PROTEINS: Structure, Function, and Genetics 38:149–164 (2000)

Optimized Representations and Maximal Information in Proteins

Armando D. Solis and S. Rackovsky*
Department of Biomathematical Sciences, Mount Sinai School of Medicine, New York, New York

ABSTRACT **In an effort to quantify loss of information in the processing of protein bioinformatic data, we examine how representations of amino acid sequence and backbone conformation affect the quantity of accessible structural information from local sequence. We propose a method to extract the maximum amount of peptide backbone structural information available in local sequence fragments, given a finite structural data set. Using methods of information theory, we develop an unbiased measure of local structural information that gauges changes in structural distributions when different representations of secondary structure and local sequence are used. We find that the manner in which backbone structure is represented affects the amount and quality of structural information that may be extracted from local sequence. Representations based on virtual bonds capture more structural information from local sequence than a three-state assignment scheme (helix/strand/loop). Furthermore, we find that amino acids show significant kinship with respect to the backbone structural information they carry, so that a collapse of the amino acid alphabet can be accomplished without severely affecting the amount of extractable information. This strategy is critical in optimizing the utility of a limited database of experimentally solved protein structures. Finally, we discuss the similarities within and differences between groups of amino acids in their roles in the local folding code and recognize specific amino acids critical in the formation of local structure. Proteins 2000;38:149–164.** © 2000 Wiley-Liss, Inc.

Key words: secondary structure; local sequence; sequence-structure relationship; information theory; optimized representations; amino acid clustering; alphabet collapse

INTRODUCTION

The foundation of bioinformatics is the processing of biomolecular sequence and structure data. Sequence and structure data are becoming available at a tremendous rate, as a result of the growth in genomic and structure-directed research. These data are not generally used in the form in which they are provided by experimentalists. For instance, in protein bioinformatics, such applications as homology searches and structure prediction frequently involve data compression in both the sequence and structure domains. This compression arises because only a finite amount of data is available. Therefore, to ensure statistical significance in these and other applications, meaningful grouping of data is necessary. Unfortunately, this processing always involves the loss of information, which compromises the accuracy of the studies in which the data are used. It thus becomes necessary in protein bioinformatic applications to balance the requirements of statistical significance and maximal conservation of information.

To date, there has been no quantitative study on the magnitude and effects of this information loss. Nor has there been an effort to optimize data processing techniques and information representation to minimize information loss. Here we use information-theoretic methods[1] to address the following areas:

- The development of a quantitative method for optimizing a protein sequence representation, given a specified representation of associated local backbone structure.
- A comparison of the amount of structural information that can be extracted from local sequence using several standard, widely used sequence and backbone structure representations to understand the compromises involved in their use, and the relative efficiency of different representations in encoding experimental data.
- Exploration of significant relationships among the amino acids with respect to backbone structure encoding, which may assist in the meaningful compression of sequence data.

Structural Representations

The conformation of the peptide backbone can be described by a range of structural representations, ranging from the most detailed (all-atom) to the most general (helix/sheet/coil). Representations that provide intermediate levels of detail include geometric descriptions of the path of the alpha-carbons in the backbone,[2] one of which is used in this study. The most popular, the canonical secondary structure representation, gives a description of local protein architecture in terms of organized structures that are universally characteristic of proteins, the alpha-helix and extended strand. These two structural features have the important characteristic

*Correspondence to: S. Rackovsky, Department of Biomathematical Sciences, Mount Sinai School of Medicine, Box 1023, One Gustave L. Levy Place, New York, NY 10029. E-mail: shelly@msvax. mssm.edu

Received 1 July 1999; Accepted 17 September 1999

that they are repeating. This means that there exist well-defined regions in the plane of the dihedral angles (ϕ, ψ) such that, when successive amino acids are assigned (ϕ, ψ) values in those regions, these structural features are automatically generated.[3] Residues that do not fall into these regions are typically said to be in a "coil" state. It should be remarked that, because chain reversals (bends) are not repeating structures, residues that occur in bends fall in one of these three states, rather than falling within a characteristic bend region.

There are various ways to transform atomic coordinates into secondary structure assignments. Not all secondary structure assignment algorithms rely exclusively on purely local atomic coordinates or on the (ϕ, ψ) description, which is the representation underlying the secondary structure approach. In fact, the most widely used method, that of Kabsch and Sander,[4] requires the detection of backbone hydrogen-bonding patterns.

Secondary structure prediction algorithms are widely used in efforts to gain insight into the three-dimensional structure of a sequence of interest.[5] Some of the earliest work in this area is that of Robson and collaborators.[6–13] In this body of work, information theory was used to construct ranking metrics, by which the decision could be made to assign a given amino acid residue to one or another of the allowed secondary structure states. The basic approach is to calculate the influence of a given sequence constraint (specification of the amino acid at the site in question, or specification of pairs of amino acids) on the tendency to assume each of the states. One then assigns the residue of interest to that state that gives the highest tendency. This method, called the GOR algorithm,[14] has become one of the most popular secondary structure prediction schemes in general use. More recent work has focused on neural network methods,[15] and on the use of multiple alignment methods[16,17] to improve the accuracy of secondary structure prediction.

The use of accuracy as a measure of success in most secondary structure prediction schemes implies that each residue must be unambiguously assigned to one state, no matter how weak the preference for that state is. Here we adopt a different viewpoint, based on the observation that a given sequence fragment can occur in different conformations in different proteins. This will lead to methods by which protein structure representations can be optimized subject to specified constraints.

Sequence Representations

The compression of protein sequences is a common informatic strategy. By sequence compression, we mean the clustering of the 20 amino acids into a smaller number of groups, each of which contains a number of amino acids related by some criterion of interest. This approach is based on the observation that some pairs of amino acids show sufficient kinship to allow substitutions and mutations while maintaining the same overall fold. Explicit sequence compression occurs, for example, in statistical studies,[18] where issues of data quantity and statistical significance arise, and in simulation studies,[19,20] where ultrasimple hydrophobic/hydrophilic (HP) sequence alphabets have been used to model aspects of protein folding. Implicit sequence compression arises in the construction of amino acid substitution matrices for use in alignments.[21–23] The statement that two (or more) amino acids are interchangeable is equivalent to the construction of a reduced amino acid sequence.

In this study, we present a novel approach to the optimization of amino acid clustering for some information-theoretic criterion. In recent work,[24] multiple sequence alignments, rather than purely structural criteria, were used to construct equivalence classes of amino acids. In contrast, the present work addresses the link between sequence and structure by optimizing sequence compression based exclusively on the distribution of structures characteristic of each sequence fragment.

THEORY AND METHODOLOGY

Our strategy is to establish a quantitative relationship between local amino acid sequence and local structure preference. We examine the effect on that relationship of varying the manner in which local sequence is described, to find that representation of sequence that leads to the *best* sequence-structure relationship. The definition of what constitutes the best relationship lies at the heart of the methodology we have developed.

The first step in our investigation is the specification of a measure of the conformational preference of any given sequence fragment. This must reflect completely the conformational properties of the fragment, and it must be so defined that optimization is a meaningful concept. A measure that satisfies these requirements is given by the structural distribution associated with a given sequence fragment. We generate these distributions as follows:

1. A nonredundant database of protein X-ray structures is established.
2. A representation of protein sequence is selected. This may be the full amino acid sequence or any compressed sequence representation of interest.
3. A representation of protein backbone structure is selected. This may be a high-resolution representation (using any convenient set of variables), a secondary-structure representation, or any other useful representation.
4. A length scale is selected. This is the length of the sequence segments whose backbone structures we want to examine.
5. For each sequence fragment of the specified length, all associated local backbone structural units in the data set are identified and placed in a histogram. This histogram is the distribution that characterizes the conformational preference of the given sequence fragment. Clearly, the shape of the histogram depends on both the sequence representation (which tells us which sequence fragments are possible) and the structure representation (which tells us which structures can be associated with each sequence fragment). Figure 1 illustrates this procedure, using the tetramer length scale as an example.

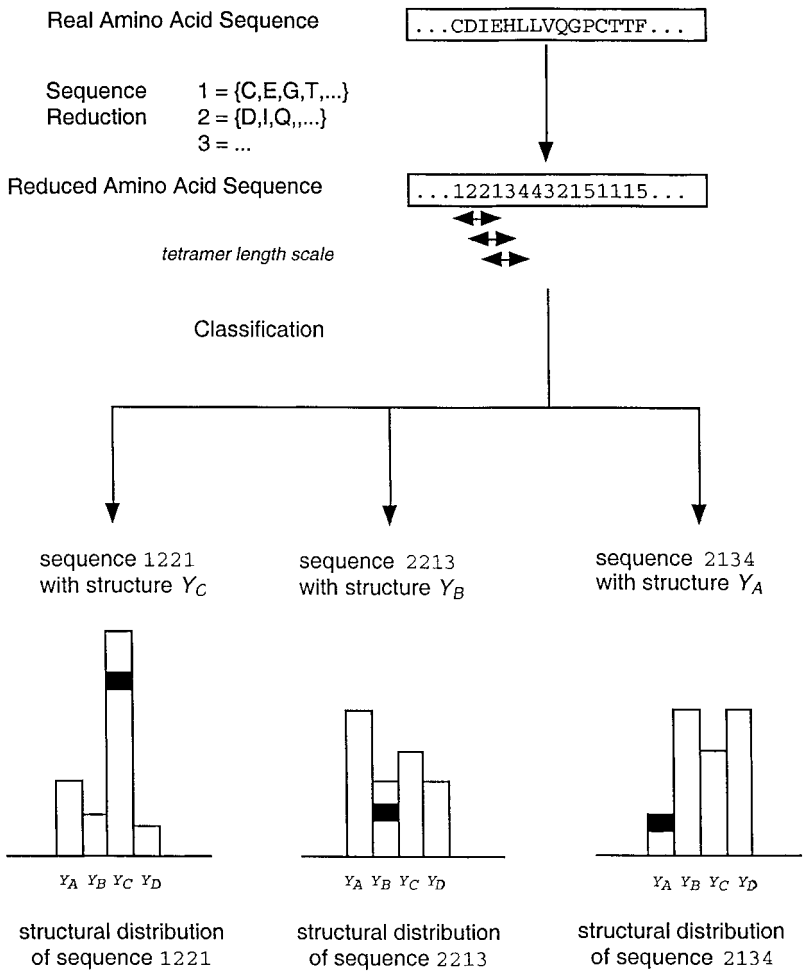


Fig. 1. Procedure used to construct a structural distribution for each reduced sequence at the tetramer length scale. The amino acid sequences of all the proteins in the data set are translated into their reduced form by alphabet collapse. The local structure of each overlapping tetramer is noted and included in the structural distribution characterizing that tetramer sequence. Once all the structures are classified, entropies and information values are calculated from these structural distributions.

