Optimally informative backbone structural propensities in proteins

ARTICLE *in* PROTEINS STRUCTURE FUNCTION AND BIOINFORMATICS · AUGUST 2002

Impact Factor: 2.63 · DOI: 10.1002/prot.10126 · Source: PubMed

2 AUTHORS:

Armando Solis
New York City College of Technology
10 PUBLICATIONS **137** CITATIONS
SEE PROFILE

Shalom Rackovsky
Icahn School of Medicine at Mount Sinai
87 PUBLICATIONS **2,075** CITATIONS
SEE PROFILE

PROTEINS: Structure, Function, and Genetics 48:463–486 (2002)

Optimally Informative Backbone Structural Propensities in Proteins

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

ABSTRACT **We use basic ideas from information theory to extract the maximum amount of structural information available in protein sequence data. From a non-redundant set of protein X-ray structures, we construct local-sequence-dependent [ϕ,ψ] distributions that summarize the influence of local sequence on backbone conformation. These distributions, approximations of actual backbone propensities in the folded protein, have the following properties: (1) They compensate for the problem of scarce data by an optimized combination of local-sequence-dependent and single-residue specific distributions; (2) They use multi-residue information; (3) They exploit similarities in the local coding properties of amino acids by collapsing the amino acid alphabet to streamline local sequence description; (4) They are designed to contain the maximum amount of local structural information the data set allows. Our methodology is able to extract around 30 cnats of information from the protein data set out of a total 387 cnats of initial uncertainty or entropy in a finely discretized [ϕ,ψ] dihedral angle space (18 × 18 structural states), or about 7.8%. This was achieved at the hexamer length scale; shorter as well as longer fragments produce reduced information gains. The automatic clustering of amino acids into groups, a component of the optimization procedure, reveals patterns consistent with their local coding properties. While the overall information gain from local sequence is small, there are some local sequences that have significantly narrower structural distributions than others. Distribution width varies from at least 20% less than the average overall entropy to at least 14% above. This spread is an expression of the influence of local sequence on the conformational propensities of the backbone chain. The optimal ensemble of local-sequence-specific backbone distributions produced is useful as a guide to structural predictions from sequence, as well as a tool for further explorations of the nature of the local protein code. Proteins 2002;48:463–486.** © 2002 Wiley-Liss, Inc.

Key words: sequence–structure relationship; local sequence; maximum information; phi-psi dihedral angle; structural propensities; protein bioinformatics

INTRODUCTION

Local amino acid sequence influences the conformation of the polypeptide backbone in a folded protein.[1] This influence is discernible in patterns seen in the protein X-ray structure database. While short polypeptide sequences take on a variety of conformations depending on sequence context, they show biases for particular sets or ranges of conformation.[2–6] When catalogued properly, these patterns may not only aid searches for the native fold of proteins in computer simulations, but may reveal how much structural information is encoded locally and how this information is distributed along the length of a native sequence.

The problem of sparse data is a fundamental barrier to detecting significant sequence–structure patterns from the protein crystal database.[7] Ordinarily, to infer the backbone conformational propensities of a particular sequence fragment S using the structural database, one identifies the occurrences of S in the database, collating the corresponding conformations in the form of a frequency distribution. One hopes that sequence segment S occurs often enough in the database, in a sampling sufficiently representative of natural propensities, to produce a representative distribution. If S is short and the structural domain is coarsely discretized, a typical data set of 500 protein crystal structures can provide sufficient data to characterize the associated backbone structures. Previous work has elucidated backbone propensities for local sequences this way, using various methods of estimation to compensate for the dearth of experimental data.[2,8–13]

Difficulty arises when longer segment lengths are considered. The 5-mer length scale, for instance, necessitates compiling structural distributions of each of 20^5 or 3.2 million unique pentamers, a significant number of which may not occur in the data set. This results in sparsely populated collections of data for many pentamers, hardly indicative of propensities or "distributions," and devoid of

Grant sponsor: National Library of Medicine of the National Institutes of Health; Grant number: LM06789; Grant sponsor: Robert Wood Johnson Pharmaceutical Research Institute; Grant sponsor: Smith Kline Beecham Pharmaceutical Corporation.

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

Received 3 August 2001; Accepted 1 February 2002

bioinformatic value. The challenge we address in this paper is to extract information latent in the local sequence in the face of limitations brought about by sparse data. We develop a strategy to construct sequence-dependent distributions that, despite limits imposed by the current size of the crystal data base, contain the maximal amount of structural information.

This study advances previous work[6] in which we investigated the factors influencing the efficiency of extracting structural information from the crystal database. In this work, we use basic information-theoretic concepts[14–16] to quantify structural information made available by knowledge of local sequence context. We demonstrated previously that the way amino acid sequence and backbone conformation are represented affects the quantity of structural information available from local sequence. The current work uses those ideas to devise an efficient procedure to construct maximally informative [ϕ,ψ] backbone dihedral angle distributions. Because we use multi-residue information and a finely discretized Ramachandran space, we are able to detect nuances in the local sequence–structure relationship, and therefore use as much structural information as is available in the local sequence.

The resulting probability distributions are designed to best approximate the true structural propensities of short fragments in a folded protein context, given a finite data set. These optimized distributions should be helpful in any application that uses probability estimates. Specifically, the optimized probability distributions may be used to improve (1) measures of local structural propensities,[3,17] (2) potentials of mean force[18,19] and threading score functions,[20,21] (3) procedures for searching the structural space in protein folding simulations,[22–24] and (4) our understanding of the way structural information is encoded locally.[6,25] Future work will explore the specific benefits of using optimized probabilities in these applications.

THEORY AND METHODOLOGY

In order to work with data sets of limited size, it is necessary to compress protein data while minimizing the amount of information loss. A frequently used method is to discretize the structural domain. Discretization is an effective way of reducing complexity while still preserving the essence of the sequence–structure relationship. We have shown previously that the level of structural discretization and the choice of structural descriptor determines the amount of structural information that can be extracted from sequence data.[6]

Another strategy is to compress the sequence. We have also shown previously that the amino acid alphabet can be collapsed in an optimally informative manner. This is consistent with the observation that amino acids exhibit similarities in the way they influence backbone conformation within a folded protein.[26] Not surprisingly, amino acids that are similar in the way they encode local structure are found to cluster together in the automatic optimization procedure.

Hybrid Structural Distributions

Given a finite data set, we wish to develop an improved method of constructing structure distributions that best approximate the true backbone propensities of sequence fragments. A conservative way to form a local sequence-specific structural probability distribution, especially for rare fragments, is to combine a background component and a local sequence-specific component. These two terms can be combined linearly to form the conditional probability of having a specified native backbone conformation given the local sequence:

$$p(C = c|S = s) = \sigma_b \pi(C = c|B) + \sigma_s \pi(C = c|S = s) \quad (1)$$

Here S and C refer to the local amino acid sequence and backbone conformation respectively, s and c are particular instances, B refers to the background distribution, p is a probability, π is an approximation of a probability by a frequency, and σ_b and σ_s are weights that quantify the respective relative contributions of the background and sequence-specific components.

The background component represents the distribution of local structure prior to any additional specification of local sequence. For instance, one can use as background set the distribution of all conformations found in the full data set. Note, however, that other forms of the background can be used. Ultimately, we would like the background to assist in forming the best approximation of natural propensities.

For discrete distributions, the background probabilities can be estimated by frequencies as follows:

$$\pi(C = c|B) = n(C = c|B) / N_B \quad (2)$$

where $n(C{=}c|B)$ is the number of occurrences of the particular local conformation c in the background set and N_B is the total number of observations in the same set.

The second component is the local sequence-specific contribution, made up of those structures from the data set corresponding to the specific local sequence of interest. The component associated with the local sequence $S = s$ is

$$\pi(C = c|S = s) = n(C = c|S = s) / n(S = s) \quad (3)$$

As the number of occurrences $n(S = s)$ in the data set approaches infinity, this equation alone becomes sufficient to describe fully the influence of local sequence on backbone structure. However, with a data set of around 500 protein structures, Eq.(3) is sorely inadequate when used alone. For rarely occurring local sequences, the sparse structural distribution derived from frequencies is uninformative and misleading.

Combining this component with a background distribution helps minimize the effect of poor sampling. If there are no occurrences for a particular local sequence in the data set, we would like the structural distribution characterizing that sequence to be identical to the background distribution. When only a few observations are available, the distribution should be dominated by the background distribution, with only slight changes to reflect the influence manifested in the few instances found in the data set. As more and more instances are found, we want the hybrid distribution to shift away from the background and move closer to the observed sequence-specific distribution, until, as the number of occurrences becomes very large, the

sequence-specific distribution should dominate. This behavior is embodied in the following equation:

$$p(C = c|S = s) = [\gamma\pi(C = c|B) + n(S = s)\pi(C = c|S = s)] / [\gamma + n(S = s)] \quad (4)$$

where γ is the hybrid coefficient, $0 < \gamma < \infty$. Eq.(4) conforms to Eq.(1) if

$$\sigma_b = \gamma / [\gamma + n(S = s)]$$

$$\sigma_s = n(S = s) / [\gamma + n(S = s)].$$

Note that γ is a positive coefficient that determines the relative importance of a single observation in the data set with respect to the background distribution. For example, if $\gamma = 100$, a single observation has a 1% effect on the total distribution p as compared to the background distribution. If $\gamma = 100$ and $n(S = s) = 100$ (i.e., there are 100 observations of local sequence fragment s in the data set), then the background and sequence-specific components have equal weights in forming the total distribution. It is easy to see that as $n(S = s)$ approaches infinity (when all the conditional probabilities become known exactly), $\pi(C = c|S = s)$ dominates, as it should.

There are a number of ways to define the background set and the hybrid coefficient γ. It is, therefore, necessary to evaluate the utility of each candidate with respect to our goal of approximating true conformational propensities of the peptide backbone. Below we describe basic tools from information theory that will help us choose the best set of conditions.

Quantifying Information Via Information Gain

The Shannon entropy provides a means to measure the "spread" of a structural distribution.[14] The Shannon entropy is

$$H(C) = -\sum_i p(C = c_i)\, ln\, p(C = c_i) \quad (5)$$

where $p(C = c_i)$ is the probability associated with structural state $C = c_i$. The basic unit for entropy is the "nat" when the natural logarithm is used. (A smaller unit is the "cnat," or centinat, which is 1/100 of a nat.) Alternatively, the unit "bit" results from using the base-2 logarithm.

When some local sequence information $S = s$ is provided, the Shannon entropy is

$$H(C|S = s) = -\sum_i p(C = c_i|S = s)\, ln\, p(C = c_i|S = s) \quad (6)$$

where $p(C = c_i|S = s)$ is the conditional probability associated with structural state $C = c_i$ given the sequence $S = s$.

These equations adequately describe the entropy of a completely defined distribution. In cases when the probabilities are estimates derived from a finite data set, a re-expression of these equations can assist in approximating these entropies. We note that entropy is defined as the expected value of the self-information,[16] or

$$H(C) = E\,(-ln\, p(C = c_i)) \quad (7a)$$

where the quantity $[-ln\ p(C = c_i)]$ is the so-called self-information, while the conditional entropy is

$$H(C|S = s) = E\,(-ln\, p(C = c_i|S = s)) \quad (7b)$$

If we have only estimates of the probabilities derived from a data set, the entropy can be approximated by averaging the self-information of each of the occurrences of interest in the data set. We include a detailed description of the precise equations used in this work in Appendix A.

In principle, information caused by a particular factor is measured by the difference in the Shannon entropies before and after the factor is considered. This quantity, called information gain, is the decrease in informational entropy or uncertainty measured from a reference state. For instance, to evaluate the conformational information gain resulting from specifying the local sequence $S = s$, one subtracts Eq.(6) from Eq.(5), to find

$$I_g(C|S = s) = H(C) - H(C|S = s) \quad (8a)$$

More generally, information gain can be computed by

$$I_g(C|S = S_0 + s) = H(C|S = S_0) - H(C|S = S_0 + s) \quad (8b)$$

where S_0 is the starting level of sequence knowledge (or reference state), and the information gain $I_g(C|S = S_0{+}s)$ results from additional sequence knowledge s. The average information gain is

$$I_g(S,C) = \sum_j I_g(C|S = S_0 + s_j)\, p(S = S_0 + s_j) \quad (9)$$

where $p(S = S_0{+}s_j)$ is the probability of occurrence of the particular local sequence $S_0{+}s_j$ in the universe of protein sequences (as defined by the data set used).

The ultimate goal of this exercise is to maximize the average information gain:

$$I_g^{max}(S,C) = max\,\{I_g\,(S,C)\} = max\left\{\sum_i [H(C|S = S_0) - H(C|S = S_0 + s_j)]\, p_s(S = S_0 + s_j)\right\} \quad (10)$$

where the conditional entropies are of the form given by the Eq.(6) (or Eq.(7b) in the current application).

Hybrid Coefficient γ

The hybrid coefficient γ is implicit in Eq.(10) because of the relationship given in Eq.(4). When $\gamma = 0$, $p(C = c|S = s) = \pi(C = c|S = s)$, which is generally an unrealistic approximation due to sparse data, while at $\gamma = \infty$, $p(C = c|S = s) = \pi(C = c|B)$, which ignores information from local sequence beyond that specified by the background set. Therefore, there must be an optimum γ between these two extremes, which maximizes information gain. In the process of optimizing local sequence and structure partitions, we should simultaneously maximize information gain with respect to the hybrid coefficient γ.

