecular conformations. Comp. Chem. 15:899–916, 1994.
35. Zhang, X., Fetrow, J., Berg, G. Design of an autoassociative neural network with hidden layer activations that were used to reclassify local protein structures. In: "Techniques in Protein Chemistry V." Crabb, J. (ed.). San Diego, CA: Academic Press, 1994:397–404.
36. Rumelhart, D.E., Hinton, G., Williams, R.J. Learning internal representations by error propagation. In: "Parallel Distributed Processing." Rumelhart, D.E., McClelland, J.L. (eds.). Cambridge, MA: MIT Press, 1986:318–362.
37. Zhang, X., Fetrow, J.S., Rennie, W.A., Waltz, D.L., Berg, G. Automatic derivation of substructures yields novel structural building blocks in globular proteins. In: "Proceedings of The First International Conference on Intelligent Systems for Molecular Biology." Washington, DC: AAAI, 1993.
38. Jones, A., Thirup, T. Using known substructures in protein model building and crystallography. EMBO J. 5:819–822, 1986.
39. Bernstein, F.C., Koetzle, T.G., Williams, G.J.B., Meyer, E.F., Brice, M.D., Rogers, J.R., Kennard, O., Shimanouchi, T., Tasumi, M. The protein data bank: A computer-based archival file for macromolecular structures. J. Mol. Biol. 112:535–542, 1977.
40. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J. Basic local alignment search tool. J. Mol. Biol. 215:403–410, 1990.
41. Hobohm, U., Scharf, M., Schneider, R., Sander, C. Selection of a representative set of structures from the Brookhaven protein data bank. Prot. Sci. 1:409–417, 1992.
42. Rumelhart, D.E., McClelland, J.L. "Explorations in Parallel Distributed Processing." Cambridge, MA: MIT Press. 1988.
43. Hartigan, J.A., Wong, M.A. A k-means clustering algorithm. Appl. Stat. 28:100–108, 1975.
44. Zimmerman, S.S., Pottle, M.S., Nemethy, G., Scheraga, H.A. Conformational analysis of the 20 naturally occurring amino acid residues using ECEPP. Macromolecules 10:1–8, 1977.
45. Hertz, J.A., Krogh, A.S., Palmer, R.G. "Introduction to the Theory of Neural Computation." Redwood City, CA: Addison-Wesley. 1991:327.
46. Sack, J.S., Quiocho, F.A. Structure of sulfate binding protein involved in active transport and novel mode of sulfate binding. To be published. 1993.
47. Knox, J.R., Moews, P.C. Beta lactamase of *Bacillus licheniformis* 749/C. Refinement at 2 « resolution and analysis of hydration. J. Mol. Biol. 220:435–455, 1991.
48. Shoemaker, K.R., Kim, P.S., York, E.J., Stewart, J.M., Baldwin, R.L. Tests of the helix dipole model for stabilization of α-helices. Nature 326:563–567, 1987.
49. Presta, L.G., Rose, G.D. Helix signals in proteins. Science 240:1632–1641, 1988.
50. Richardson, J.S., Richardson, D.C. Amino acid preferences for specific locations at the ends of α helices. Science 240:1648–1652, 1988.
51. Fairman, R., Shoemaker, K.R., York, E.J., Stewart, J.M., Baldwin, R.L. Further studies of the helix dipole model: Effects of a free α-NH3+ or α-COO-group on helix stability. Proteins 5:1–7, 1989.
52. Sali, D., Bycroft, M., Fersht, A.R. Stabilization of protein structure by interaction of α helix dipole with a charged side chain. Nature 335:740–743, 1988.
53. Nicholson, H., Becktel, W.J., Matthews, B.W. Enhanced protein thermostability from designed mutations that interact with α-helix dipoles. Nature 336:651–656, 1988.
54. Harper, E.T., Rose, G.D. Helix stop signals in proteins and peptides: The capping box. Biochemistry 32:7605–7609, 1993.
55. Seale, J.W., Srinivasan, R., Rose, G.D. Sequence determinants of the capping box, a stabilizing motif at the N-termini of α-helices. Prot. Sci. 3:1741–1745, 1994.
56. Aurora, R., Srinivasan, R., Rose, G.D. Rules for α helix termination by glycine. Science 264:1126–1130, 1994.
57. Schellman, C. The αL conformation at the ends of helices. In: "Protein Folding." Jaenicke, R. (ed.). New York: Elsevier/North-Holland, 1980:53–61.
58. Colloc'h, N., Cohen, F.E. β-Breakers: An aperiodic secondary structure. J. Mol. Biol. 221:603–613, 1991.
59. Baker, E.N., Hubbard, R.E. Hydrogen bonding in globular proteins. Prog. Biophys. Mol. Biol. 44:97–179, 1984.
60. Leszczynski, J.S.F. Loops: a novel class of protein secondary structure (The Pennsylvania State University College of Medicine, 1986). PhD Thesis.
61. Kabsch, W., Sander, C. On the use of sequence homologies to predict protein structure: Identical pentapeptides can have completely different conformations. Proc. Natl. Acad. Sci. U.S.A. 81:1075–1078, 1984.
62. Cohen, B.I., Presnell, S.R., Cohen, F.E. Origins of structural diversity within sequentially identical hexapeptides. Prot. Sci. 2:2134–2145, 1993.
63. Serrano, L., Fersht, A.L. Capping and α helix stability. Nature 342:296–299, 1989.
64. Bruch, M.D., Dhingra, M.M., Gierasch, L.M. Side chain-backbone hydrogen bonding contributes to helix stability in peptides derived from an α helical region of carboxypeptidase. Proteins 10:131–139, 1991.
65. Lecomte, J.T.J., Moore, C.D. Helix formation in apocytochrome b5: The role of a neutral histidine at the N-cap position. J. Am. Chem. Soc. 113:9663–9665, 1991.
66. Lyu, P.C., Zhou, H.X., Jelveh, N., Weemer, D.E., Kallenbach, N.R. Position-dependent stabilizing effects in α helices: N-terminal capping in synthetic model peptides. J. Am. Chem. Soc. 114:6560–6562, 1992.
67. Serrano, L., Sancho, J., Hirshberg, M., Fersht, A.R. α helix stability in proteins: I. Empirical correlations concerning substitution of side-chains at the N- and C-caps and the replacement of alanine by glycine or serine at solvent-exposed surfaces. J. Mol. Biol. 227:544–559, 1992.
68. Lyu, P.C., Wemmer, D.E., Zhou, H.X., Pinker, R.J., Kallenbach, N.R. Capping interactions in isolated α helices: Position-dependent substitution effects and structure of a serine-capped peptide helix. Biochemistry 32:421–425, 1993.
69. Forood, B., Feliciano, E.J., Nambiar, K.P. Stabilization of α helical structures in short peptides via end capping. Proc. Natl. Acad. Sci. U.S.A. 90:838–842, 1993.
70. Argos, P., Palau, J. Amino acid distribution in protein secondary structures. Int. J. Pept. Prot. Res. 19:380–393, 1982.