Note that this distribution completely specifies the conformational preference of the sequence fragment under consideration, as reflected in experimental data. Furthermore, use of this distribution opens the door to optimization of the sequence-structure relationship. This follows intuitively from the observation that changing the sequence representation, given a specified representation of structure, changes the ensemble of possible sequence fragments, and, therefore, the ensemble of structure distributions. Optimization of the sequence-structure relationship has occurred when a sequence representation that gives the narrowest possible ensemble of structural distributions is chosen. That sequence representation has then been found which carries the most information about structure.

Structural Entropy and Information Gain

We now make these intuitive remarks quantitative. A useful measure of the size of a structural distribution Y characteristic of a local sequence $X = x$ is its entropy H,[1] given by

$$H(Y \mid X = x) = -\Sigma_i\, pi\, log\, pi. \quad (1)$$

Here p_i is the fraction of occurrences of the distributed structural characteristic in bin i. The index i summarizes in convenient shorthand all the variables that are used to describe that structural characteristic (cf. Refs. 18,25). (Strictly, Y is a random variable for structure characterized by an underlying probability distribution; by citing Y as a distribution, we simply refer to this underlying probability distribution.) The units for entropy are "nats" when the natural logarithm is used, and "bits" when the base 2 logarithm is used.

Note that we have no a priori standard by which to decide whether a given distribution is broad or narrow. We fulfill this requirement by generating a large ensemble of random distributions in association with each actual structural distribution. For a distribution Y that describes the structures associated with n_Y occurrences of a particular sequence fragment $X = x$ in the data set, we generate random structural distributions containing n_Y structures of the selected length. These are chosen without replacement from the total distribution of structure fragments in our database, without reference to sequence. From this ensemble, we can determine an average entropy of random structural distributions as a function of n_Y, $EH(Y \mid N = n_Y)$. We then consider the entropy difference

$$I_g\,(X = x, Y) = EH\,(Y \mid N = n_Y) - H(Y \mid X = x). \quad (2)$$

This function is positive if the experimentally observed structure distribution is narrower than the average of random distributions of the same size, and negative if it is broader. It therefore provides an intrinsically calibrated measure of distribution size. Furthermore, it can form the basis for a suitable objective function for optimization, as we shall show below.

The function $I_g(X = x,Y)$, which we call information gain, is actually a special case of mutual information, a function well known in information theory. This connection is described in the Appendix. One can think of $I_g(X = x,Y)$ as measuring quantitatively the additional information available to us as a result of specifying the local sequence of a protein backbone fragment to be x.

The local sequence and structure spaces may be subdivided and discretized in various ways. A measure of the success of a particular subdivision for the goal of optimizing the local sequence-structure relationship is the average information gain

$$I_g\,(X,Y) = EI_g\,(X = x,Y) = \Sigma_x\, Ig(X = x,\, Y)p_X(x) \quad (3a)$$

where $p_X(x)$ is the probability of occurrence of local sequence x in proteins. This equation reduces to

$$I_g\,(X,Y) = EH(Y|N) - H(Y|X). \quad (3b)$$

The quantity $EH(Y|N)$ can be taken as the effective structural entropy of the entire data set. A critical advantage of using the average information gain is that it automatically corrects for small sample sizes arising from use of a limited data set.

We want to find those subdivisions of the sequence and structure spaces that lead to the greatest information gain. This objective is achieved by a direct maximization:

$$I_g^{\,max}(X,Y) = max\{I_g(X,Y)\}. \quad (4)$$

Definitions of Local Sequence and Backbone Structure and Construction of Sequence-Dependent Structural Distributions

Two structural representations are considered here: the three-state secondary structural assignment scheme of Kabsch and Sander[4] and the generalized bond matrix representation of Rackovsky,[2] hereafter referred to as DSSP (Dictionary of Secondary Structure of Proteins) and GBMR, respectively. DSSP is the most widely used automatic secondary classification scheme in the analysis of structures and evaluation of putative prediction algorithms and, therefore, has become the de facto standard in protein science. While DSSP relies on the detection of hydrogen-bonding patterns in the polypeptide backbone, GBMR is a direct representation of local backbone structure. It describes the path of the alpha-carbon backbone through space, without regard to hydrogen bonding patterns.[2,26–28] This chain (Fig. 2) is completely characterized by three parameters: bond length, bond angle (formed by adjacent bonds), and bond dihedral angle (arising from three consecutive bonds.

Each complete protein structure in the data set is decomposed into backbone structural segments by using either the GBMR or DSSP representation. A complete structural unit in the GBMR tetramer representation, used throughout this study, involves four consecutive alpha-carbons and the three virtual bonds v_i, v_{i+1}, v_{i+2} connecting them, and can be fully described by three bond lengths (l_i,l_{i+},l_{i+2}), two bond angles (θi between v_i and v_{i+1}, and θ_{i+1} between v_{i+1} and $v_{i+2}$and one bond dihedral angle (γ_ι, formed by all three virtual bonds) (Fig. 2). A structural unit in the DSSP representation is composed of the assigned secondary structural state at one residue position. Each of these backbone segments is then associated with the local amino acid sequence, of the desired length, in which they occur. The local sequence block can be one residue in length or longer, depending on how much sequence information is desired for analysis or prediction. Here we analyze structural information arising from sequence 2-mers up to 9-mers, with the tetramer length scale as the basis for the most extensive studies. Regardless of sequence length scale, the size of the structural units (GBMR and DSSP) are held constant. We use the assigned DSSP backbone structure of the *centermost* position of the sequence block as the DSSP structural unit associated with the local sequence. This is unambiguous for odd length scales, where a middle position exists; for even length scales, we have found that the informational quantities of interest do not depend significantly on which of the two centermost positions is chosen. Similarly, the centermost GBMR four alpha-carbon unit is used as the GBMR structural unit.

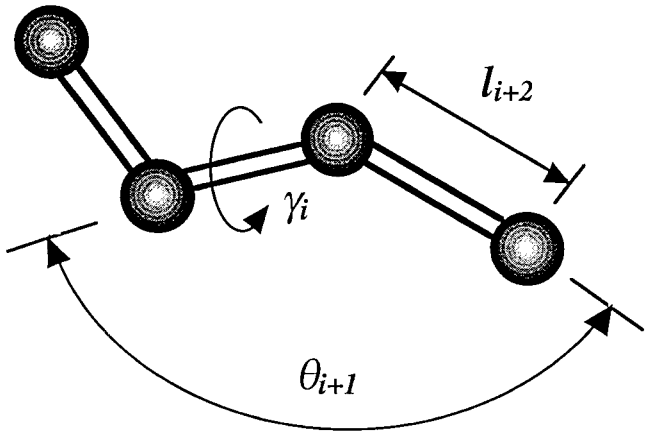


Fig. 2. The basic tetramer unit of the GBMR structural representation.[2] Four consecutive alpha-carbons are connected by virtual bonds. These virtual bonds can be completely characterized by six parameters: three bond lengths (l), two bond angles (θ), and one dihedral angle (γ).

The complete structural data set contains 114 protein structures used previously.[2] The local GBMR structures of the proteins in the data set were generated by using algorithms developed previously.[2] In the GBMR partition used, there are a total of 17,496 structural states, of which only 587 are occupied. The DSSP structures were obtained from processed files available from EMBL at: ftp.embl.heidelberg.de/pub/databases/protein_extras-/dssp.

Search for the Maximum Information Gain

Our fundamental conjecture is that clustering structurally similar amino acids should result in narrow local structural distributions for each tetramer sequence. Compact distributions result in lower structural entropies and higher information gains. Therefore, maximizing the information gain $I_g(X,Y)$ (Eq. 4) by searching across the ensemble of possible amino acid clusterings should produce amino acid clusters whose members show the strongest local coding similarities, for a given structural descriptor.

Because the number of possible cluster sets is enormous, an exhaustive search is not feasible. Rather, we use a Monte Carlo algorithm with an empirically chosen sampling procedure and decision criterion. The search cycle has six parts:

1. An initial amino acid clustering is randomly generated. The information gain is computed from the resulting structural distributions using Eq. 3.
2. To generate the next trial clustering, the current grouping of amino acids is altered by a random change in membership for one, or for two, three, or four amino acids simultaneously. The number of amino acids involved in the change is randomly chosen, with a sampling frequency of (0.4, 0.3, 0.2, 0.1), respectively.
3. From this trial clustering, a new information gain is computed and compared with the old value. If the new value is higher, the trial clustering is kept, and another iteration is made from the second step. Otherwise, the old grouping is kept.
4. When no trial grouping is accepted after 10,000 trials, the criterion for acceptance is relieved slightly, by accepting any trial grouping with an information gain that is at least 10% below the old information gain.
5. After another set of iterations, if the maximum information gain does not increase further, the algorithm is restarted with a new random amino acid grouping.
6. This process is iterated 1,000 times.