An important result of this procedure is that, compared to $\pi(C|S = s)$, which can be jagged and sparse, the new distributions $p(C|S = s)$ are smoother and defined every-

where. The difficulty in constructing structural distributions from a limited data set is overcome in a natural way that preserves as much structural information from local sequence as possible. It should be recalled that in previous work, we overcame this problem by using the term $EH(C|N)$, an averaged sequence-independent property of the entire data set, in calculating information quantities.[6,22] Here, we integrate a single-residue-dependent correction into the local sequence-dependent distributions themselves, which makes them more convenient for further use. Furthermore, the objective function (Eq.10) detects cases when the discretization of local sequence and structure become too fine. For instance, one can be overzealous in discretizing the [ϕ,ψ] conformation domain into 360 × 360 regions (to make 129,600 unique structural states for a description with 1° accuracy). Faced with a data set of less than 200,000 residues, such extreme cases result in very few, if any, structural observations per state, simplifying Eq.(4) into approximately

$$p(C = c_i|S = s) = \pi(C = c_i|B)$$

Eq.(6) then becomes

$$H(C|S = s_j) = -\sum_i \pi(C = c_i|B)\, ln\, \pi(C = c_i|B) = H(C|B)$$

This equation says that in such situations, the entropy of the local-sequence-specific distributions $[H(C|S = s_j)]$ is approximately equal to the entropy of the background distribution. Therefore, no information is gained by using local sequence knowledge whenever the level of structural discretization is too fine.

In this work, we use a constant hybrid coefficient (i.e., $\gamma = k$). However, this is not the only way to construct these sequence-dependent distributions. The hybrid coefficient γ can take other forms that take into account the number of observations $n(S = s)$ for the particular local sequence $S = s$. For instance, one can tune out the background distribution as the number of occurrences increases by a linear function

$$\begin{aligned} \gamma(S = s) &= \gamma_0 - \rho\, n(S = s) \quad &if\ n(S = s) < \gamma_0/\rho \text{ or} \\ &= 0 \quad &\text{otherwise} \end{aligned} \tag{11}$$

where ρ is a positive coefficient. This requires that two variables be optimized (γ_0 and ρ). In future work, we will investigate these higher-order functions to discover whether they give a significant increase in information gain.

We note that the hybrid coefficient γ is related to a quantity used by Sippl[2] to form interaction potentials of mean force between amino acid pairs. In his work, Sippl defines a quantity similar to γ, called the information quantum, which anchors a pair-specific distribution with a reference distribution. There are two critical differences.

- The information quantum was limited to backgrounds that are sequence-*in*dependent distributions. Here, we allow use of γ in conjunction with various types of background, including local-sequence-dependent distributions. From these, we choose the ensemble of background distributions which maximizes the amount of structural information revealed by the hybrid distributions.
- We do not limit the value of the hybrid coefficient to a preset value; instead, the value of γ is optimized. We integrate the search for an optimum hybrid coefficient into the global search for maximally informative local-sequence-specific structural distributions.

Local Sequence Description

Given a finite protein data set, structural information encoded in any type of local sequence configuration can be measured using the general optimization equation (Eq.10). Information latent in any local sequence length, specific amino acid clustering (amino acid alphabet reduction), and backbone structural descriptor can be calculated. For instance, we could examine the influence of the trimer sequence unit on the [ϕ,ψ] conformation of the middle residue. We denote the trimer sequence configuration by

$$S_3 = X^{i-1}_{n(i-1)} - X^{i}_{n(i)} - X^{i+1}_{n(i+1)} \tag{12a}$$

where $n(i-1)$, $n(i)$, $n(i+1)$ are the numbers of amino acid clusters (or alphabet size) into which the 20 amino acids are grouped at each position (indicated by $i-1$, i, and $i+1$), and X^k is a specific cluster chosen from the alphabet at position k. The amino acid clustering at each position can be optimized so that the largest possible amount of information about the [ϕ,ψ] dihedral conformation at the ith position is revealed. The amino acid collapse into the particular cluster configuration $n(i-1)$, $n(i)$, $n(i+1)$ reduces the number of unique tripeptides from 8,000 to a more manageable number. For instance, if one collapses the complete 20-letter amino acid alphabet into 10 clusters at each position,

$$S_3 = X^{i-1}_{10} - X^{i}_{10} - X^{i+1}_{10},$$

the number of unique tripeptides is reduced to 1,000. The information gain can then be computed from $I_g^{max}(S=X^{i-1}_{10}\text{-}X^{i}_{10}\text{-}X^{i+1}_{10},C=[\phi,\psi])$ (Eq.10), after an appropriate background distribution is chosen to assist in estimating the necessary probabilities. (There are a number of possible background distributions, a point we explore in Results and Discussion. As the choice of background system affects the information output of the system, this choice has to be optimized as well.)

For ease in discussion, we represent the local sequence configuration as follows: for a particular fragment length m,

$$S_m = X^{i-p}_{n(i-p)} - \cdots - X^{i-1}_{n(i-1)} - X^{i}_{n(i)} - X^{i+1}_{n(i+1)} - \cdots - X^{i+q}_{n(i+q)} \tag{12b}$$

where $m = p+q+1$, p is the number of positions on the amino side of i, q is the number of positions on the carboxyl side, and $n(r)$ is the number of groups of amino acids (alphabet size) at position r. As a convenient notation to specify the alphabet size at each position, we define the cluster configuration index set. This is denoted by $\Omega^m_{b=i}$, where

TABLE I. Optimal Clustering of Amino Acids $\Xi^3_{b=2}$ in the Trimer Configuration $\Omega^3_{b=2}$ = {4.20.4}†

POSITION	i − 1	CLUSTER	1	GLY
POSITION	i − 1	CLUSTER	2	PRO
POSITION	i − 1	CLUSTER	3	CYS PHE ILE LEU MET VAL TRP TYR
POSITION	i − 1	CLUSTER	4	ALA ASP GLU HIS LYS ASN GLN ARG SER THR
POSITION	i	CLUSTER	1	ALA
POSITION	i	CLUSTER	2	ASP
POSITION	i	CLUSTER	3	CYS
POSITION	i	CLUSTER	4	GLU
POSITION	i	CLUSTER	5	PHE
POSITION	i	CLUSTER	6	GLY
POSITION	i	CLUSTER	7	HIS
POSITION	i	CLUSTER	8	ILE
POSITION	i	CLUSTER	9	LYS
POSITION	i	CLUSTER	10	LEU
POSITION	i	CLUSTER	11	MET
POSITION	i	CLUSTER	12	ASN
POSITION	i	CLUSTER	13	PRO
POSITION	i	CLUSTER	14	GLN
POSITION	i	CLUSTER	15	ARG
POSITION	i	CLUSTER	16	SER
POSITION	i	CLUSTER	17	THR
POSITION	i	CLUSTER	18	VAL
POSITION	i	CLUSTER	19	TRP
POSITION	i	CLUSTER	20	TYR
POSITION	i + 1	CLUSTER	1	GLY
POSITION	i + 1	CLUSTER	2	PRO
POSITION	i + 1	CLUSTER	3	CYS PHE ILE LEU MET VAL TRP TYR
POSITION	i + 1	CLUSTER	4	ALA ASP GLU HIS LYS ASN GLN ARG SER THR

†The reference distribution is the single-residue-specific distribution at the middle position (i) of the trimer segment.

$$\Omega^m_{b=i} = \{n(i-p)\ldots n(i-1)\cdot n(i)\cdot n(i+1)\ldots n(i+q)\} \tag{12c}$$

Here b refers to the position of interest within the backbone segment and m is the sequence fragment length. For instance, in a trimer configuration where the first position has alphabet size 7, the second 20, and the third 8, $\Omega^3_{b=2}$ = {7.20.8}. The notation $\Omega^3_{b=2}$ = {20.20.20} indicates that the full amino acid sequence of the trimer is used.

There are many ways to cluster the 20 amino acids at each position of a given index set. We refer to a specific amino acid clustering or reduced sequence alphabet as $\Xi^{\mathrm{m}}_{\mathrm{b}=i}$, where b refers to the position within the segment of the backbone structure of interest and m is the sequence fragment length. For instance, Table I shows a particular clustering $\Xi^3_{b=2}$ for the index set $\Xi^3_{b=2}$ = {4.20.4}. Lastly, we indicate a particular sequence fragment by a_{i-p}-. . . ,-a_{i-1}-a_i-a_{i+l}-· · ·-a_{i+q}, where a_b refers to the particular cluster at position b. For instance, both trimer sequences Cys-Met-Gln and Ile-Met-Ser are reduced to *3-11-4* when the $\Xi^3_{b=2}$ given in Table I is used to collapse the amino acid alphabet. (Note that if the full 20-letter alphabet is used in any position, as in the middle position in Table I, we can refer to the actual amino acid in that position to describe the sequence of a fragment. Thus, we can also refer to the two trimers just mentioned as 3-Met-4).

Calculation of $I_g(S,C)$

1. We begin with a particular sequence configuration, as in Eq.(12b), with index set $\Omega^m_{b=i}$ (Eq.12c), and a backbone conformation descriptor with a specified level of discretization. To illustrate, we might take the trimer sequence configuration S_3 (Eq.12a) with index set $\Omega^3_{b=2}$ = {4.20.5}, and take the [ϕ,ψ] dihedral angles of the middle residue, discretized into 18 × 18 unique structural states, as the structural descriptor of the trimer segment. In addition, we are also provided with a particular background as well as a value of the hybrid coefficient γ.
2. Given the index set $\Omega^3_{b=2}$ = {4.20.5}, there are many ways to arrange the 20 amino acids in 4 and 5 clusters at positions $i-1$ and $i+1$ respectively. (At position i, since each amino acid populates its own cluster, there is only one clustering configuration.) Each particular $\Xi^3_{b=2}$ has a characteristic information gain $I_g(S,C)$. We first discuss the calculation of information gain brought about by one specified $\Xi^3_{b=2}$.
3. The sequences of all overlapping trimers in the proteins of the data set are translated into a given reduced alphabet $\Xi^3_{b=2}$. Each reduced trimer sequence is associated with the backbone conformation of the middle residue.
4. We classify all backbone structures in the data set with

respect to their associated reduced local sequences. In our example, there are $4 \times 20 \times 5 = 400$ unique trimer sequences. We therefore have 400 sets of structures, one for each trimer sequence. Each of these sets gives a value for $\pi(C = c|S = s)$ [to be used in Eq.(4)].

5. The probability $p(C = c|S = s)$ is easily computed via a slight modification of Eq.(4), using the background distribution $\pi(C = c|B)$ and γ, both predetermined (see Appendix A for a description of the equations used). This probability enables us to first calculate the Shannon entropy (Eq.7b) and then the average information gain $I_g(S,C)$ (Eq.9).

There is a value of $I_g(S,C)$ for every amino acid clustering $\Xi^m_{b=i}$, hybrid coefficient γ, local sequence configuration S (Eq.12a), backbone structure definition, and background distribution. The maximization of information gain is therefore a multi-dimensional search, and can be carried out as a function of some or all of these parameters.

Search Algorithm for Finding Maximum Information Gain $I_g^{\max}(S,C)$ and Its Associated Optimal Reduced Alphabet $\Xi^m_{b=i}$ and Hybrid Coefficient γ

Our optimization procedure involves a search for maximum information gain as a function of amino acid alphabet membership $\Xi^{\mathrm{m}}_{b=\mathrm{i}}$ and value of γ, for a chosen (fixed) local sequence configuration (fragment length m and amino acid cluster configuration $\Omega^m_{b=i}$), backbone structure descriptor and level of discretization, and background distribution B. First, we maximize information gain by manipulating amino acid membership $\Xi^m_{b=i}$ within the predetermined cluster configuration $\Omega^m_{b=i}$ for a particular γ. We then repeat the process for different values of γ in order to find the γ and $\Xi^m_{b=i}$, which simultaneously optimize information gain.

Since an exhaustive search for the optimal $\Xi^m_{b=i}$ is not feasible, we use a Monte Carlo algorithm with an empirically chosen sampling procedure and decision criterion. The complete search cycle has 5 parts:

1. An initial amino acid clustering $\Xi^m_{b=i}$ is randomly generated. The information gain is computed from the resulting structural distributions using Eq.(10).
2. To generate the next trial clustering, $\Xi^m_{b=i}$ is altered by a random change in membership for one, two, three, or four (or more) amino acids simultaneously. The number of amino acids involved in the change is randomly chosen, with a preset sampling frequency. In this work, a typical sampling frequency of {0.4, 0.3, 0.2, 0.1} was used for 1, 2, 3, and 4 membership changes. This sampling frequency may be altered for more complex searches (e.g., longer sequence fragment lengths, higher cluster numbers, etc.).
3. The information gain given by this trial clustering is computed, and compared to the old value. If the new value is higher, the trial clustering is kept, and another iteration is made from the Step 2. Otherwise, the old clustering is kept.
4. When no trial $\Xi^m_{b=i}$ is accepted after 1,000 trials, the criterion for acceptance is relieved slightly, by accepting any trial grouping with an information gain that is greater than a given function of the old value. The fraction can vary depending on the complexity of the search; for typical searches, a fraction of 0.97–0.99 is adequate. Choosing to ease the acceptance criterion is done randomly, and only 10% of the time.
5. If after 2,000 iterations the absolute value of the maximum information gain does not increase further, the algorithm is restarted with a new random amino acid grouping.