It should be remarked that we are conducting Monte Carlo optimization, rather than generating statistical ensembles. Therefore, the search algorithm need not follow a realistic, energetically derived sampling procedure (such as a Metropolis-type criterion) and can be designed and altered with great flexibility.

This optimization procedure works well. To ensure convergence, we performed at least 100 cycles for each optimization case (specified local sequence length, structural description, number of amino acid clusters, etc.). Every pass through the six steps outlined above, using a given set of sequence and structural representations, produces the *same* optimal grouping of amino acids.

RESULTS AND DISCUSSION

Backbone Structural Information From Local Sequence

Amount of backbone structural information found in the local tetramer sequence

Tables IA and IB show the optimal amino acid clusters, characterized by the maximum information gain (Eq. 4), resulting from an analysis of the local structures of four-residue segments. Initial inspection reveals that these optimal groupings are strikingly similar to the consensus of a mass of amino acid similarity studies.[29] More analyses of amino acid coding relationships will be found below.

To analyze the effect of the size of the database in determining the information values of interest, we applied the clustering procedure to partial sets of the data. In Figure 3, maximum information gain is plotted for tetramers in the GBMR and DSSP scheme against the number of amino acid clusters n_c using different size data sets (100%, 75%, 50%). The smaller sets (75%, 50%) were organized by randomly picking out subsets of complete proteins from the full data set. A number of observations are clear from the graph. Foremost, $I_g^{max}(X,Y)$ reaches a higher maximum for GBMR at 0.24 nats, around 60% higher than that of DSSP (0.15 nats) for both 50% and 100% protein sets. Also, it is clear that the size of the data set has a significant effect on maximum information gain in both DSSP and GBMR. With smaller data sets, I_g^{max}falls noticeably at higher amino acid cluster numbers. The apparent loss of information, observed at the right tails of the plots, is undoubtedly caused by the inadequacy of the data set. At higher cluster numbers (leading to higher numbers of unique tetramers), there are very few observations per tetramer, causing the partition to move toward complete memorization and diminished predictive power. (The concept of complete memorization is easily understood by illustration; see Fig. 4). Our objective function detects such undesired results and reflects this clearly in the plots.

The maximum value of I_g^{max} increases slightly as the size of the data set is increased, along with a minute shift in its location with respect to the cluster number (Table II). We expect that in the limiting case of an infinite protein set, I_g^{max} should monotonically increase up to $n_c = 20$, or at least plateau at a constant level. This is because recognizing the unique local coding properties of each amino acid, as one goes to higher cluster numbers should enhance the amount of structural information available from sequence. The upward trend summarized in Table II, as the number of proteins in the data set is increased, reflects this expectation.

Fraction of structural information found locally

The maximum fraction of backbone structural information encoded in the local sequence is obtained by normalizing the information gain $I_g^{max}(X,Y)$ by $EH(Y|N)$, the effective structural entropy of the data set:

$$f_I(n_c) = Ig^{max}(X,Y)n_c/EH(Y|N)n_c$$
$$= [EH(Y|N) - H(Y|X)]/EH(Y|N) \quad (5)$$

where n_c is the number of amino acid clusters used to simplify the sequence space. Plots of f_I against the number of amino acid clusters at the tetramer length scale, across different data set sizes, for both GBMR and DSSP (Fig. 5), show the tails increasing almost linearly at higher cluster numbers.

The value of f_I is expected to converge to its true value as

TABLE IA. Partition of the Amino Acids in Terms of Similarities in the Distribution of GBMR Structures They Encode†

2 (0.123)	Ala, Asp, Cys, Glu, Phe, His, Ile, Lys, Leu Met, Asn, Gln, Arg, Ser, Thr, Val, Trp, Tyr															Gly Pro				
3 (0.182)	Gly															Ala, Asp, Cys, Glu, Phe, His Ile, Lys, Leu, Met, Asn, Gln Arg, Ser, Thr, Val, Trp, Tyr				Pro
4 (0.228)	Gly												Ala, Asp, Glu, Lys Asn,Gln,Arg,Ser,Thr						Cys, Phe, His, Ile Leu, Met, Val, Trp, Tyr				Pro
5 (0.241)	Gly								Asp Asn							Ala, Glu, His Lys, Gln, Arg Ser, Thr				Cys, Phe, Ile Leu, Met, Val Trp, Tyr				Pro
6 (0.243)	Gly						Asp Asn						Ala, Glu, Phe His, Ile, Lys Leu, Met, Gln Arg, Val, Trp Tyr						Cys Thr				Ser				Pro
7 (0.243)	Gly				Asp Asn						Ala, Glu Phe, Ile Lys, Leu Met, Gln Arg, Val Trp, Tyr						Cys His				Thr				Ser				Pro
8 (0.243)	Gly						Asp						Asn				Ala, Glu Phe, Ile Lys, Leu Met, Gln Arg, Val Trp, Tyr						Cys His				Thr				Ser				Pro
9 (0.241)	Gly						Asp						Asn						Ala, Glu Phe, Ile Lys, Leu Met, Gln Arg, Val Trp, Tyr						His				Cys				Thr				Ser				Pro
10 (0.234)	Gly						Asp						Asn						Ala,Glu Phe,Ile Lys,Le uMet,G ln Arg,Val Trp						Tyr				His				Cys				Thr				Ser				Pro
11 (0.226)	Gly						Asp						Asn						Ala Glu Phe Ile Lys Leu Met Gln ArgVal						Trp				Tyr				His				Cys				Thr				Ser				Pro
12 (0.212)	Gly						Asp						Asn						Ala Glu Phe Ile Lys Leu Met Gln Val						Arg				Trp				Tyr				His				Cys				Thr				Ser				Pro
13 (0.192)	Gly						Asp						Asn						Ala Glu Phe Ile Lys Leu Met Val						Gln				Arg				Trp				Tyr				His				Cys				Thr				Ser				Pro
14 (0.171)	Gly						Asp						Asn						Ala Glu Phe Ile Lys Leu Val						Met				Gln				Arg				Trp				Tyr				His				Cys				Thr				Ser				Pro

†These partitions of amino acids have been found to give the optimal structural information gain (Eq. 4) using the GBMR backbone structural representation at the tetramer length scale.

TABLE IB. Partition of the Amino Acids in Terms of Similarities in the Distribution of DSSP Structures They Encode[†]

2 (0.066)	Ala, Cys, Glu, Phe, His, Ile, Lys Leu, Met, Gln, Arg, Val, Trp, Tyr											Asp, Gly, Asn, Pro, Ser, Thr	
3 (0.106)	Ala, Glu, His, Lys, Gln, Arg					Cys, Phe, Ile, Leu, Met, Val, Trp, Tyr						Asp, Gly, Asn, Pro, Ser, Thr	
4 (0.122)	Ala, Glu, His, Lys, Gln, Arg				Cys, Phe, Ile, Leu, Met, Val, Trp, Tyr					Asp, Asn, Ser, Thr		Gly, Pro	
5 (0.133)	Ala, Glu, His, Lys, Gln, Arg			Phe, Ile, Leu, Met, Val, Trp, Tyr				Cys, Ser, Thr			Asp, Asn		Gly, Pro
6 (0.145)	Ala, Glu, Lys, Gln, Arg		Phe, Ile, Val			Leu, Met, Trp, Tyr			His, Cys, Thr			Asp, Asn, Ser	Gly, Pro
7 (0.152)	Glu, Lys, Gln, Arg		Phe, Ile, Val		Leu, Met, Trp, Tyr			Ala, Cys, His			Ser, Thr	Asp, Asn	Gly, Pro
8 (0.155)	Glu, Lys, Gln, Arg		Ile, Val		Leu, Trp, Tyr		Ala, Met		Cys, Phe		His, Thr	Asp, Asn, Ser	Gly, Pro
9 (0.154)	Glu, Lys, Gln, Arg		Ile, Val		Leu		Phe		Ala, Met, Trp	Cys, Tyr		His, Thr	Asp, Asn, Ser	Gly, Pro
10 (0.152)	Glu, Lys, Gln, Arg		Ile, Val		Leu, Tyr		Phe		Ala, Met	Trp		His, Thr	Cys	Asp, Asn, Ser	Gly, Pro
11 (0.150)	Ala, Glu, Lys, Gln, Arg		Ile, Val		Leu, Met		Phe		Tyr	Trp		Cys	His	Thr	Asp, Asn, Ser	Gly, Pro
12 (0.148)	Ala, Glu, Lys, Gln, Arg		Ile		Val		Leu, Met		Phe	Tyr		Trp	Cys	His	Thr	Asp, Asn, Ser	Gly, Pro
13 (0.144)	Ala, Glu, Lys, Gln, Arg		Ile		Val		Leu		Met	Phe		Tyr	Trp	Cys	His	Thr	Asp, Asn, Ser	Gly, Pro
14 (0.137)	Asp, Glu, Lys, Gln		Arg		Ala		Ile		Val	Leu		Met	Phe	Tyr	Trp	Cys	His	Thr	Gly, Asn, Pro, Ser

[†]These partitions of amino acids have been found to give the optimal structural information gain (Eq. 4) using the DSSP backbone structural representation at the tetramer length scale.

the number of protein structures in the data set increases. In the limit, when observed frequencies approach the true probabilities they attempt to estimate, $H(Y|X)$ approaches its true value, and $EH(Y|N)$ approaches $H(Y)$. Thus,

$$f_I\,(N \dashrightarrow \infty, n_c) = lim\,[EH\,(Y|N) - H(Y|X)]/EH(Y|N) = I(X;Y)n_c/H(Y). \quad (6)$$

In other words, f_I at the limit is just the Shannon mutual information $I(X;Y)$ (derived in the Appendix) normalized by the entropy of structures before sequence information.