The Monte Carlo search is terminated when the same optimized reduced alphabet is achieved by at least 10 randomly chosen starting amino acid groupings. This procedure, undertaken for a particular value of γ, gives the optimal amino acid clustering that maximizes $I_g(S,C)$, resulting in $I_g^{\max}(S,C)$. The next level of the search procedure is accomplished by going through the 5-part cycle for different γ's, comparing the resulting $I_g^{max}(S,C)$, and then choosing the maximum among them. This search finds the parameters (amino acid clustering $\Xi^m_{b=i}$ and γ) which give the maximum information output $I_g^{max}(S,C)$ of the system $\Omega^m_{b=i}$. Note that the hybrid coefficient γ can take on an infinite number of values. In order to make the search feasible, we discretize the γ-domain.

It should be remarked that we are carrying out a 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 criterion), and can be designed and altered with great flexibility. For instance, a wider search scheme (e.g., longer segment lengths) may necessitate a modified sampling frequency in Step 2 to allow larger changes in amino acid clustering and more efficiently overcome local minima in the objective function.

Comparison With Previous Work

We previously investigated the information gain brought about by the knowledge of local sequence.[6] In this paper, we improve our methodology in order to construct local sequence-dependent structural distributions for rare sequence fragments that are smooth and defined everywhere. Moreover, information gain is measured without resorting to a statistical term $H(Y|N = n)$, which is the average entropy of an ensemble of n structures randomly collected from the universe of structures [see Eq.(2) of ref. 6]. Instead, we use a more standard equation for information gain (Eq.8). In addition, clustering of amino acids at each position within a fragment is done independently, which, we find, increases information gain.

RESULTS AND DISCUSSION

The tripeptide sequence configuration (Eq. 12a) was chosen to illustrate the capabilities of our procedure and to analyze general amino acid clustering patterns resulting from the optimization of information gain. We use [ϕ,ψ] dihedral angles as the structural descriptor. The [ϕ,ψ] plane is discretized as an 18×18 grid, with grid boundaries positioned at the following angles for both phi and psi:

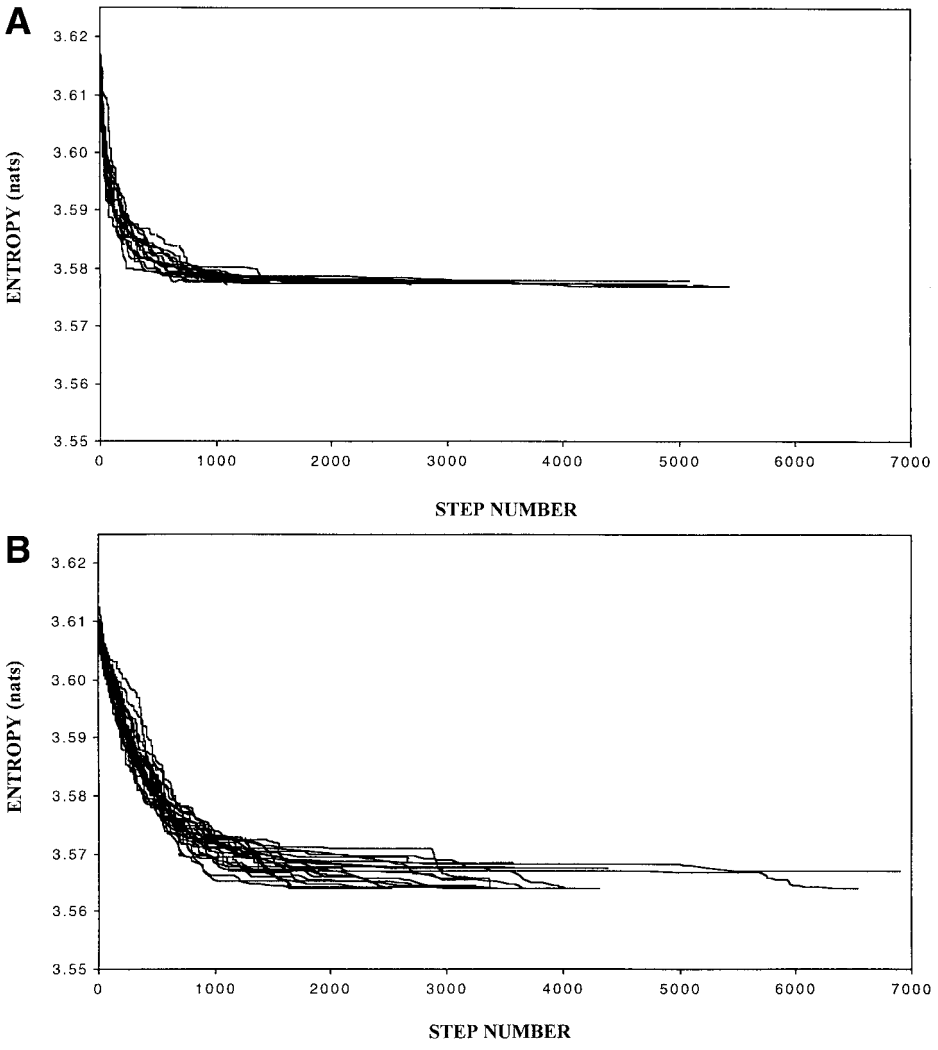


Fig. 1. The course of optimization of amino acid clustering $\Xi^{3}_{b=2}$ from different randomly chosen starting points. **A:** Fifteen distinct starting positions and their respective courses in the optimization algorithm for trimer $\Omega^{3}_{b=2}$ = {7.20.7}, γ = 475, with single-residue-specific distributions as background. **B:** The course of optimization (from 25 distinct starting points) at the pentamer length scale with the following conditions: $\Omega^{5}_{b=3}$ = {4.6.20.6.4} at γ = 275, with single-residue-specific distributions as background.

$n\pi/9$, $0<n<18$, giving 324 unique backbone structural states. In later sections, other segment lengths beyond the trimer configuration are investigated and discussed.

Construction of the data set of protein X-ray structures used in this work is based on the algorithm for generating a representative list of protein chains (of varying sequence homologies) by Hobohm and colleagues.[27,28] We use the protein list generated on May 6, 1997 containing proteins up to 25% sequence-homologous. This list of 536 proteins included in the data set can be obtained from the authors via e-mail request.

Search Procedure Is Able to Find the Optimum Amino Acid Clustering $\Xi^{m}_{b=i}$

The progress of the amino acid clustering procedure at the trimer level is illustrated in Figure 1(A), which plots entropy as a function of step number. (A step is made when the amino acid clustering $\Xi^{m}_{b=i}$ is randomly altered by changing the cluster membership slightly, as described in Materials and Methods.) The conditions for this particular optimization example are: $\Omega^{3}_{b=2}$ = {7.20.7}, γ = 475, background = single-residue distributions (the choice of background is discussed in a later section).

The number of different initial amino acid clustering $\Xi^{m}_{b=i}$ sampled for a trial run of 200,000 steps was 28, an average of more than 7,000 steps per complete cycle, where a cycle is begun by an initial randomly generated $\Xi^{m}_{b=i}$. Out of the 28 different descents, the maximum information gain was reached three times (success rate of 10.7%). Figure 1(A) shows a representative 15 starting positions and their respective courses in the optimization algorithm. A rapid descent within the first 1,000 steps after the initial clustering configuration is followed by a correction phase where minute decreases in entropy (and, therefore, increases in information gain) are achieved with small rearrangements in the amino

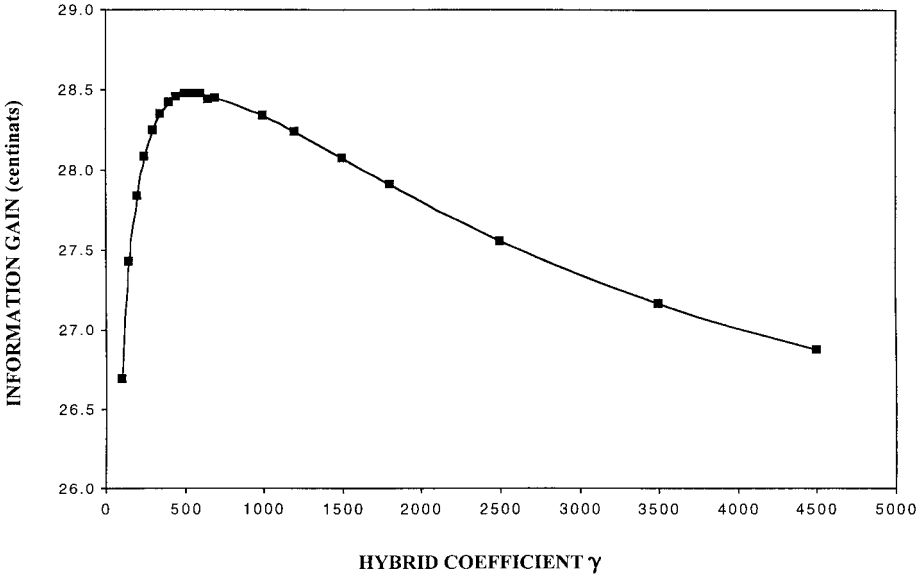


Fig. 2. The maximum information gain from local sequence at cluster configuration $\Omega^3_{b=2}$ = {4.20.4} for various hybrid coefficients γ.

acid grouping. A majority of local minima were found within 5,000 steps.

In Figure 1(B), we illustrate the behavior of the optimization procedure for a longer local sequence, the pentamer, with the following conditions: $\Omega^5_{b=3}$ = {4.6.20.6.4} at γ = 275. A typical run of 400,000 steps allows the sampling of 45 different initial $\Xi^5_{b=3}$'s. Out of these 45 independent cycles, the maximum information gain was reached three times (6.7% success rate). Figure 1(B) shows a representative 25 starting points and their paths towards their respective local minima, most of which are found within 5,000 to 6,000 steps. The optimization search at the pentamer level differs from that at the shorter trimer level in the range of local entropy minima and the speed and efficiency with which they are found. This is expected, since there are many more ways to cluster the 20 amino acids in the pentamer configuration $\Omega^5_{b=3}$ = {4.6.20.6.4} than there are in the trimer configuration $\Omega^3_{b=2}$ = {7.20.7}.

Because the search algorithm is heuristic, finding the global extremum is never guaranteed. In order to ensure confidence in the optimization result, we implement the procedure as many times as it takes to confirm each result from at least 10 different starting amino acid cluster configurations. In the case of the trimer configuration, 92 randomly selected starting amino acid cluster configurations were needed to meet this requirement, while the pentamer configuration necessitated 128 complete cycles.

An Optimal Hybrid Coefficient γ Exists That Maximizes the Information Output of the Backbone Structural Distributions

We have postulated that an optimum hybrid coefficient exists (in the range of $0 < \gamma < \infty$) for combining the background distribution with the sequence-dependent distribution to produce the most informative ensemble of structural propensities. A hybrid coefficient γ = 0 means the resulting distributions are composed purely of the sequence-dependent component, while a high positive γ means that the sequence-dependent component is virtually disregarded in favor of the background distribution.

In order to simplify the search for the optimum γ and accomplish the optimization in reasonable computation time, the search space for γ was discretized. We find that information gain does not vary greatly as γ is varied within a range of γ / 4 to γ / 10. Therefore, γ, which typically fall into the range $100 < \gamma < 1{,}000$ in this [ϕ,ψ] system, was discretized into values separated by at most 50.

The optimization at cluster configuration $\Omega^3_{b=2}$ = {4.20.4} over a range of γ's exhibits a characteristic pattern (Fig. 2). The function is peaked around one optimum γ, which in this particular cluster configuration is 600. At this optimal γ, the clustering of the amino acids results in the separation of Gly and Pro, as well as a distinction among the rest of the amino acids in terms of polarity. These patterns are analyzed further in later sections.

Backbone Structural Distributions Specific to Each of the 20 Amino Acids Are the Optimal Set of Background Distributions

Our goal to maximize information gain is equivalent to finding the ensemble of structural distributions that best approximate the true structural propensities of peptide segments in a folded protein. This is achieved in part by a well-chosen background distribution. Among the many options for the background, we focus on three in particular: (1) a uniform, equiprobable distribution, (2) a sequence-independent (universal) distribution, and (3) a distribution specific to the particular amino acid at the position *i* (or single-residue specific distribution), irrespective of sequence environment.

The naive choice for background distribution is the uniform distribution. Such a choice assumes that prior to any local sequence information, the best approximation can be built from a distribution in which every structural

state has an equal probability of occurrence. This is equivalent to having a constant background probability in Eq.(4)

$$\pi(C = c|B) = 1/n_{str} \tag{13}$$

where n_{str} is the number of structural states.

A more sophisticated choice for background is the sequence-independent distribution, which is just the distribution of [ϕ,ψ] given by the entire structural database (or the universe of local structures):

$$\pi(C = c|B) = \pi(C = c) \tag{14}$$

The universe of protein structures, which characterizes this particular $\pi(C = c|B)$, is shown in Figure 3(A).