From the plots in Figure 5, we observe that a reasonable estimate for the measure f_I can be achieved with our current data set for both GBMR and DSSP structural descriptions. Furthermore, an extrapolation of f_I to cluster number 20, at which point the identities of all 20 amino acids are recognized, is possible by simply extending this plot from its rightmost point. We can then estimate the absolute limit of the fractional information $f_I\,(n_c = 20)$, and the absolute mutual information (from Eq. 6) as

$$I(X;Y)_{n_c = 20} f_I\,(n_c = 20) \cdot H(Y). \quad (7)$$

This calculation is straightforward because $H(Y)$ can be sufficiently approximated by the structural entropy of the entire data set. $(H(Y)$ for DSSP is 1.04 nats, whereas it is 4.08 nats for GBMR.) In our study of tetramer sequences, the extrapolated fraction of structural information f_I is 0.26 for DSSP and 0.13 for GBMR (Table III). These result in estimates of the potential information gain, ranging from 0.24 to 0.27 nats using the DSSP descriptor and from 0.41 to 0.53 nats for GBMR, roughly a 100% difference.

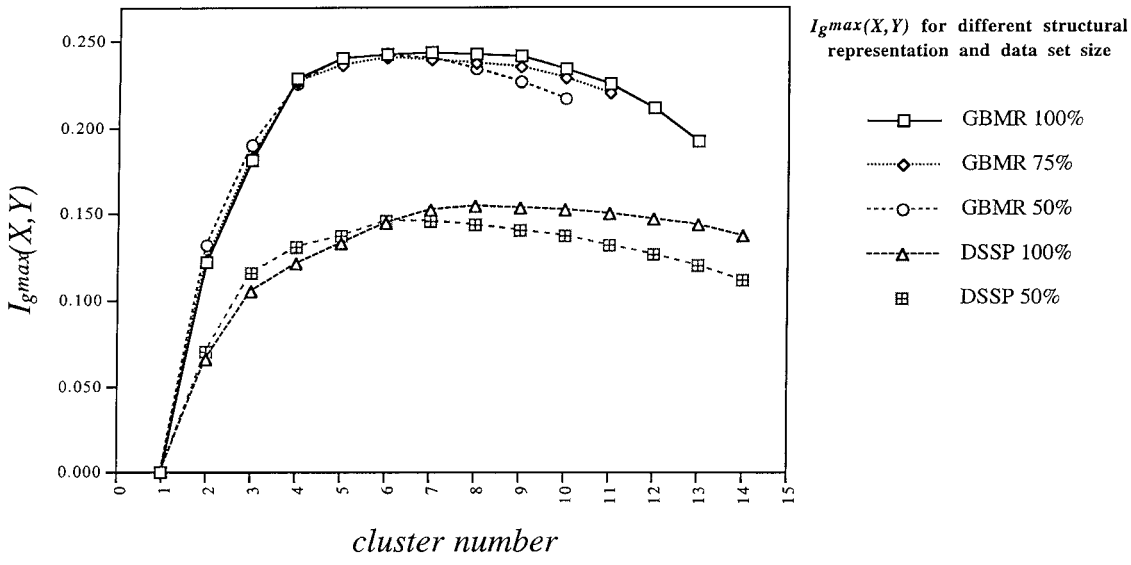


Fig. 3. Maximum information gain I_g^{max} (in nats) for GBMR and DSSP structural representations, for different cluster numbers, and using different data set sizes.

Different local sequence lengths

Similar analyses were applied at the dimer and trimer sequence length scales. Because there are fewer possible dimer and trimer sequences than tetramer sequences given the same n_c, the calculation for I_g^{max} was extended to larger numbers of clusters. Associated with each unit of GBMR structure is a local sequence block made up only of the amino acids occupying positions i, $i+1$, and $i+2$ for the trimer analysis, and positions $i+1$ and $i+2$ for the dimer analysis, excluding knowledge of the amino acids at position $i+3$ and positions i and $i+3$, respectively. The dimer local sequence block associated with each unit of DSSP structure is composed of the identities of the amino acids occupying that position and the one preceeding it in the chain, whereas for the trimer analysis, the identity of the succeeding amino acid is also considered. As an illustration, data from the dimer length scale in DSSP are plotted in Figure 6. These plots confirm the stability of f_I as a function of database size and show that f_I is a strong measure of the real information gain. Similar behavior was observed in the trimer length scale (data not shown).

Calculations for f_I were also made for segments larger than four residues. As always, the GBMR structural unit is in the centermost position of its corresponding local sequence block. For 5-mer [$i \ldots i+4$] and 6-mer [$i \ldots i+5$] sequence segments, the alpha-carbons comprising the GBMR structural unit traverse [$i+1 \ldots i+4$]. The DSSP structural unit of the same sequence block is the assigned secondary structure of the centermost residue of the segment. For the 5-mer, it is the secondary structure of the 3rd residue, and for the 6-mer and the 7-mer, it is that of the 4th residue.

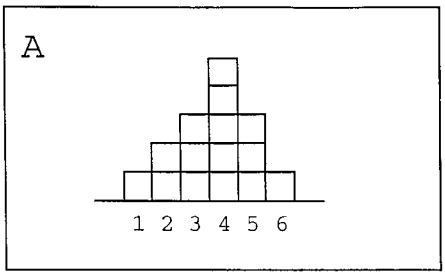
Figure 7 shows plots of f_I against cluster number for sequence block lengths up to 9. Because of the exponential increase in the number of unique sequence segments, the number of clusters was severely limited at longer segments. It should be no surprise that increasing the segment length generates a greater amount of accessible information. Note that while f_I , the maximum *fraction* of information gain, is always greater for DSSP than for GBMR, f_I describes the *ratio* of information gain to *H(Y)* of the same structural descriptor. Because the *H(Y)* of GBMR is roughly four times that of DSSP (4.08 vs. 1.04 nats), the GBMR descriptor results in consistently higher values for $I_g(X,Y)$ (from Eq. 7). These results are summarized in Table III.

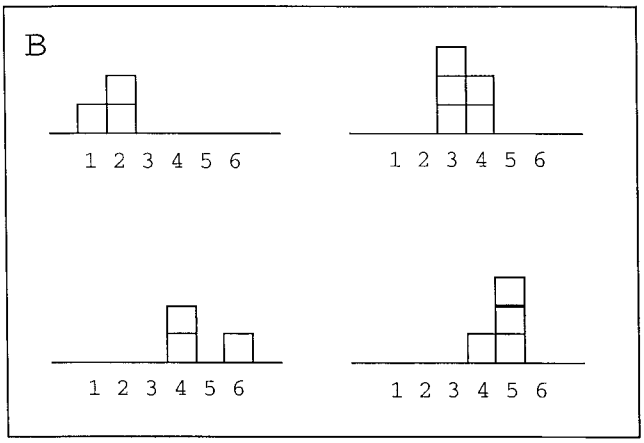
Quantity and quality of information as a function of structural representation

The foremost reason for the difference in information gain between the DSSP and GBMR schemes lies in the nature of secondary structure definitions. The major drawbacks of the three-state DSSP are the following:

1. It is easy to visually identify the core of regularly occurring backbone structures (helices and strands), but there is often some ambiguity in classifying *every* residue, particularly at the ends of helices and sheets.[30]
2. The cutoff distance and orientation defining hydrogen bonding are arbitrary, and consequently, a slight change in criteria may produce different assignments.[4]
3. The secondary "structure" loop is defined as the absence of hydrogen bonding regularity, even though some "loop" residues occur in the same Ramachandran regions as helical and strand residues, and exhibit sequence-dependent backbone structural distributions similar to those of their "ordered" counterparts.[31]

Failure to detect differences in fine structure within the helical and extended secondary structural classes may be another source of difference between DSSP and GBMR.[32] The rigidity Pro imparts to the backbone, and its associa-





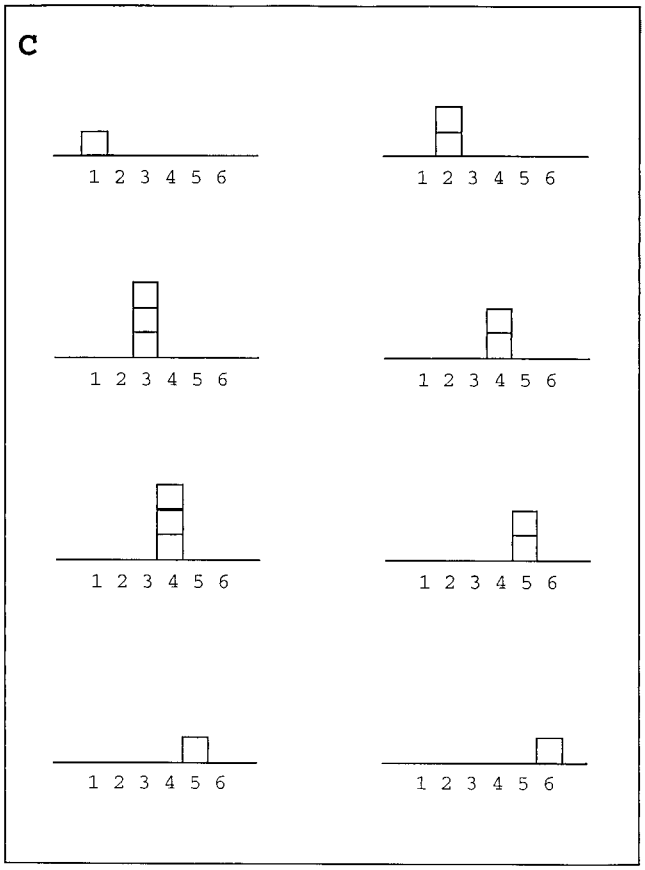


TABLE II. Maximum Values of I_g^{max} for Tetramer Length Scale

	GBMR		DSSP	
	I_g^{max} (nats)	No. of clusters	I_g^{max}	No. of clusters
Data set size (%)				
100	0.2434	7	0.1547	8
75	0.2403	6		
50	0.2398	6	0.1465	7

tion with bends and kinks in helices and turns linking secondary structures, are not explicitly indicated in any 1D secondary structure string. Similarly, the peculiarities Gly imparts to sheets, helices, and turns in which it participates are not considered in a coarse three-state DSSP classification. Indeed, in the GBMR representation (but not in DSSP) these two amino acids quickly cluster out of the group of amino acids to maintain their identities as distinctly different from the rest (Table I). Other amino acids, particularly the small ones (Asn, Asp, Thr, Ser) that are frequently involved in special turns and other side chain-main chain interactions also cluster out early in GBMR, reflecting their significance for the formation of local structure.

Amino Acid Similarities: Locally Informative Amino Acids

In a search for the most structurally informative simplified description of short sequences, the amino acids were partitioned on the basis of similarities in local structural coding behavior. For instance, a pair of amino acids *B* and *J* are said to be locally related if they encode distributions of local structures that do not differ significantly. Detecting similarity via structural propensities and distributions explicitly takes into account the fact that local sequences can have not one but a range of backbone structures in folded proteins.[33–35] Equally important, local coding similarity between a pair of amino acids is determined not only by the local structures that align, or within any particular secondary structure, but rather by their relative propensities for *all* types of structures.

Fig. 4. An illustration of the phenomenon of complete memorization. Two different partitions of the local sequence space *X* generate, for each particular local sequence, a distribution of structures. **A:** A hypothetical structural data set before any sequence-dependent partition is made. **B** and **C:** Local sequence is used to partition the data set. In the first partition (B), four unique local sequence blocks were considered, resulting in a partition of the full data set into four subsets, whereas in the second partition (C), eight unique sequence blocks were used to partition the data. In the first partition, the structural entropy given by each local sequence, $H(Y|X = x)$, is greater than zero. In the second partition, $H(Y|X = x')$ for all x' is zero (because only *one* type of structure occurs in the data set for each local sequence x', and thus by Eq. 1, $-1 \ln 1 = 0$). Consequently, its apparent *average* structural entropy is zero, whereas that of the first partition is greater than zero. As the data are partitioned into more and more subsets, the resulting apparent average structural entropy approaches zero. Eventually, complete memorization of the data is reached when the number of partitions equals or exceeds the number of distinct data points in the full data set.

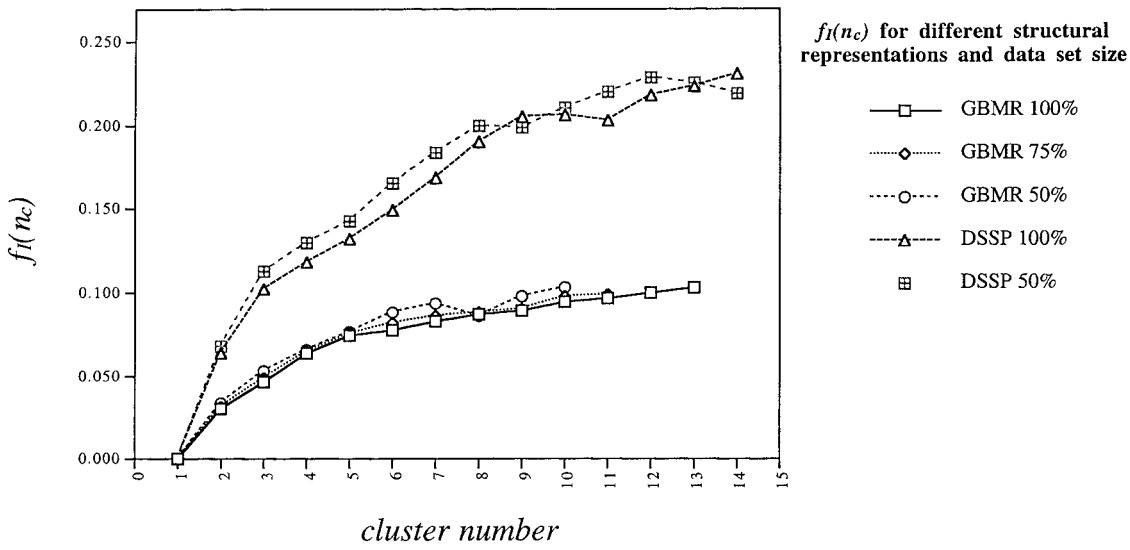


Fig. 5. Plots of f_I, the ratio of the maximum information gain to the effective structural entropy of the database, for both GBMR and DSSP, at different data set sizes.

TABLE III. Estimates for Maximum Fractional Information f_I and Information Gain I_g^{max}, Extrapolated to $n_c = 20$†

Local sequence length	GBMR		DSSP	
	Fractional info range f_I	Info gain range I_g^{max}	Fractional info range f_I	Info gain range I_g^{max}
Trimer	0.07–0.08	0.29–0.33	0.17	0.18
Tetramer	0.10–0.13	0.41–0.53	0.23–0.26	0.24–0.27
Pentamer	0.11–0.17	0.45–0.69	0.25–0.38	0.26–0.40

†Approximation of ranges was performed by linear extrapolation using those data points for f_I at lower n_c where the slope appears to be linear. The lower bound of a given range is the actual value observed at the highest n_c measured for the particular length scale, whereas the upper bound is given by the linear extrapolation to $n_c = 20$. For the GBMR representation, the data points used for extrapolation in the trimer, tetramer, and pentamer length scales are [$n_c = 8 \ldots 16$], [$n_c = 6 \ldots 13$], and [$n_c = 6 \ldots 8$], respectively. For the DSSP representation, the points used for the tetramer and pentamer length scales are [$n_c = 8 \ldots 14$] and [$n_c = 6 \ldots 10$], respectively. Because the actual calculation for the trimer length scale using DSSP was able to reach $n_c = 20$, no extrapolation was required, and the actual values are indicated instead of an approximate range. The extrapolation ranges, which involve values of f_I, are transformed into $I(X, Y)_{n_c=20}$ using Equation (7).

This similarity measure is different from other amino acid clustering or relatedness studies in a number of ways. First, no physicochemical variables are taken into account a priori. Some studies use arbitrarily chosen sets of physicochemical attributes to characterize each amino acid, which are then integrated with arbitrary weights.[36,37] These methods rest on the assumption that similarities in those characteristics indicate kinship among amino acids with respect to local structural coding or mutability. The procedure developed herein classifies amino acids solely in terms of their local structural propensities, and thus eliminates any artificial bias brought about by arbitrarily quantifying their physicochemical features. Other studies make use of multiple sequence alignment of homologous structures to derive amino acid substitution matrices.[21] These methods are dependent on the availability and quality of alignments of related sequences, which may bias the classification. We note that work that considered local structural similarities in the clustering of amino acids,[32] based on analyses at a sequence length scale of 1, gave results that are generally consistent with the results we present here. The present work is more general, in that we allow sequence fragments of longer length, and are therefore able to include the influence of multi-residue correlations on the optimal clustering.

Clustering in the GBMR structural representation

An examination of the course of amino acid clustering using the GBMR representation at the tetramer sequence length scale (Table IA) reveals that the two most unusual amino acids, Gly and Pro, separate out first. Such behavior is readily understood in light of the unusual structural features of these two amino acids. Gly, having no side chain, allows maximum torsional flexibility of the backbone, permitting access to less populated regions of the Ramachandran space (positive phi values). Analysis of tetramer virtual bond structural preferences (Fig. 8A) demonstrates that Gly enables the tetramer backbone to adopt conformations in the irregular regions of the virtual dihedral angle space, or outside the two major peaks of regular secondary structures (130° for helices and −30° for strands). In particular, tetramer sequences containing a Gly at position 2 are able to stabilize less common conformations in the vicinity of dihedral angle 50°, and similarly those with a Gly at position 3 have a higher preference than expected for angles around 170°. On the other hand, because the side chain of Pro forms a ring with the peptide backbone, the backbone conformations of adjacent residues are severely affected. The effect of Pro on the preceding residue in the chain is well understood: such residues

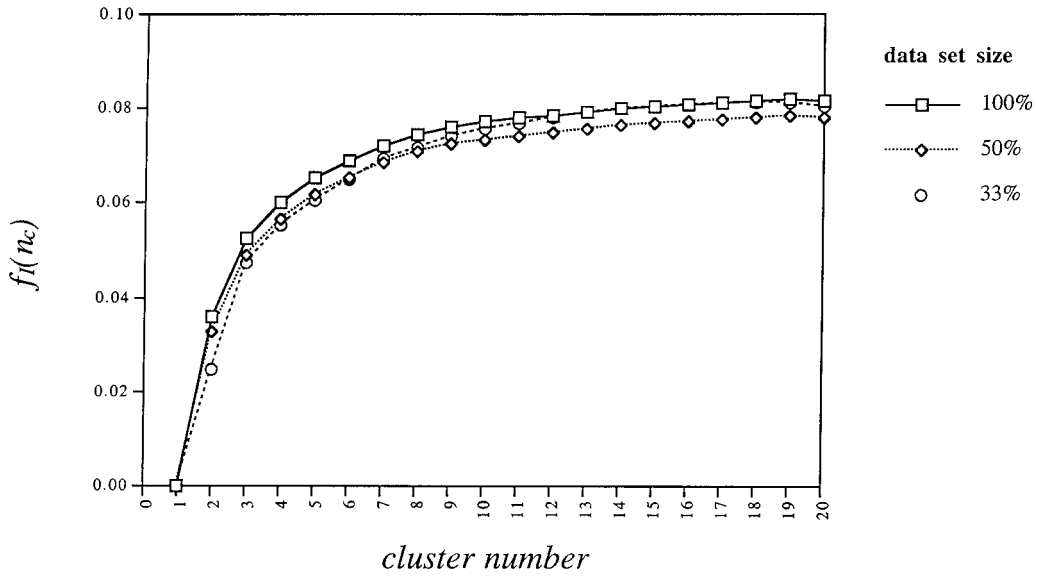


Fig. 6. The f_I plot for the dimer local sequence, using the DSSP structural description, shows stability across different subsets of the data set (100%, 50%, 33%).

are known to have a marked preference for the β region.[39] This is amplified in an analysis of tetramer virtual bonds (Fig. 8B), which shows the influence of Pro on the backbone conformation not of the residue immediately preceding it, but of the full tetramer. The virtual dihedral angle profiles of tetramers containing Pro in the first or second positions are similar to the profile given by the entire data set, whereas the profiles of those having Pro in either the third or fourth position deviate significantly, having an increased preference for extended as well as irregular conformations.