The third option is to use a distribution specified by the identity of the amino acid at the backbone site of interest, or $\Xi^i_{n(i)=20}$. That is,

$$\pi(C = c|B) = \pi(C = c|S = X^i_{n(i)=20}) \tag{15}$$

In this case, there are 20 different background distributions, one for each amino acid. For a given sequence fragment, the amino acid identity of the residue at position i determines which background distribution to use in constructing its backbone structure distribution. For instance, if one wants to derive the distribution of the middle [ϕ,ψ] dihedral angle for the trimer Gly-Pro-Ser, we use as background the probability distribution of [ϕ,ψ] angles specific to proline, or $\pi(C = c|X = \mathrm{Pro})$. Figure 3(B–E) shows the single-residue specific background distributions for four amino acids Gly, Pro, Asp, and Ile.

Figure 4 illustrates the difference in the average information gains characterizing the hybrid distributions derived from the three kinds of background. The maxima occur at 16.3, 24.1, and 29.0 cnats when the uniform, sequence-independent, and single-residue distributions are used respectively, clearly proving that use of single-residue background produces the most informative hybrid distributions. (We describe a way to understand the relative magnitudes involved in these informatic units in Appendix B).

We see that when the uniform distribution is used as background, the entropy of the resulting hybrid distributions in specifying the backbone conformation remains high. This is because, for reasonable values of γ, such hybrid distributions, because of the highly entropic uniform distribution component, assign significantly high probabilities to regions of the Ramachandran space not normally populated by natural polypeptide chains. Specifically, only three major areas of the [ϕ,ψ] space are populated in significant proportions, outside of which steric clashes limit the occurrence of such backbone geometries [Fig. 3(A)]. Even glycine, which lacks a side chain, occupies only a fraction of the Ramachandran space [Fig. 3(B)]. Ignoring these patterns, by assigning equal prior probabilities to physically possible as well as improbable states, makes the hybrid distributions flat and more entropic. Such an ensemble of distributions is a poor approximation to the true structural propensities, and therefore results in the observed lower information gain.

On the other hand, though the information gain is almost 50% higher than the uniform background case, the disadvantage of employing the sequence-*in*dependent distribution is that it ignores the unique structural influence of the amino acid on its own backbone. Glycine has the widest [ϕ,ψ] dihedral angle propensities [Fig. 3(B)] of all amino acids, while others have less distinct but still significantly unique propensities [e.g., see Asp and Ile backbone propensities in Fig. 3(D,E)]. Since the sequence-*in*dependent [ϕ,ψ] distribution [Fig. 3(A)] contains those conformations mostly accessible only to glycine, approximating the structural propensity of a fragment containing a non-glycyl amino acid using this background is inefficient. (The same is true of proline.) We conclude that analyzing the information brought about by local sequence should be measured from the baseline of the single-residue structural distribution, because the backbone [ϕ,ψ] of an amino acid is most strongly influenced by its own side chain.

There are two ways to estimate these single-residue background distributions. From the data set, one can simply compile the backbone conformations of each of the 20 amino acids and form a frequency distribution. A better method is to use the same strategy of deriving a hybrid distribution to approximate true single-residue propensities. In this case, we would like to anchor the single-residue-specific distribution to the most appropriate background distribution. Results in this section suggest that the best background component should most closely resemble the propensities of the sequence being estimated. Therefore, to construct the most informative hybrid distributions, we employ the universal distribution as the background distribution (Eq.4) and the raw single-residue frequency distribution as the sequence-specific component. Such a hybrid distribution needs a hybrid coefficient, γ_U, which can be optimized by the same procedure described previously. (Note that this hybrid coefficient is independent of the hybrid coefficients used to derive the local sequence-specific distributions.) Using Eq.(10) to optimize γ_U, we find that for [ϕ,ψ] structural space subdivided into an 18×18 grid, the optimal hybrid coefficient has a value of 779. The resulting ensemble of single-residue-specific hybrid distributions with this value of γ_U was then used as the set of background distributions to approximate local sequence-specific distributions in this work.

An Optimal Amino Acid Cluster Configuration Exists That Maximizes Information Gain

Optimization was done for various configurations $\Omega^3_{b=2} = \{n.20.n\}$ for $n = 1,\ldots,12$, and $n = 20$. Figure 5 summarizes the optimization results and shows that the information gain is maximal at the cluster number $n = 7$ configuration. This observation demonstrates the balance between the desirability of having a greater number of amino acid clusters and the limitations set by sparse data. Naturally, extractable information should increase as the alphabet size is increased, since a greater number of "kinds" of residues exposes more of the unique identity of each amino acid. On the other hand, the amount of data limits alphabet size. At higher cluster number, distributions are

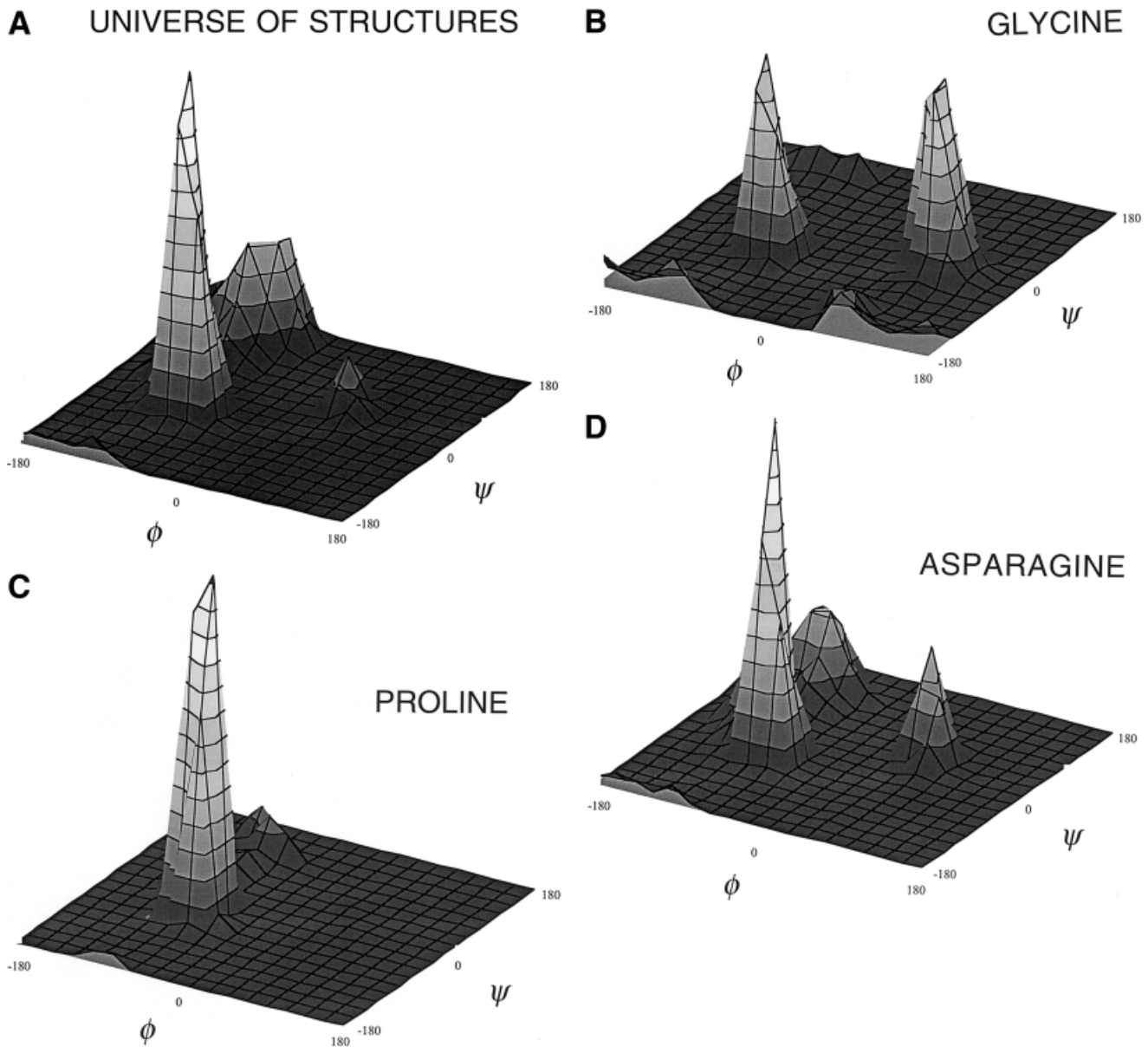


Fig. 3. 3-D [ϕ, ψ] plots. The (*x,y*)-planar axes refer to the two backbone dihedral angles phi and psi, while the vertical *z*-axis reflects the frequency of occurrence in the data set. **A:** Summary of all [ϕ, ψ] conformations found in the data set. **B–E:** Single-residue-specific [ϕ, ψ] distributions for amino acids Gly, Pro, Asn, Ile, respectively.

dominated by the background. Information gain is therefore diminished, and further increases in cluster number are penalized. At the extreme, the partition at $\Omega^3_{b=2}$ = {20.20.20}, which recognizes each amino acid in all positions in the trimer fragment, does not perform as well as the clustered configuration $\Omega^3_{b=2}$ = {7.20.7}, suggesting that judiciously clustering residues may reveal *more* local structural information from a limited data set.

Clustering Patterns of Amino Acids, Derived by Maximizing Information Gain, Reflect Both Similarities and Differences in Their Influence on Local Backbone Conformation

Table II shows the most informative reduced amino acid alphabets $\Xi^3_{b=2}$ for $\Omega^3_{b=2}$ = {*n*.20.*n*}, *n* = 1,…,20. General patterns, consistent with previous work,[6] are readily observed:

1. Gly and Pro are the first to separate into distinct clusters. It is well-known that the presence of glycine or proline in a sequence fragment drastically alters the conformational space of the backbone.
2. In general, the remaining amino acids cluster with respect to hydrophobicity. Non-polar amino acids (Phe, Ile, Leu, Met, Val, and Trp) usually cluster together, as do polar residues (Glu, Lys, Gln, His, and Arg).
3. There is a separation of small, polar amino acids like Thr, Asp, and Ser. Since their side chains are able to form hydrogen bonds with the backbone, backbone

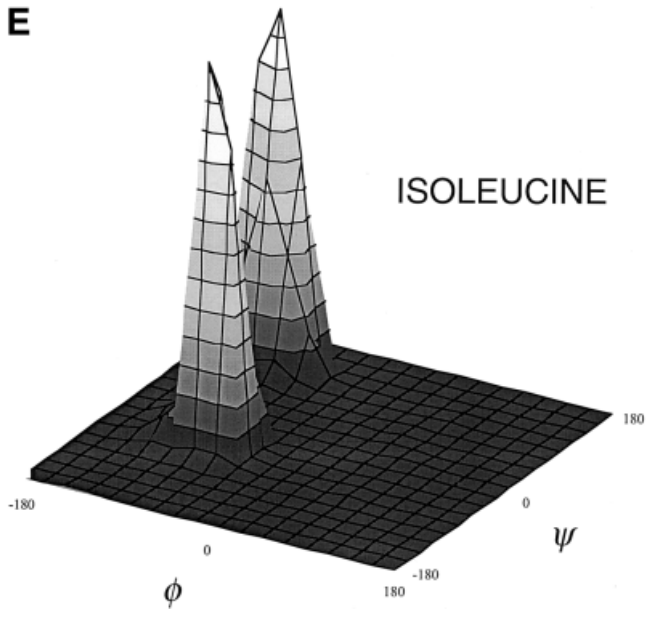


Figure 3. (Continued.)

conformations otherwise energetically unfavored are stabilized.

4. Cys separates to its own cluster in some cases. The formation of disulfide bonds is known to affect the conformation of the surrounding backbone.
5. Independent clustering is observed in the two neighboring positions in the trimer, indicating that amino acids have distinct effects on backbone conformation at different locations within the segment.

These cluster patterns are also seen in results for fragment lengths longer than the trimer level (data not shown).

Residues on the Carboxyl Side Exert Greater Influence on the Conformation of the Backbone Than Residues on the Amino Side

Structural information arising from knowledge of the surrounding tetramer sequence is included in Table III at varying $\Omega^4_{b=2}$ and $\Omega^4_{b=3}$. [Note that the tetramer (and fragments with an even number of residues in general) does not have a central residue; instead, either of the two middle residues may be taken as the location of the [ϕ,ψ] parameter of interest and can be used as the reference point for constructing sequence-dependent distributions.] Comparing the information output of two similar partitions $\Omega^4_{b=3}$ = {5.5.20.5} and $\Omega^4_{b=2}$ = {5.20.5.5} shows that residues found on the carboxyl side of the dihedral angles of interest have a greater informatic value than residues in the amino side. Since the two sequence partitions give the same number of distinct tetramer sequences (2,500), we conclude that the identities of the two amino acid residues in the carboxyl direction provide more structural information than those of the two amino acids residues in the amino direction. These observations are confirmed by simpler dimer partitions, as shown in Table III. For instance, an information gain of 28.49 cnats is achieved for $\Omega^2_{b=1}$ = {20.16} while the equally complex $\Omega^2_{b=2}$ = {16.20} is only able to extract 26.17 cnats. This result is consistent with previous observations,[29] which show a substantial anisotropic influence of amino acids on adjacent backbone conformation.