Other amino acids which cluster out early are Asp, Asn, Ser, Thr, and Cys. Apart from Cys, these amino acids are small and have been implicated as significant in the formation of turns and other conformations, because of the ability of their side chains to form hydrogen bonds with backbone groups.[40–44] Local roles of these amino acids include stabilizing β-turns by Gly and Pro[45] and Asp, Asn, Ser, and Thr;[46] helix capping by Asn, Gly, Ser, and Thr;[47] and N-terminal capping box motif formation by Ser, Thr, Asp, and Asn.[48] Cys stands out because it is able to stabilize less populated backbone conformations by constraints introduced through formation of disulphide bridges with other spatially adjacent cysteines.

At intermediate stages of clustering, the 20 amino acids are divided into several groups, each containing single amino acids, together with a larger group containing the rest. The latter group contains those amino acids with large side chains. In later stages of clustering, there is a differentiation between those amino acids that are able to induce unique local structural propensities, and those that show very similar and relatively standard local coding propensities. The uniqueness of local structural coding shown by sequence fragments containing the more informative amino acids may be essential for stabilizing structures that are critical in the folding process. The remaining larger group contains several amino acids whose primary characteristic appears to be their hydrophobicity, which do not exhibit a propensity for particular local backbone conformations, and which may therefore more readily adopt conformations dictated by nonlocal forces (e.g., hydrophobic interactions).

Clustering in the DSSP structural representation

The clustering behavior of the amino acids in the DSSP scheme exhibits a greater emphasis on the delineation between hydrophobic and polar/hydrophilic residues, alongside the segregation of the small amino acids (Table IB). The emergence of the first two groups, the hydrophilic/polar and hydrophobic amino acid sets, reflects the prominence of canonical hydrophobicity patterns in the stability of standard secondary structural elements. Helices exhibit a general tetrameric [H-H-P-P] pattern, whereas extended forms like sheets show an alternating [H-P-H-P] pattern. Residues comprising the third group (Asp, Asn, Gly, Pro, Ser, and Thr) carry the highest preferences for loop or irregular structures and are especially implicated in various turn configurations that bound helices and strands. Further differentiation consists of splitting the latter group into subgroups {Ser, Thr, Asp, Asn}, and {Gly, Pro} and distinguishing between the generic hydrophobic and hydrophilic amino acids. It is important to note that Gly and Pro, although shown to have dramatic and distinct influences on virtual backbone conformation, persist in clustering together in the DSSP scheme. We account for this observation by noting that neither amino acid resides particularly comfortably in standard helices or strands. As one would expect, the basis for clustering in the DSSP structural description is the role of each amino acid in the propagation or disruption of the three secondary struc-

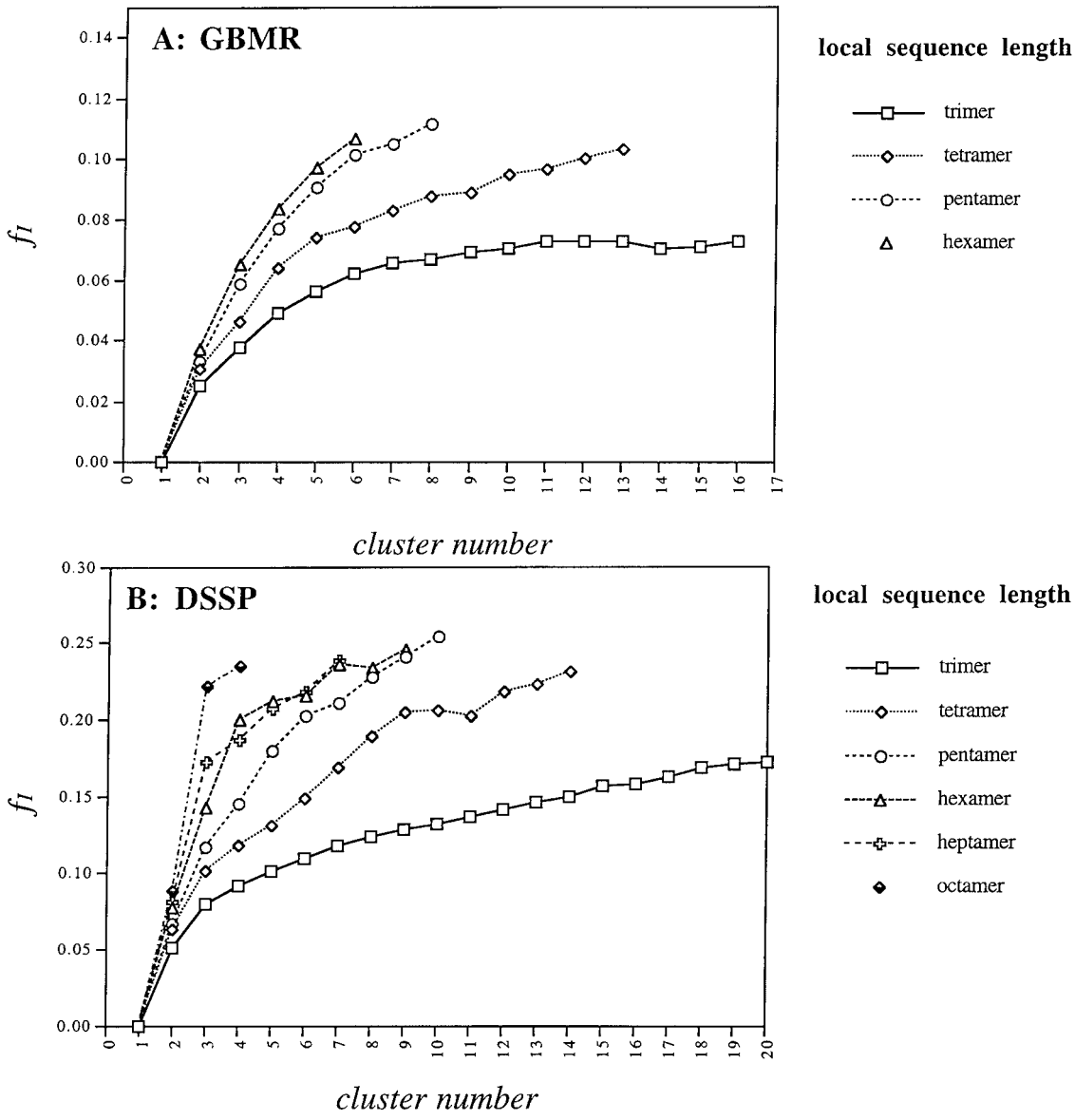


Fig. 7. Plots of f_i for (**A**) GBMR and (**B**) DSSP structural representations. These f_i were computed as a function of the number of amino acid clusters, as well as sequence length scale.

tures, rather than the actual structural properties of the amino acid.

Detection of locally critical residues in the GBMR representation

The results above demonstrate that the backbone structural descriptor chosen determines the quality of structural information that may be derived from local sequence. For the GBMR structural representation, the operative factor in the amino acid clustering is the similarity in the fine structure of the alpha-carbon backbone encoded by the local sequence. Using a purely geometric description of the backbone facilitates the detection of unique effects of Gly, Pro, Ser, Thr, Asp, Asn, and to a lesser extent Cys, His, Trp, and Tyr (Table IA) on the local structures of those sequence segments in which they are found. These results assist in our understanding of the determinants of protein structure in a number of ways. Foremost, recognizing the influence of these locally critical residues on local conformation facilitates the extraction of more information from sequence. These results also suggest that some key amino acids, because of their unique structural propensities, may be necessary agents in the successful folding and stability of real proteins. It may be that sequences that have relatively strong structural propensities act as nucleation sites of folding.

On the other hand, because the three-state DSSP structural description is coarse, details of the relationship