Sequence Fragments Longer Than the Trimer Length Scale Are Able to Extract More Information From the Data Set

Since the number of amino acid cluster configurations $\Omega^{\mathrm{m}}_{b=\mathrm{i}}$ for a particular fragment length m is 20^{m-1}, an exhaustive search for the optimal cluster configuration becomes increasingly computationally intensive for longer m's. For instance, there are 400 different patterns of clustering in the $i-1$ and $i+1$ positions of the trimer, ranging from {1.20.1} to {20.20.20}. The former extreme describes the situation wherein no sequence information is introduced beyond the identity of the amino acid at position i, which provides no increase in information beyond the single-residue information gain. The latter configuration, with 20 clusters at positions $i-1$ and $i+1$, recognizes the specific identities of the flanking residues. A search for the optimal partition, which exists somewhere between these two extremes, is possible but computationally challenging, since each clustering configuration $\Omega^m_{b=i}$ requires a multi-dimensional search over different amino acid clusterings $\Xi^{\mathrm{m}}_{b=i}$ as well as different values of γ.

With these issues in mind, we accomplish the search by subjecting to the maximization procedure only those configurations we believe are in the vicinity of the optimum. We use two principles to guide this search:

1. In general, larger information gains are found when the number of clusters is larger at positions close to i in the segment. This is not surprising, since we know that closer residues have a greater influence on backbone conformation. This principle is illustrated by the two instances $\Omega^5_{b=3}$ = {4.6.20.6.4} and {6.4.20.4.6}, both of which partition the pentamer universe into 11,520 unique fragments. The information gain for the former configuration is 29.59 cnats while the latter is 29.02 cnats.
2. The contribution to the over-all information gain of the neighboring residues on the carboxyl side is larger than that of residues on the amino side, as noted above.

Though we did not undertake an exhaustive search of all possible clustering configurations at the tetramer length scale, the considerations listed above have allowed us to selectively explore some of them. Table III summarizes these results. Of those configurations, the partition $\Omega^4_{b=2}$ = {3.20.9.4} provides the most information gain, at 29.63 cnats. Similarly, results for local sequence lengths greater than 4 are summarized in Table III. Figure 6 compares the information gains at different sequence lengths. In Figure 6, the extent of clustering at every position determines the number of unique sequence fragments that can be formed.

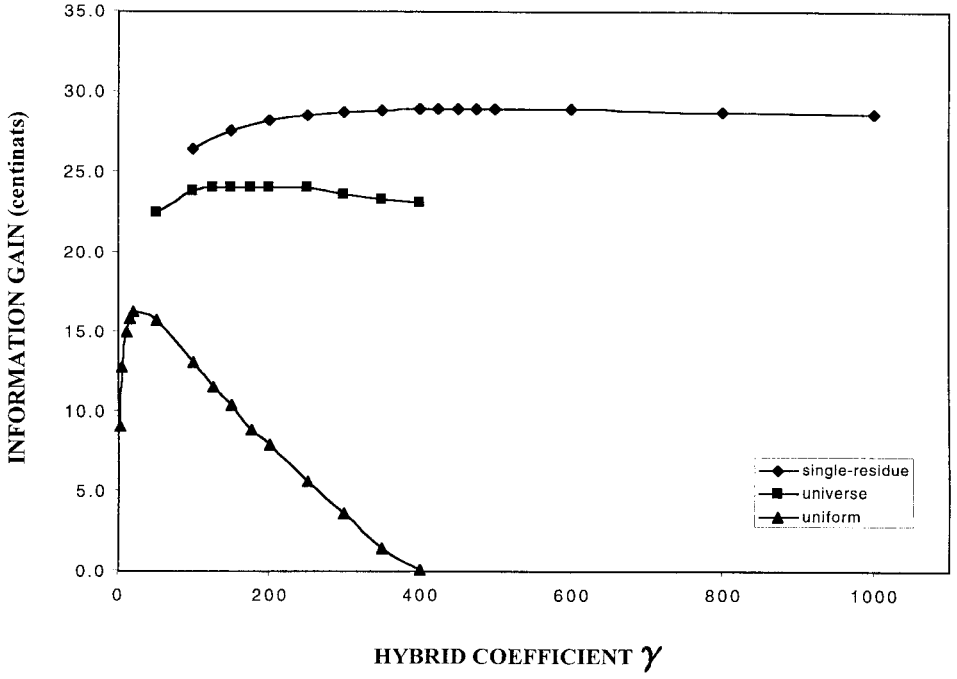


Fig. 4. Information gain at the trimer sequence length scale using three kinds of background distributions: uniform, sequence-independent (universe), and single-residue distributions at cluster configuration $\Omega^3_{b=2}$ = {7.20.7}, over a range of γ's.

Maximum Length Exists Beyond Which Extending the Fragment Degrades Information

The length of the sequence fragment affects the level of pattern detection, and therefore the information gain, because the number of distinct sequence fragments grows exponentially as the length is increased. For example, the pentamer configuration {4.4.20.4.4}, which results in 5,120 unique sequences, has, on average, more than 29 observations per unique reduced pentamer sequence in a dataset of 150,000 residues. A longer segment, the heptamer {4.4.4.20.4.4.4}, gives rise to 81,920 distinct sequence segments, with less than 2 observations per unique sequence on average. This discrepancy limits the power of pattern detection at longer segment lengths, and is reflected in lower information gain (29.36 cnats for the pentamer and 29.19 cnats for the heptamer), even though more local sequence information is used in the latter.

We show another example in Table IV. Amino acids at residues $i-1$ and $i+1$ were clustered into 5 and 7 groups respectively, while the rest were clustered into 3 groups. Initially, the progress of increasing the local sequence knowledge brings about a gradual increase in information gain. A maximum is reached at the hexamer level, after which further increases in length degrade local information. The balance between the need to include as much sequence information as possible and the pressure of a limited database leads to an optimum segment length, at which maximum information is obtained.

Table V summarizes the factors that affect the efficiency of extracting backbone structural information from local sequence. These factors can be divided into three groups: sequence domain, structural domain, and probability approximation. The optimum level of each of these factors is determined by a general principle: the desirability of including as much detail in the representation of the sequence and structure domains is balanced by the pressure of limited data.

Nearest Neighbor Residues Can Have a Substantial Impact on the Informational Entropy of the Backbone Structure

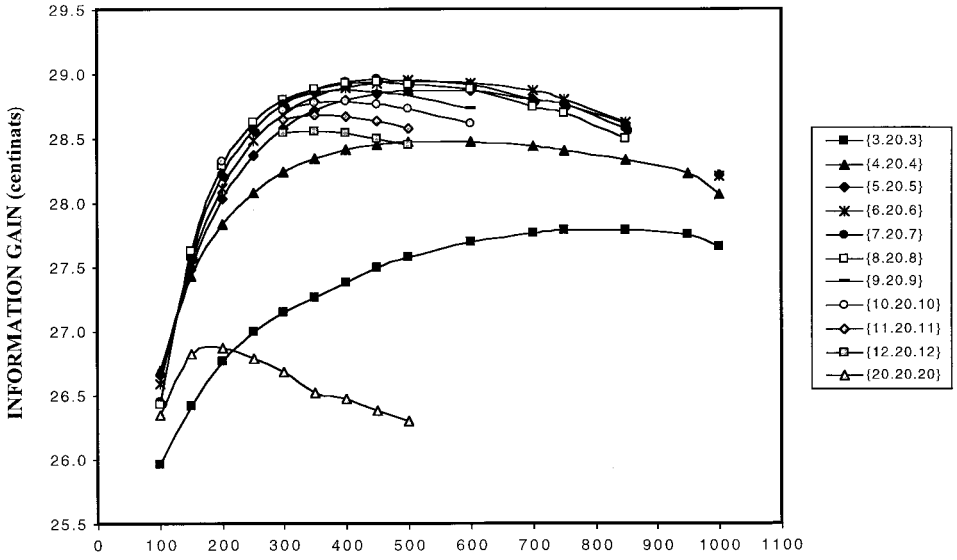
The overall average information gain $I_g(S,C)$, which we have been maximizing, can be re-expressed as the average of individual contributions by all unique sequence fragments using a variant of Eq.(9):

$$I_g(S,C) = \sum I_g(C|S = s)p(S = s) \qquad (16a)$$

where $I_g(C|S = s)$ is the information gain for a specific sequence $S = s$ and $p(S = s)$ is its probability of occurrence in the data set. We can break down the former quantity into two distinct components (using Eqs.8a and 8b) to recognize the separate influences of the amino acid residue s_i and its neighboring residues $s_{neighbor}$ on the backbone conformation:

$$\begin{aligned} I_g(C|S = s) &= H(C) - H(C|S = s) \\ &= H(C) - H(C|S = s_i) + H(C|S = s_i) - H(C|S = s) \\ &= I_g(C|S = s_i) + I_g(C|S = s_{neighbor}) \end{aligned} \qquad (16b)$$

where $s = s_i + s_{\text{neighbor}}$, $H(C)$ is the overall entropy of structures without any sequence-information (i.e., the entropy of the universe of structures), $H(C|S = s)$ is the average entropy when local sequence fragment $S = s$ is



TRIMER $\Omega^3_{b=2}$	n_{seq}	γ	I_g^{max}
{2.20.2}	80	2000	26.34 cnats
{3.20.3}	180	850	27.80
{4.20.4}	320	600	28.47
{5.20.5}	500	600	28.87
{6.20.6}	720	500	28.95
{7.20.7}	980	450	28.97
{8.20.8}	1280	450	28.95
{9.20.9}	1620	400	28.88
{10.20.10}	2000	400	28.79
{11.20.11}	2420	350	28.69
{12.20.12}	2880	350	28.56
{20.20.20}	8000	200	26.87

Fig. 5. Information gain for various amino acid cluster configurations $\Omega^3_{b=2} = \{n.20.n\}$, $n \geq 2$, across different hybrid coefficients γ. The table shows the maximum information gain value for each configuration. (The configuration $\Omega^3_{b=2} = \{2.20.2\}$ is not pictured.) The number of unique trimer sequences, n_{seq}, produced by the reduced alphabet is the product of the cluster numbers at all positions in the segment.

revealed, and $H(C|S = s_i)$ is the entropy of structures when only the amino acid at position i is known (where position i is the location of the [ϕ,ψ] angles in question). These entropy differences can be summarized into two information contributions: $I_g(C|S = s_i)$ is the single-residue specific contribution to the information gain, while $I_g(C|S = s_{\text{neighbor}})$ is the residual information gain brought about by the residues adjacent to the amino acid at position i.

We would like to compare these two contributions. The overall information gain $I_g^{max}(S,C)$ for the trimer length scale is 29.10 cnats (at cluster configuration $\Omega^3_{b=2} = \{4.20.10\}$, in Table III). The average information gain

$$I_g(S = s_i,C) = \sum I_g(C|S = s_i)p(S = s_i) \qquad (16c)$$

caused by the residue at position i is 20.91 cnats, leaving an average nearest-neighbor information gain of 8.19 cnats. Thus, with the current data set, the average contribution of an amino acid to its backbone conformation is 72% of the total local structural information, while the detectable contribution of two adjacent residues is 28%.

The amino acid grouping at each position of the optimal clustering configuration $\Omega^3_{b=2}$, shown in Table VI, simplifies the alphabet from the full amino acid representation to a 4 by 20 by 10 clustering, reducing the number of distinct trimers from 20^3 or 8,000 to 800, out of which 799 occur at least once in the dataset. The distribution of values of the sequence-specific information gains of each of these reduced trimer sequences are plotted in Figure 7 [local

TABLE II. Partition of the 20 Amino Acids Into Clusters at the Trimer Sequence Configuration

	CLUSTER	POSITION $i-1$	POSITION $i+1$
{3.20.3}	1	GLY	GLY PRO
$Ig = 27.80$ cnats	2	CYS PHE ILE LEU MET VAL TRP TYR	CYS PHE ILE LEU MET VAL TRP TYR
	3	ALA ASP GLU HIS LYS ASN PRO GLN ARG SER THR	ALA ASP GLU HIS LYS ASN GLN ARG SER THR
{4.20.4}	1	GLY	GLY
$Ig = 28.47$ cnats	2	PRO	PRO
	3	CYS PHE ILE LEU MET VAL TRP TYR	CYS PHE ILE LEU MET VAL TRP TYR
	4	ALA ASP GLU HIS LYS ASN GLN ARG SER THR	ALA ASP GLU HIS LYS ASN GLN ARG SER THR
{5.20.5}	1	GLY	GLY
$Ig = 28.88$ cnats	2	PRO	PRO
	3	THR	ASP HIS ASN SER THR
	4	CYS PHE ILE LEU MET VAL TRP TYR	CYS PHE ILE LEU MET VAL TRP TYR
	5	ALA ASP GLU HIS LYS ASN GLN ARG SER	ALA GLU LYS GLN ARG
{6.20.6}	1	GLY	GLY
$Ig = 28.95$ cnats	2	PRO	PRO
	3	THR	ASP HIS ASN SER
	4	HIS	CYS THR
	5	CYS PHE ILE LEU MET VAL TRP TYR	PHE ILE LEU MET VAL TRP TYR
	6	ALA ASP GLU LYS ASN GLN ARG SER	ALA GLU LYS GLN ARG
{7.20.7}	1	GLY	GLY
$Ig = 28.97$ cnats	2	PRO	PRO
	3	THR	THR
	4	HIS	LYS GLN
	5	CYS	ASP HIS ASN SER
	6	PHE ILE LEU MET VAL TRP TYR	CYS PHE ILE LEU MET VAL TRP TYR
	7	ALA ASP GLU LYS ASN GLN ARG SER	ALA GLU ARG
{8.20.8}	1	GLY	GLY
$Ig = 28.95$ cnats	2	PRO	PRO
	3	THR	THR
	4	HIS	ASP
	5	CYS	LYS GLN
	6	TRP	HIS ASN SER
	7	PHE ILE LEU MET VAL TYR	CYS PHE ILE LEU MET VAL TRP TYR
	8	ALA ASP GLU LYS ASN GLN ARG SER	ALA GLU ARG
{9.20.9}	1	GLY	GLY
$Ig = 28.88$ cnats	2	PRO	PRO
	3	THR	THR
	4	HIS	HIS
	5	CYS	ASN SER
	6	TRP	ASP
	7	ASP ASN SER	LYS GLN
	8	PHE ILE LEU MET VAL TYR	CYS PHE ILE LEU MET VAL TRP TYR
	9	ALA GLU LYS GLN ARG	ALA GLU ARG
{10.20.10}	1	GLY	GLY
$Ig = 28.79$ cnats	2	PRO	PRO
	3	THR	THR
	4	HIS	HIS
	5	CYS	CYS
	6	TRP	ASP
	7	MET	LYS GLN
	8	ASP ASN SER	ASN SER
	9	PHE ILE LEU VAL TYR	PHE ILE LEU MET VAL TRP TYR
	10	ALA GLU LYS GLN ARG	ALA GLU ARG
{11.20.11}	1	GLY	GLY
$Ig = 28.69$ cnats	2	PRO	PRO
	3	THR	THR
	4	HIS	HIS
	5	CYS	CYS
	6	TRP	ASP
	7	MET	ASN
	8	TYR	SER
	9	ASP ASN SER	LYS GLN
	10	PHE ILE LEU VAL	PHE ILE LEU MET VAL TRP TYR
	11	ALA GLU LYS GLN ARG	ALA GLU ARG