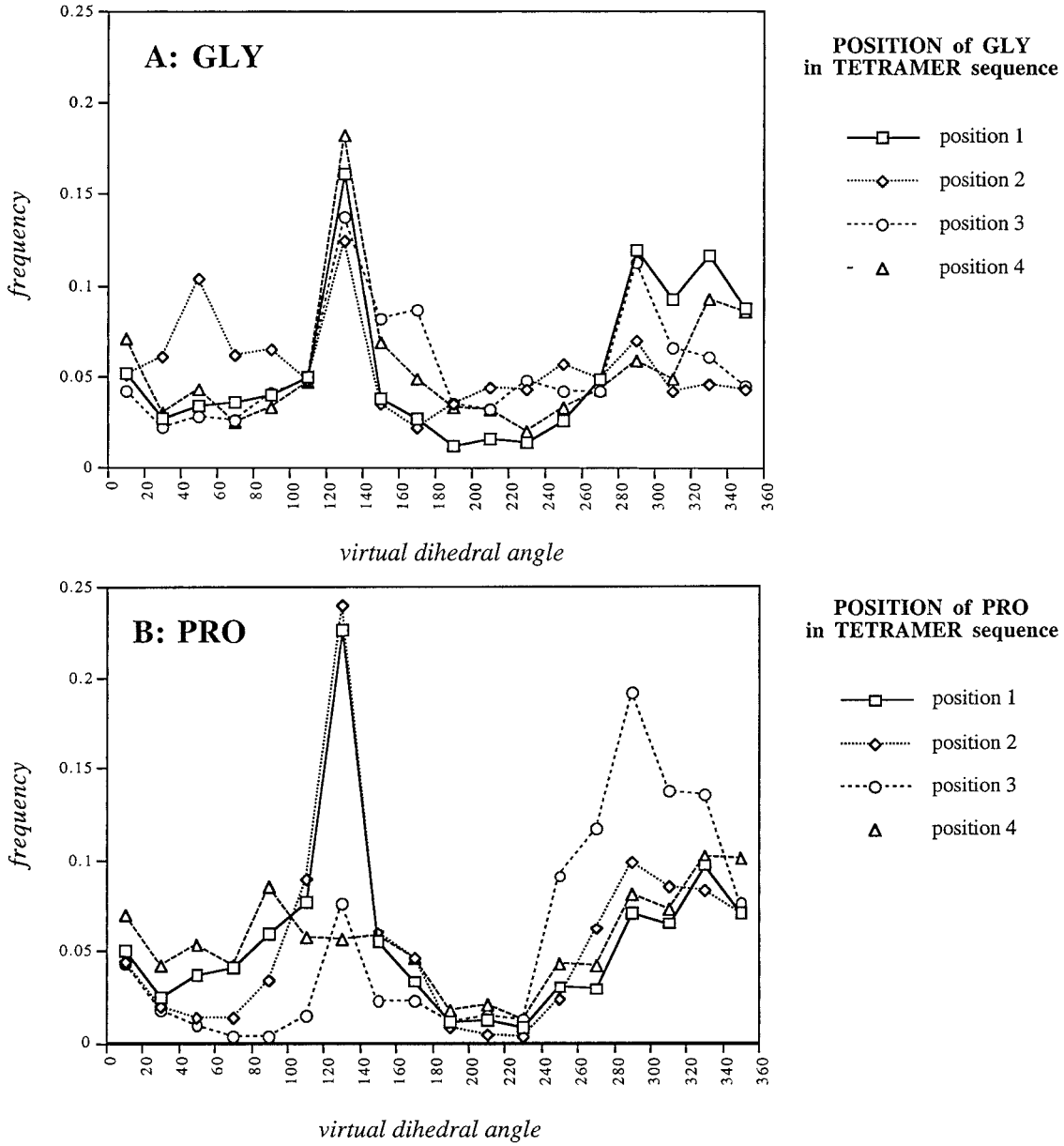


Fig. 8. Dihedral angle profiles of tetramer sequences containing (**A**) glycine or (**B**) proline in each of four possible positions.

between sequence and structure are obscured, resulting in an amino acid clustering based on the most general properties of the three secondary structural types. These are dominated by hydrophobicity patterns necessary for the formation of secondary structural elements.

Recently, there have been experiments that seek to delineate those aspects of protein sequence that are absolutely essential to preserving the fold.[49,50] These studies conclude that although most positions in a sequence can tolerate substitution by an amino acid from a smaller representative amino acid set (or alphabet), some key amino acids generally resist any change without significantly compromising the ability and/or stability to fold. Gly and Pro were invariably found to be irreducible, along with Asp and Asn (kept because they also participate in enzyme activity), Ser, Ala, and representative hydrophobic (Leu and Ile) and polar (Glu, Gln, and Lys) residues. Similarly, in the simplification of the sequence code of 4-helix bundles, the binary HP code, successfully modeling the correct sequence pattern of the helix arms, had to be supplemented by specifically designed linker sequences of locally critical residues like Gly, Pro, Ser, and Asp residues to produce the turns.[51] These and other experimental observations confirm the importance of locally critical residues identified here, and raise the point that simplistic models relying on two general kinds of amino acids (HP)

may not be sufficient to simulate the folding behavior of real proteins.[52–54]

CONCLUSION

Using information-theoretic concepts, we have developed a novel scheme of data compression that determines the maximum amount of structural information available in local protein sequences. We draw the following conclusions:

1. *On the quantity of structural information encoded in local sequence.* The amount of structural information that can be extracted from local sequence is dependent on the kind of structural description used. We find that potentially at least 10% of the information encoded in the GBMR alpha-carbon backbone conformation (0.41–0.53 nats for information gain I_g^{max}over 4.08 nats before any sequence information, Table 3) and 23% for the DSSP conformation (0.24–0.27 nats) may be extracted using knowledge of the local tetramer sequence alone. Although the full fraction is not accessible given the current size of the experimental structure data set, it is already possible to extract at least 50% of this by a collapse of the amino acid alphabet into six or seven clusters (Table II). The effect of database size on this result has been investigated (Table II).
2. *On sequence representations.* Collapsing the amino acid alphabet, by recognizing strong similarities in the structural characteristics of amino acids, is a viable strategy for analyzing the relationship between local sequence and backbone structure. This collapse partially alleviates problems arising from the use of a limited data set of protein structures to derive empirical relationships. For example, reduction of alphabet size from 20 to 6 or 7 reduces the structural information found in local sequences by only half, from 0.41– 0.53 nats (Table III; tetramer length scale, GBMR) to 0.24 nats (Table II). The reduction in the number of unique local sequences, however, is dramatic: from 160,000 tetramers in the 20-letter alphabet to only 1,296 tetramers using the six reduced amino acid sets derived here.
3. *On structural representations.* The choice of structural descriptor affects the quality of structural information that may be derived from local sequence. Use of three-state secondary structural assignment like DSSP captures the general propensities of local sequences to occur in secondary structure elements (sheet, helices, and turns) and places a primary emphasis on the hydrophobicity pattern of the amino acid sequence. On the other hand, use of alpha-carbon backbone geometries to represent secondary structure, as exemplified by GBMR, detects finer conformational nuances encoded in local sequences, leading to a greater information gain. Through GBMR (but not DSSP) the importance of locally critical amino acids and of the local sequences in which they participate is recognized. The placement of these amino acids imparts specificity to the folding process as seen in the formation of necessary microdomains such as helical caps and turns.

ACKNOWLEDGMENTS

We are grateful for support of A.D.S. by the Robert Wood Johnson Pharmaceutical Research Institute and by Smith-Kline Beecham Pharmaceutical Corporation.

REFERENCES

1. Shannon CE. A mathematical theory of communication. Bell Syst Tech J 1948;27:379–423.
2. Rackovsky S. Quantitative organization of the known protein x-ray structures. I. methods and short-length-scale results. Prot Struct Funct Genet 1990;7:378–402.
3. Ramachandran GN, Sasisekharan V. Conformation of polypeptides and proteins. Adv Prot Chem 1968;23:283–437.
4. Kabsch W, Sander C. Dictionary of rotein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
5. Barton GJ. Protein secondary structure prediction. Curr Opin Struct Biol 1995;5:372–376.
6. Robson B. Analysis of the code relating sequence to conformation in globular proteins. Theory and application of expected information. Biochem J 1974;141:853–867.
7. Robson B, Pain RH. Analysis of the code relating sequence to conformation in proteins: possible implications for the mechanism of formation of helical regions. J Mol Biol 1971;58:237–259.
8. Robson B, Pain RH. Analysis of the code relating sequence to conformation in globular proteins: development of a stereochemical alphabet on the basis of intra-residue information. Biochem J 1974;141:869–882.
9. Robson B, Pain RH. Analysis of the code relating sequence to conformation in globular proteins: an informational analysis of the role of the residue in determining the conformation of its neighbours in the primary sequence. Biochem J 1974;141:883–897.
10. Robson B, Pain RH. Analysis of the code relating sequence to conformation in globular proteins: the distribution of residue pairs in turns and kinks in the backbone chain. Biochem J 1974;141:899–904.
11. Robson B, Suzuki E. Conformational properties of amino acid residues in globular proteins. J Mol Biol 1976;107:327–356.
12. Garnier J, Osguthorpe DJ, Robson B. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. J Mol Biol 1978;120:97–120.
13. Gibrat JF, Garnier J, Robson B. Further developments of protein secondary structure prediction using information theory: new parameters and consideration of residue pairs. J Mol Biol 1987;198:425–443.
14. Garnier J, Gibrat JF, Robson B. GOR method for predicting protein secondary structure from amino acid sequence. Methods Enzymol 1996;266:540–553.
15. Rost B, Sander C. Combining evolutionary information and neural networks to predict protein secondary structure. Prot Struct Funct Genet 1994;19:55–72.
16. Zvelebil MJJM, Barton GJ, Taylor WR, Sternberg MJE. Prediction of protein secondary structure and active sites using the alignment of homologous sequences. J Mol Biol 1987;195:957–961.
17. Russell RB, Barton GJ. The limits of protein secondary structure prediction accuracy from multiple sequence alignment. J Mol Biol 1993;234:951–957.
18. Rackovsky S. On the nature of the protein folding code. Proc Natl Acad Sci 1993;90:644–648.
19. Shakhnovich EI. Theoretical studies of protein-folding thermodynamics and kinetics. Curr Opin Struct Biol 1997;7:29–40.
20. Yue K, Dill KA. Forces of tertiary structural organization in globular proteins. Proc Natl Acad Sci 1995;92:146–150.
21. Henikoff S, Henikoff JG. Amino acid substitution from protein blocks. Proc Natl Acad Sci 1992;89:10915–10919.
22. Wako H, Blundell TL. Use of amino acid environment-dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologous proteins. I. Solvent accessibility classes. J Mol Biol 1994;238:682–692.
23. Wako H, Blundell TL. Use of amino acid environment-dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologous proteins. II. Secondary structures. J Mol Biol 1994;238:693–708.
24. Thompson MJ, Goldstein RA. Constructing amino acid residue