sequence information gain $I_g(C|S = s)$ in Figure 7(A) and neighbor-residue information gain $I_g(C|S = s_{neighbor})$ in Fig. 7(B)]. We make two important observations. First, the range of values for $I_g(C|S = s)$ and $I_g(C|S = s_{\text{neighbor}})$ are [−66,99] and [−20,41] cnats, respectively, with the *spread* of the information gain values of 165 cnats and 61 cnats, respectively. Comparison of these values with the entropy prior to any sequence knowledge (at 3.87 nats or 387

TABLE IIIA. Information Gain at Different Amino Acid Cluster Configuration and Fragment Length†

Cluster configuration		n_{seq}	I_g^{max}	Cluster configuration		n_{seq}	I_g^{max}
Dimer $\Omega_{b=2}^2$	{5.20}	100	25.85		{5.20.8.4}	3,200	29.58
	{8.20}	160	26.04		{6.20.7.4}	3,360	29.53
	{10.20}	200	26.10		{7.20.8.3}	3,360	29.48
	{12.20}	240	26.14		{5.20.7.5}	3,500	29.56
	{14.20}	280	26.16		{4.20.9.5}	3,600	29.55
	{16.20}	320	26.17*		{6.20.6.5}	3,600	29.46
	{18.20}	360	26.16		{6.20.8.4}	3,840	29.54
	{20.20}	400	26.03		{8.20.8.3}	3,840	29.41
Dimer $\Omega_{b=1}^2$	{20.2}	40	24.90		{7.20.7.4}	3,920	29.46
	{20.5}	100	27.84		{5.20.8.5}	4,000	29.54
	{20.7}	140	28.12		{6.20.7.5}	4,200	29.49
	{20.8}	160	28.22		{6.20.6.6}	4,320	29.51
	{20.10}	100	28.36		{7.20.8.4}	4,480	29.48
	{20.12}	240	28.42		{6.20.8.5}	4,800	29.50
	{20.14}	280	28.46	Pentamer $\Omega_{b=2}^5$	{6.20.8.4.3}	11,520	29.57
	{20.15}	300	28.47	Pentamer $\Omega_{b=3}^5$	{4.4.20.4.4}	5,120	29.36
	{20.16}	320	28.49*		{3.6.20.6.3}	6,480	29.64
	{20.17}	340	28.49*		{3.6.20.7.3}	7,560	29.64
	{20.18}	360	28.48		{4.5.20.5.4}	8,000	29.57
	{20.20}	400	28.42		{3.4.20.7.5}	8,400	29.71*
Trimer $\Omega_{b=2}^3$	{2.20.2}	80	26.34		{3.6.20.8.3}	8,640	29.65
	{3.20.3}	180	27.80		{3.7.20.7.3}	8,820	29.63
	{4.20.4}	320	28.48		{3.6.20.7.4}	10,080	29.66
	{5.20.5}	500	28.88		{3.7.20.8.3}	10,080	29.59
	{5.20.7}	700	29.03		{3.6.20.8.4}	11,520	29.67
	{6.20.6}	720	28.95		{4.6.20.6.4}	11,520	29.59
	{4.20.9}	720	29.09		{6.4.20.4.6}	11,520	29.02
	{4.20.10}	800	29.10*		{5.5.20.5.5}	12,500	29.46
	{5.20.8}	800	29.06		{3.6.20.7.5}	12,600	29.63
	{6.20.7}	840	29.00		{3.7.20.8.4}	13,440	29.61
	{7.20.7}	980	28.97		{4.7.20.7.4}	15,680	29.55
	{6.20.9}	1,080	29.04		{5.6.20.6.5}	18,000	29.47
	{7.20.8}	1,120	29.00		{6.6.20.6.6}	25,920	29.30
	{8.20.7}	1,120	28.92		{10.10.20.10.10}	200,000	28.13
	{8.20.8}	1,280	28.95	Hexamer $\Omega_{b=3}^6$	{3.3.20.3.3.3}	4,860	29.01
	{9.20.9}	1,620	28.88		{3.4.20.4.3.3}	8,640	29.44
	{10.20.10}	2,000	28.79		{3.5.20.5.3.3}	13,500	29.66
	{11.20.11}	2,420	28.69		{3.4.20.7.3.3}	15,120	29.76*
	{12.20.12}	2,880	28.56		{3.5.20.7.3.3}	19,200	29.74
	{20.20.20}	8,000	26.87		{3.6.20.6.3.3}	19,440	29.67
Tetramer $\Omega_{b=3}^4$	{5.5.20.5}	2,500	28.88		{4.4.20.4.4.4}	20,480	29.39
	{5.6.20.5}	3,000	28.87		{3.5.20.6.4.3}	21,600	29.76*
	{4.6.20.6}	2,880	29.01		{3.6.20.7.3.3}	22,680	29.65
	{4.7.20.7}	3,920	29.02		{4.5.20.5.4.3}	24,000	29.62
	{5.7.20.7}	4,900	28.96		{3.5.20.7.4.3}	25,200	29.77
	{4.7.20.8}	4,480	29.05		{4.6.20.6.4.3}	34,560	29.62
	{3.8.20.8}	3,840	29.08*		{5.5.20.5.5.5}	62,500	29.33
Tetramer $\Omega_{b=2}^4$	{6.20.7.2}	1,680	29.32	Heptamer $\Omega_{b=4}^7$	{3.3.3.20.3.3.3}	14,580	28.95
	{3.20.9.4}	2,160	29.63		{4.4.4.20.4.4.4}	20,480	29.19
	{6.20.6.3}	2,160	29.52		{3.3.4.20.4.3.3}	25,920	29.36
	{3.20.8.5}	2,400	29.60		{3.3.5.20.5.3.3}	40,500	29.57
	{5.20.5.5}	2,500	29.48		{3.3.5.20.6.3.3}	48,600	29.63
	{6.20.7.3}	2,520	29.52		{3.3.5.20.7.3.3}	56,700	29.64*
	{4.20.8.4}	2,560	29.61*	Octamer $\Omega_{b=4}^8$	{3.3.3.20.3.3.3.3}	43,740	28.90
	{3.20.9.5}	2,700	29.60		{3.3.4.20.4.3.3.3}	77,760	29.29
	{4.20.7.5}	2,800	29.59		{3.3.5.20.5.3.3.3}	121,500	29.48
	{4.20.9.4}	2,880	29.58		{3.3.3.20.8.5.3.2}	129,600	29.62
	{6.20.6.4}	2,880	29.53		{3.3.4.20.7.5.3.2}	151,200	29.63*
	{6.20.8.3}	2,880	29.52		{3.3.5.20.5.4.3.3}	162,000	29.52
	{7.20.7.3}	2,940	29.47		{3.3.4.20.5.5.3.3}	162,000	29.48
	{5.20.6.5}	3,000	29.55		{3.3.4.20.8.5.3.2}	172,800	29.62
	{4.20.8.5}	3,200	29.58		{3.3.5.20.6.4.3.3}	194,440	29.58

†The cluster configurations that give the maximum information gain within a specified sequence segment length are marked by (*).

cnats), emphasizes the fact that even if the overall *average* information gain is relatively small ($I_g^{max}(S,C)$ = 29.10 cnats out of an overall entropy of $H(C)$ = 387 cnats), the influence of specific local sequence $S = s$ on the conformational propensity of the backbone can be quite dramatic. Second, informatic values for specific sequence fragments can be *negative*. Such values, which imply higher conformational entropy, characterize local sequences which confer significant conformational flexibility on the backbone.[22]

We cite examples of particular trimer sequences at amino acid partition $\Omega^3_{b=2}$ = {4.20.10} with unusual entropic and informatic properties. The sequence with the highest conformational entropy [$H(C|S = s)$ = 453 cnats] is reduced sequence 3-Gly-1 (Table VI), shown in Figure 8(A), while the sequence with the lowest conformational entropy [$H(C|S = s)$ = 287 cnats, more than 36% lower than the highest] is reduced sequence 4-Pro-10 [Fig. 8(B)]. These entropy values exemplify the difficulty of predicting fragment conformations.

We now turn to the $I_g(C|S = s_{\text{neighbor}})$ values of these trimer sequences. Reduced sequence 4-Ala-10 gives the highest individual nearest-neighbor information gain (41.4 cnats), while sequence *4-Glu-2* has the lowest, at -20.3 cnats [see Fig. 8(C,D)]. The nearest neighbor residues can have a dramatic narrowing effect on the conformational landscape of the middle residue, as in the case of 4-Ala-10. On the other hand, a negative nearest-neighbor informa-

TABLE IIIB. Summary of Amino Acid Cluster Configurations That Yield the Maximum Information Gain Within Each Segment Length

Cluster configuration		n_{seq}	$I_g^{\max}$
Dimer $\Omega^2_{b=2}$	{16.20}	320	26.17
Dimer $\Omega^2_{b=1}$	{20.16}	320	28.49
	{20.17}	340	28.49
Trimer $\Omega^3_{b=2}$	{4.20.10}	800	29.10
Tetramer $\Omega^4_{b=3}$	{3.8.20.8}	3,840	29.08
Tetramer $\Omega^4_{b=2}$	{4.20.8.4}	2,560	29.61
Pentamer $\Omega^5_{b=2}$	{6.20.8.4.3}	11,520	29.57
Pentamer $\Omega^5_{b=3}$	{3.4.20.7.5}	8,400	29.71
Hexamer $\Omega^6_{b=3}$	{3.4.20.7.3.3}	15,120	29.76
	{3.5.20.6.4.3}	21,600	29.76
Heptamer $\Omega^7_{b=4}$	{3.3.5.20.7.3.3}	56,700	29.64
Octamer $\Omega^8_{b=4}$	{3.3.4.20.7.5.3.2}	151,200	29.63

TABLE IV. $I_g^{\max}$ at Different Cluster Configurations[†]

m	$\Omega^{\mathrm{m}}_{b=i}$	n_{seq}	I_g^{max}
2	{20.7}	140	28.12 cnats
3	{5.20.7}	700	29.03
4	{5.20.7.3}	2,100	29.57
5	{5.20.7.3.3}	6,300	29.59
5	{3.5.20.7.3}	6,300	29.71
6	{3.5.20.7.3.3}	18,900	29.74
7	{3.5.20.7.3.3.3}	56,700	29.64
7	{3.3.5.20.7.3.3}	56,700	29.64
8	{3.3.5.20.7.3.3.3}	170,100	29.54

[†]Position i in each configuration is where cluster number is 20.

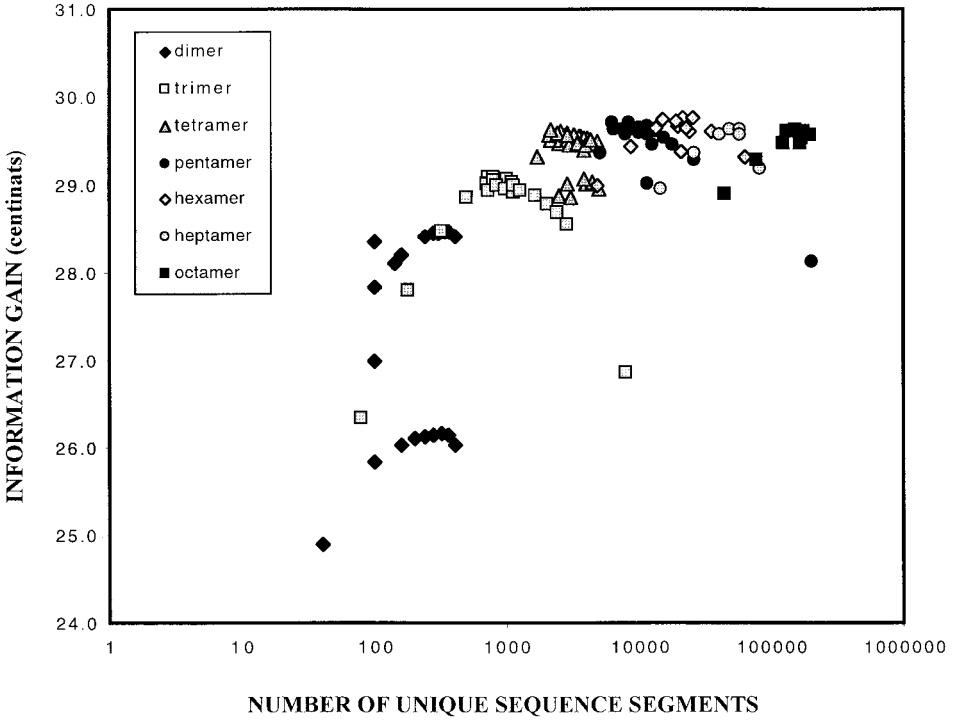


Fig. 6. Maximum information gain at different sequence fragment lengths and different cluster configurations. For instance, at the trimer length scale, maximum information gain was computed for different configurations $\Omega^3_{b=2}$ = {*m*.20.*n*} for various *m* and *n* between 1 and 20. Each point in this plot indicates the maximum information gain at a particular configuration optimized across different hybrid coefficients γ (some of which are listed in Table III).

TABLE V. Factors That Affect the Value of $I_g(S,C)$

Factor	Domain	*Notes*
Sequence		
Sequence length	$m = 1, 2, 3, \ldots$ [single, dipeptide, tripeptide, etc.]	Desirability of longer fragments competes with limited data set size.
Extent of amino acid clustering per position	$\Omega^{m}_{b=i}$	Reducing the amino acid alphabet may result in higher information gain than when full local sequence is considered, due to data set limitations. Higher information gain is extracted when cluster numbers are higher at positions closer to i. More information is stored in the residues on the carboxyl side of i than on the amino side.
Amino acid cluster membership	$\Xi^{m}_{b=i}$	Some locally critical amino acids separate into their own clusters in most optimizations.
Structure		
Choice of structural descriptor	[ϕ,ψ], virtual bond conformation, DSSP classification, etc.	In this work, we use [ϕ,ψ] of residue i to describe backbone conformation, but the methodology can be extended to optimize other discrete forms of structural description.
Level of discretization	{H,E,C} classification, GBMR classification, 18 × 18 grid discretization for [ϕ,ψ], etc.	In this work, the Ramachandran space was discretized into an 18 × 18 grid.
Sequence-dependent probability approximation		
Background component	Single-residue specific distribution, sequence-independent distribution, uniform distribution, etc.	In this work, we show that single-residue-specific distribution at position i gives most information gain.
Functionality of γ	Constant γ, linear, etc.	Higher order functions need more parameters, each of which can be optimized. In this work, we use γ = k.
γ	$\gamma > 0$	Positive integer values of γ are discretized by at least 50 in order to simplify optimization search (e.g., γ = 50, 100, 150, . . .).

tion "gain," similar to that of sequence 4-Glu-2, shows that the adjacent residues also have the ability to override the natural propensity of the middle residue and make a wider range of backbone conformations more energetically accessible.

Comparing the structural propensity profile of trimer sequence 4-Ala-1 [Fig. 8(E)] with that of 4-Ala-10, mentioned above, provides a qualitative insight into the effect of neighboring residues. The central residues of these two trimer fragments are identical, while their amino terminal residues (position $i-1$) belong to the polar amino acid cluster (see Table VI). The only difference in sequence occurs in the carboxyl terminal (position $i+1$): the former has Gly at this position, while the latter has one from the cluster {Ala, Glu, Arg}. This small sequence change translates into a large difference in informational entropies, from a high 355 cnats for trimer 4-Ala-1 to a low 295 cnats for trimer 4-Ala-10 (a difference of 60 cnats).

Analysis of Maximally Informative Sequence-Specific Structural Distributions Reveals Unique Influences by Certain Amino Acids

Quantitative comparison of the influence of each neighboring amino acid on the [ϕ,ψ] dihedral angles of the central residue in a trimer differentiates locally critical amino acids from locally neutral amino acids. We would like to examine further the specific effects of the flanking amino acids on the backbone of the middle residue in the trimer. For each amino acid (except Gly and Pro), we compute the nearest-neighbor information gain $I_g(C|S = s_{\text{neighbor}})$ (via Eq.16) for cluster configuration $\Omega^{3}_{b=2}$ = {5.20.5}, where the first four clusters are identical to the optimal clustering at $\Omega^{3}_{b=2}$ = {4.20.4} (same cluster pattern as in Table I) but with the fifth cluster made up solely by the amino acid of interest. That is, to find the effect of a flanking Ala on the backbone conformation of a trimer, we set up the clustering configuration shown in Table VII. The first two clusters contain Gly and Pro, respectively, followed by the third cluster of non-polar amino acids and the fourth cluster of polar amino acids, and finally the fifth cluster solely occupied by Ala.

Such a clustering will result in some information gain, which may be attributed to the unique interaction of the amino acid of interest (e.g., Ala in the configuration shown in Table VII) with the two flanking residues. The second column of Table VIII tallies this value for the 18 non-glycyl/non-prolyl residues, arranged in order of decreasing information gain. A critical value to examine is the relative contribution to the average information gain of those trimers that contain the particular amino acid as compared to the contribution of those trimers that do not. If we represent the amino acid of interest as X, the nearest-neighbor information gain can be re-expressed as a sum of two contributions to consider the presence and absence of the particular amino acid at position i-1 (the amino terminal residue in the trimer) as

TABLE VI. Optimal Clustering of Amino Acids at the Trimer Configuration $\Omega^3_{b=2}$ = {4.20.10}†

POSITION	$i-1$	CLUSTER	1	GLY
POSITION	$i-1$	CLUSTER	2	PRO
POSITION	$i-1$	CLUSTER	3	CYS PHE ILE LEU MET VAL TRP TYR
POSITION	$i-1$	CLUSTER	4	ALA ASP GLU HiS LYS ASN GLN ARG SER THR
POSITION	i	CLUSTER	1	ALA
POSITION	i	CLUSTER	2	ASP
POSITION	i	CLUSTER	3	CYS
POSITION	i	CLUSTER	4	GLU
POSITION	i	CLUSTER	5	PHE
POSITION	i	CLUSTER	6	GLY
POSITION	i	CLUSTER	7	HIS
POSITION	i	CLUSTER	8	ILE
POSITION	i	CLUSTER	9	LYS
POSITION	i	CLUSTER	10	LEU
POSITION	i	CLUSTER	11	MET
POSITION	i	CLUSTER	12	ASN
POSITION	i	CLUSTER	13	PRO
POSITION	i	CLUSTER	14	GLN
POSITION	i	CLUSTER	15	ARG
POSITION	i	CLUSTER	16	SER
POSITION	i	CLUSTER	17	THR
POSITION	i	CLUSTER	18	VAL
POSITION	i	CLUSTER	19	TRP
POSITION	i	CLUSTER	20	TYR
POSITION	$i+1$	CLUSTER	1	GLY
POSITION	$i+1$	CLUSTER	2	PRO
POSITION	$i+1$	CLUSTER	3	ASP
POSITION	$i+1$	CLUSTER	4	ASN
POSITION	$i+1$	CLUSTER	5	SER
POSITION	$i+1$	CLUSTER	6	HIS
POSITION	$i+1$	CLUSTER	7	LYS GLN
POSITION	$i+1$	CLUSTER	8	CYS THR
POSITION	$i+1$	CLUSTER	9	PHE ILE LEU MET VAL TRP TYR
POSITION	$i+1$	CLUSTER	10	ALA GLU ARG

†The reference distribution is the single-residue-specific distribution at the middle position of the trimer segment.

$$I_g(S = s_{neighbor}, C) = I_g(S_{i-1} = X,C)p(S_{i-1} = X) + I_g(S_{i-1} \neq X,C)p(S_{i-1} \neq X) \quad (17)$$

where S_{i-1} is the residue at position $i-1$ and $p(.)$ refers to a probability. A similar equation can be constructed for the amino acid X in position $i+1$. The computation of each of these quantities is straightforward; computed values are shown in Table VIII for each non-prolyl/non-glycyl amino acid. The third column is the nearest-neighbor information gain of the trimers that contain the particular amino acid X in position $i-1$, or $I_g(C,S_{i-1} = X)$, while the fourth column is that of those trimers that do not, $I_g(C,S_{i-1} \neq X)$. The information ratio

$$I_r = I_g(S_{i-1} = X,C)/_g(S_{i-1} \neq X,C) \quad (18)$$

appears on the fifth column. The right side of Table VIII are the corresponding values for position $i+1$. A high I_r would indicate that trimers with the particular amino acid X at either position have on average a *narrower* structural distribution (lower conformational entropy). We notice that I_r is low (<0.20) for some amino acids, particularly those that generate high average information gains (second column in the table). These amino acids—Thr, His, Asp, Cys, Ser Trp, and Asn—have been observed to separate early in the progress of enlarging the amino acid alphabet (for instance, see Table II). Thr, Asp, and Ser have significantly low information ratios only at position $i+1$, while His, Cys, Trp, and Asn have low ratios on both sides of the central residue.

The nearest-neighbor information gain resulting from identifying these amino acids in the clustering scheme is high, which means that recognizing them apart from the generic hydrophobic/polar clusters results in an *average* narrowing of the entire ensemble of sequence-dependent structural distributions. Curiously, however, specific trimers that do contain these amino acids have only a minimal increase in information gain $I_g(C,S_{i-1} = X)$ and $I_g(C,S_{i+1} = X)$ compared to trimers that do not (resulting in low information ratios). Therefore, these amino acids by themselves do not, on average, significantly increase the ease of predicting the backbone conformation of adjacent positions. However, because their separation in the reduced amino acid alphabet results in a significant increase in overall information gain, the local backbone propensities of short fragments containing them, which may be as entropic (or as wide) as fragments containing another amino acid, must be significantly distinct from the hydro-

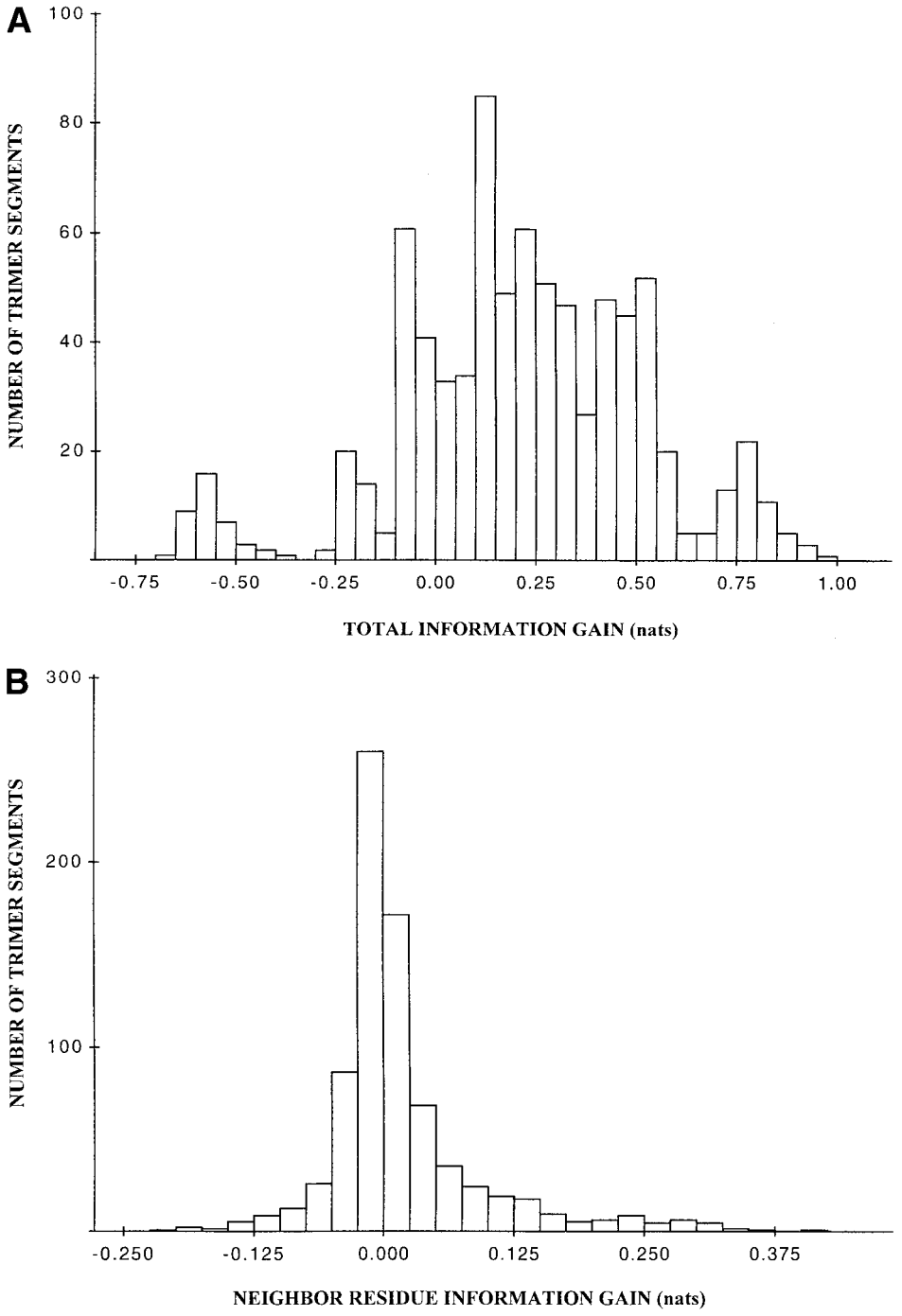


Fig. 7. The distribution of values of the local-sequence-specific information gain given by every unique trimer sequence fragment. The optimal amino acid clustering configuration is described in Table VI. **A:** The distribution of total information gain $I_g^{max}(C|S=S_{trimer})$ resulting from the knowledge of the trimer sequence. **B:** The neighbor-residue information gain $I_g(C|S=S_{neighbor})$.

phobic and polar clusters. The act of removing a divergent member from these clusters increases the kinship among the other members, resulting in an over-all increase in the prediction power of the reduced alphabet. The main influence of some locally critical short sequence fragments is to shift the range of backbone conformations away from typical propensities towards peculiar conformations by making ordinarily less accessible geometries more stable, even though the conformational entropies of those sequence fragments, on average, may be the same (or even higher) than the rest.