substitution classes maximally indicative of local protein structure. Prot Struct Funct Genet 1996;25:28–37.
25. Rackovsky S. On the sxistence and implications of an inverse folding code in proteins. Proc Natl Acad Sci 1995;92:6861–6863.
26. Rackovsky S, Scheraga HA. Differential geometry and protein folding. Acc Chem Res 1984;17:209–214.
27. DeWitte RS, Shakhnovich EI. Psudodihedrals: Simplified protein backbone representation with knowledge-based energy. Prot Sci 1994;3:1570–1581.
28. Oldfield TJ, Hubbard RE. Analysis of Cα geometry in protein structures. Prot Struct Funct Genet 1994;18:324–337.
29. Taylor WR. The classification of amino acid conservation. J Theor Biol 1986;119:205–218.
30. Rufino SD, Donate LE, Canard LHJ. Predicting the conformational class of short and medium size loops connecting regular secondary structures: application to comparative modelling. J Mol Biol 1997;267:352–367.
31. Gibrat JF, Robson B, Garnier J. Influence of the local amino acid sequence upon the zones of the torsional angles phi-psi adopted by residues in proteins. Biochemistry 1991;30:1578–1586.
32. Unger R, Sussman JL. The importance of short structural motifs in protein structure analysis. J Comput Aided Mol Des 1993;7:457–472.
33. 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 1984;81: 1075–1078.
34. Argos P. Analysis of sequence-similar pentapeptides in unrelated protein teritary structures. J Mol Biol 1987;197:331–348.
35. Mezei M. Chameleon sequences in the PDB. Prot Eng 1998;11:411–414.
36. Mocz G. Fuzzy cluster analysis of simple physicochemical properties of amino acids for recognizing secondary structure in proteins. Prot Sci 1995;4:1178–1187.
37. Stanfel LE. A new approach to clustering the amino acids. J Theor Biol 1996;183:195–205.
38. Naor D, Fischer D, Jernigan RL, Wolfson HJ, Nussinov R. Amino acid pair interchanges at spatially conserved location. J Mol Biol 1996;256:924–938.
39. MacArthur MW, Thornton JM. Influence of proline residues on protein conformation. J Mol Biol 1991;218:397–412.
40. Baker EN, Hubbard RE. Hydrogen bonding in globular proteins. Prog Biophys Mol Biol 1984;44:97–179.
41. Richardson JS, Richardson DC. Amino acid preferences for specific locations at the ends of alpha-helices. Science 1988;240:1648–1652.
42. Wilmot CM, Thornton JM. Beta-turns and their distortions: a proposed new nomenclature. Prot Eng 1990;3:479–493.
43. Ring CS, Kneller DG, Langridge R, Cohen FE. Taxonomy and conformational analysis of loops in proteins. J Mol Biol 1992;224: 685–699.
44. Kwasigroch J-M, Chomilier J, Mornon J-P. A global taxonomy of loops in globular proteins. J Mol Biol 1996;259:855–872.
45. Yang A-S, Hitz B, Honig B. Free energy determinants of secondary structure formation. III. β -Turns and their role in protein folding. J Mol Biol 1996;259:873–882.
46. Wilmot CM, Thornton JM. Analysis and prediction of the different types of β-turn in proteins. J Mol Biol 1988;203:221–232.
47. Chakrabartty A, Doig AJ, Baldwin RL. Helix capping propensities in peptides parallel those in proteins. Proc Natl Acad Sci USA 1993;90:11332–11336.
48. Harper ET, Rose GD. Helix stop signals in proteins and peptides: the capping box. Biochemistry 1993;32:7605–7609.
49. Riddle DS, Santiago JV, Bray-Hall ST, Doshi N, Grantcharova VP, Yi Q, Baker D. Functional rapidly folding proteins from simplified amino acid sequences. Nat Struct Biol 1997;4:805–809.
50. Schafmeister CE, LaPorte SL, Miercke LJW, Stroud RM. A designed four helix bundle protein with native-like structure. Nat Stuct Biol 1997;4:1039–1046.
51. Kamtekar S, Schiffer JM, Xiong H, Babik JM, Hecht MH. Protein design by binary patterning of polar and nonpolar amino acids. Science 1993;262:1680–1685.
52. Kolinski A, Madziar P. Collapse transitions in protein-like lattice polymers: the effect of sequence patterns. Biopolymers 1997;42: 537–548.
53. Wolynes PG. As simple as can be? Nat Struct Biol 1997;4:871–874.
54. Honig B, Cohen FE. Adding backbone to protein folding: why proteins are polypeptides. Fold Design 1996;1:R17–R20.
55. Sippl MJ. Calculation of conformational ensembles from potentials of mean force: an approach to the knowledge-based prediction of local structures in globular proteins. J Mol Biol 1990;213:859–883.
56. Goldman S. Information theory. New York: Dover Publications, Inc; 1968. 385 p.

APPENDIX

Further Notes on Mutual Information and Information Gain

Advantages of information gain over mutual information

The Shannon mutual information

$$I(X;Y) = H(Y) - H(Y|X)$$
$$= E[IX = x;Y)] = \Sigma_x[H(Y) - H(Y|X = x)]\, p_X(x) \quad \text{(A1)}$$

measures the expected reduction in uncertainty in one variable caused by the knowledge of another. (For a detailed derivation and discussion, please refer to Refs. 1 and 55.) The concept of mutual information has been used with modest success to detect informative sequence-structure patterns in proteins from multiple sequence alignments and to relate these to secondary structures and solvent accessibility.[24] Although it is a powerful tool, its essential requirement is a complete knowledge of *exact* probability distributions of both the sequence and the structure random variables X and Y. If frequencies are to be used as estimates for probabilities, the structural data set should be large enough for a sufficiently accurate analysis. This is not achievable in many applications, even when the most current structural databases are used. The size of the data set will become especially crucial as the length of the local sequence fragments used to predict backbone structure is increased to include more and more neighboring residues surrounding the site of analysis and prediction.

We have formulated an alternative measure, the information gain (Eq. 3), which compares $H(Y|X = x)$ with the *expected* entropy of a *random* set of structures of the same size picked from the data set. Two important properties of $I_g(X = x,Y)$ make it especially useful: (a) $I_g(X = x,Y)$ is a definite maximum when $H(Y|X = x) = 0$, i.e., if all the occurrences of the specified sequence exhibit a single conformation k, then $p_k = 1$ and $H(Y|X = x) = 0$, reducing Eq. 2 into $I_g(X = x,Y) = E[H(Y|N = n)]$. (b) $I_g(X = x,Y)$ resists the ubiquitous problem of complete memorization (see below and Fig. 4). The general behavior of $E[H(Y|N = n_Y)]$ is illustrated in Figure 9. Notice that as n increases, $E[H(Y|N = n_Y)] \rightarrow H(Y)$. This is a consequence of making frequencies stronger estimates of probabilities. Equally important, if the number of observations n is low for a particular state, the expected structural entropy of the random set is greatly lowered. The effect of a sparsely occurring

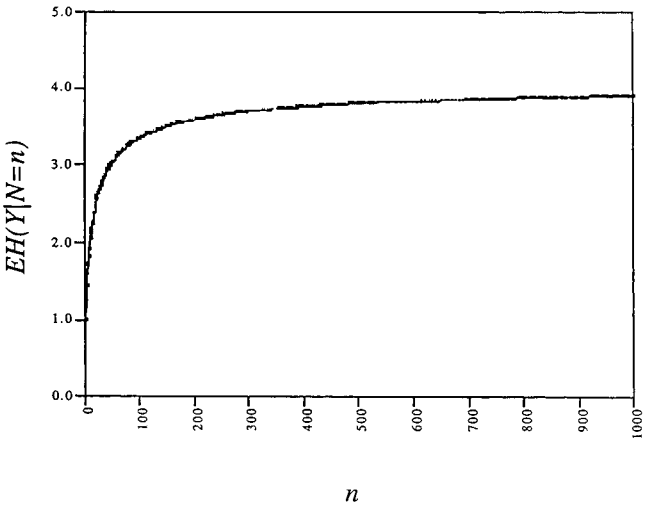


Fig. 9. The behavior of $EH(Y|N = n)$ as n becomes large. A total of 10,000 random distributions were generated from the full data set for each n. The structural entropies for each of the random distributions were computed as in Eq. 1, and means were calculated for each n. The structural entropy $H(Y)$ of the entire data set is 4.08 nats for the GBMR structural representation, which is asymptotically approached as n becomes large.

local sequence on its information gain is thus properly weighed.

The average information gain is obtained by weighing each sequence-specific $I_g(X = x,Y)$ with the probability distribution function for the occurrence of each unique local sequence (Eq. 3). Referring to Eq. 3b, notice that as the size of the data set approaches infinity, $EH(Y|N)$ approaches $H(Y)$, $H(Y|X)$ approaches its true value, and thus information gain $I_g(X,Y)$ approaches Shannon's mutual information $I(X;Y)$.

Maximum information gain and complete memorization

The maximized information gain (Eq. 4) is analogous to another information-theoretic quantity derived by Shannon[1] called channel capacity $C = max\{I(X;Y)\}$, referring to the maximum volume of information a communications channel may carry from information source to its destination. However, Shannon's channel capacity is applicable only when the probability distribution functions of X, Y, and (X,Y) are known completely. Both quantities I_g^{max} *and* C provide an estimate of the upper limit on the amount of information available when considering the random variables X and Y simultaneously.

The difference between Eqs. 3 and A1 becomes readily apparent upon their application in a maximization problem. Mutual information (Eq. A1), in a maximization problem, tends to favor sparse groupings given a limited data set. In the extreme case, groups with just one observation, according to Eq. A2, apparently reduce the uncertainty $H(Y|X = x)$ to zero (refer to Fig. 4). A simple calculation shows that a grouping with just one observation for some set of conditions ($p_m = 1$, $p_{I<>m} = 0$) gives an apparent entropy of $H(Y|X = x) = -1\ ln\ 1 = 0$, bringing about a large gain in "information" of $I(X = x;Y) = H(Y) - 0 = H(Y)$ (from Eq. 3). Thus, in a maximization algorithm, such an objective function will favor partitions of the sequence-structure data set composed of very small numbers of observations, which may mask any real correlations between local sequence and backbone conformation. This phenomenon of complete memorization is encountered in some applications of mutual information, arising from an insufficient sample size.[31,56]