CONCLUSION

We set out to determine the influence of local sequence on backbone conformation by designing a procedure topro-

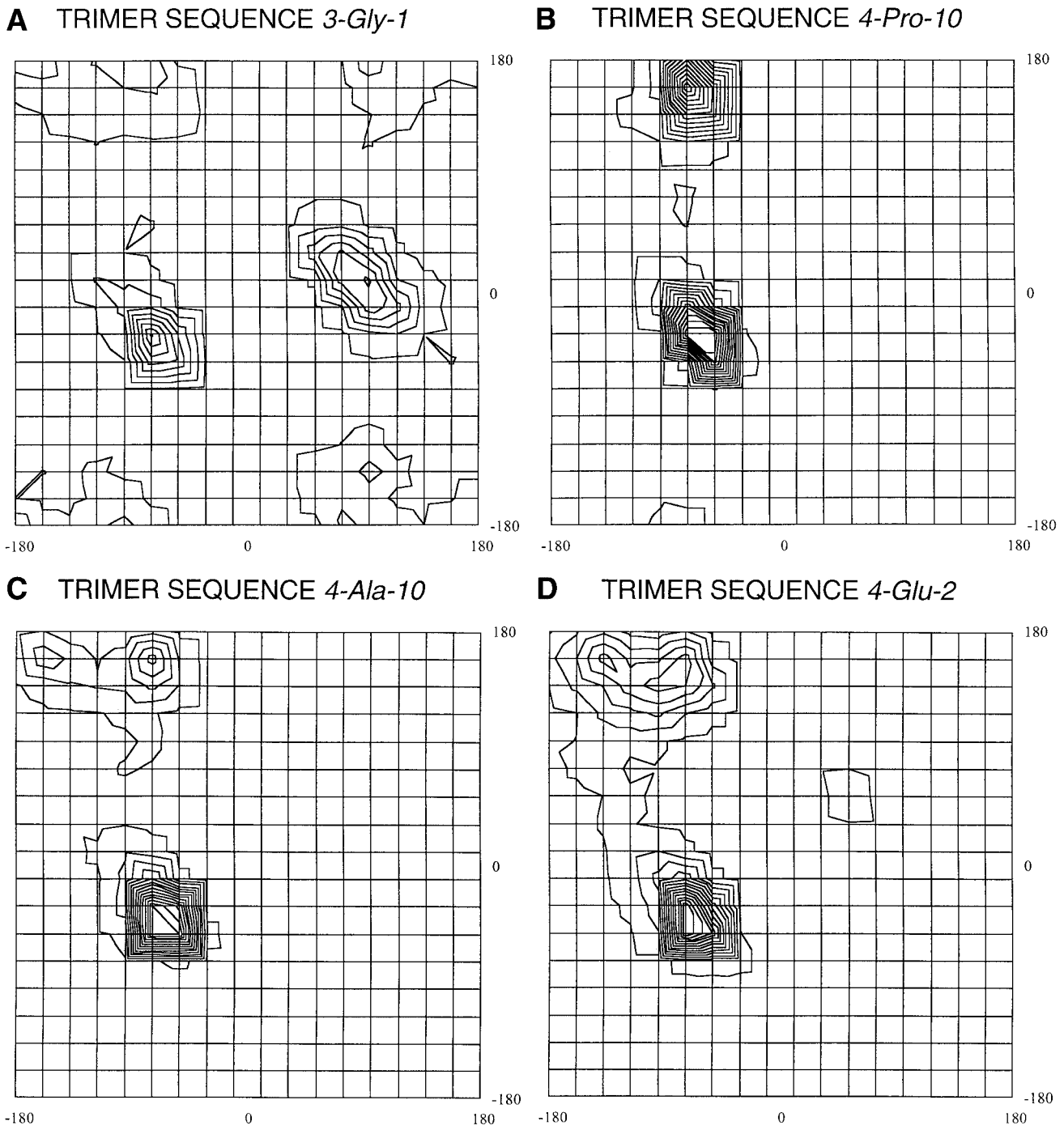


Fig. 8. Optimized conformational probability distributions for specific trimer sequences in the clustering scheme described by Table VI. These contour plots were derived by smoothing out the discrete (18×18 grid) phi-psi probability distributions to approximate a continuous surface. **A:** *3-Gly-1.* **B:** *4-Pro-10.* **C:** *4-Ala-10.* **D:** *4-Glu-2.* **E:** *4-Ala-1.*

duce local sequence-specific structural distributions. These distributions, one for each local sequence fragment, have been constructed to maximize the information obtainable from the current protein structural database. Limitations imposed by scarce data are overcome by a strategy designed to simplify sequence and structure data with minimal loss of information.

We use structural propensity profiles partitioned into a background distribution and a local sequence-specific component. The procedure involves the following:

1. Shannon entropy and information gain, basic information-theoretic quantities, are used to measure the information content of a given set of local sequence-specific

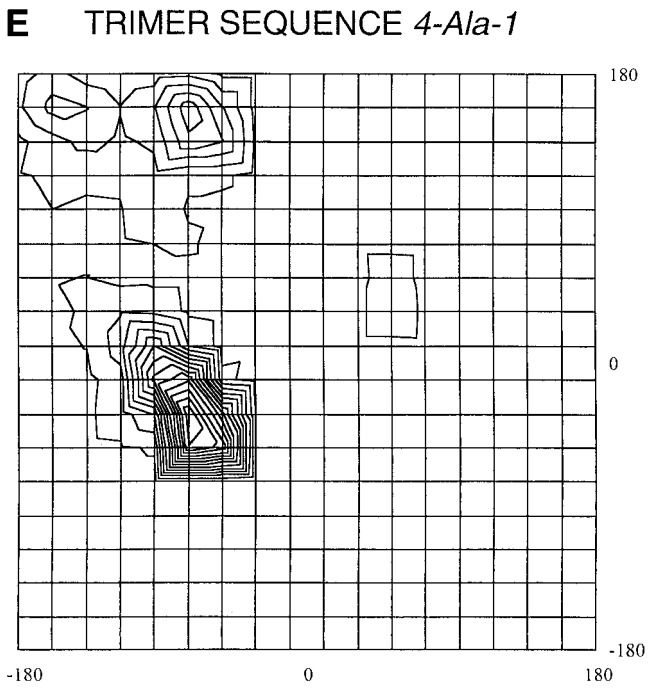


Figure 8. (Continued.)

structural distributions. We maximize information gain by adjusting the way in which these structural distributions are constructed.

2. We find that the background distribution which preserves the most information is the distribution of backbone conformations specific only to the most proximate amino acid (called the single-residue distribution). This background distribution is combined with the local sequence-specific structural component, which is the set of occurrences of the particular sequence fragment in the protein data set. The relative proportions of these two components is determined by a hybrid coefficient γ, which can be optimized.
3. Since the full amino acid representation of even short fragments places undue demands on the limited data set of available protein structures, the amino acid alphabet is reduced from the full 20 to a smaller size. Since this reduction has a direct influence on the formation of the ensemble of structural distributions, we seek a reduction that preserves as much information as possible. The two parameters which determine the optimal reduction are: (a) amino acid alphabet size and (b) cluster membership in each position of the fragment.
4. A Monte-Carlo search is employed to find the set of conditions that maximize the information gain.

TABLE VII. Test Clustering $\Omega^3_{b=2}$ {5.20.5} to Measure the Effect of Alanine on the Backbone Dihedral Angles of the Middle Residue†

POSITION	$i-1$	CLUSTER	1	GLY
POSITION	$i-1$	CLUSTER	2	PRO
POSITION	$i-1$	CLUSTER	3	CYS PHE ILE LEU MET VAL TRP TYR
POSITION	$i-1$	CLUSTER	4	ASP GLU HIS LYS ASN GLN ARG SER THR
POSITION	**i − 1**	**CLUSTER**	**5**	**ALA**
POSITION	i	CLUSTER	1	ALA
POSITION	i	CLUSTER	2	ASP
POSITION	i	CLUSTER	3	CYS
POSITION	i	CLUSTER	4	GLU
POSITION	i	CLUSTER	5	PHE
POSITION	i	CLUSTER	6	GLY
POSITION	i	CLUSTER	7	HIS
POSITION	i	CLUSTER	8	ILE
POSITION	i	CLUSTER	9	LYS
POSITION	i	CLUSTER	10	LEU
POSITION	i	CLUSTER	11	MET
POSITION	i	CLUSTER	12	ASN
POSITION	i	CLUSTER	13	PRO
POSITION	i	CLUSTER	14	GLN
POSITION	i	CLUSTER	15	ARG
POSITION	i	CLUSTER	16	SER
POSITION	i	CLUSTER	17	THR
POSITION	i	CLUSTER	18	VAL
POSITION	i	CLUSTER	19	TRP
POSITION	i	CLUSTER	20	TYR
POSITION	$i+1$	CLUSTER	1	GLY
POSITION	$i+1$	CLUSTER	2	PRO
POSITION	$i+1$	CLUSTER	3	CYS PHE ILE LEU MET VAL TRP TYR
POSITION	$i+1$	CLUSTER	4	ASP GLU HIS LYS ASN GLN ARG SER THR
POSITION	$\boldsymbol{i}+1$	**CLUSTER**	**5**	**ALA**

†ALA is shown in bold to signify that the partition is made to measure its effect on the conformation of the middle residue. It may be replaced by any other amino acid (except Gly and Pro) in order to measure their informatic contribution to the structure of the middle residue.

TABLE VIII. The Influence of Recognizing Each Amino Acid as a Separate Cluster on the Nearest-Neighbor Information Gain†

X	$I_g(S = S_{neighbor})$ (cnats)	$I_g(S_{i-1} = X)$ (cnats)	$I_g(S_{i-1} \neq X)$ (cnats)	I_r	$I_g(S_{i+1} = X)$ (cnats)	$I_g(S_{i+1} \neq X)$ (cnats)	I_r
ASN	9.01	1.19	9.40	**0.13**	0.49	9.43	**0.05**
THR	8.97	3.10	9.33	0.33	0.35	9.50	**0.04**
HIS	8.94	1.07	9.12	**0.12**	1.01	9.13	**0.11**
CYS	8.93	0.76	9.05	**0.08**	0.65	9.05	**0.07**
SER	8.91	2.66	9.31	0.29	0.26	9.46	**0.03**
TRP	8.76	1.45	8.88	**0.16**	1.50	8.88	**0.17**
ASP	8.64	3.98	8.94	0.45	1.22	9.12	**0.13**
TYR	8.45	2.07	8.69	0.24	3.68	8.63	0.43
PHE	8.34	2.62	8.58	0.31	3.77	8.53	0.44
MET	8.36	3.56	8.46	0.42	3.89	8.45	0.46
LYS	8.19	3.81	8.45	0.45	5.68	8.34	0.68
GLN	8.14	4.09	8.30	0.49	4.92	8.27	0.59
ARG	8.13	3.94	8.33	0.47	5.43	8.26	0.66
VAL	7.96	4.21	8.25	0.51	4.67	8.21	0.57
ILE	7.93	4.37	8.14	0.54	5.92	8.05	0.74
GLU	7.72	6.15	7.82	0.79	6.88	7.77	0.88
LEU	7.34	8.70	7.22	1.20	10.89	7.01	1.55
ALA	7.18	9.19	7.00	1.31	9.42	6.98	1.35

†The information ratios lower than 0.20 are shown in bold. They refer to amino acids with critical roles in determining local structure.

We highlight the following observations:

- We are able to extract around 30 cnats of information from the protein data set out of a total 387 cnats of initial uncertainty in the [ϕ,ψ] dihedral angles of the protein backbone, at the hexamer length scale. More information is extracted when there are more amino acid clusters at sites adjacent to the dihedral angles in question. We find that the residues on the carboxyl side of the dihedral angles carry more information than the residues on the amino side. We observe that shorter fragments, as expected, have lower information gains. On the other hand, information is degraded as the sequence fragment is lengthened beyond the heptamer level.
- The clustering of amino acids into groups is a viable strategy to maximize information. Greater information gain is achieved in a finite database when a collapsed alphabet is used instead of the full amino acid representation in representing the sequence of a peptide fragment (Fig. 5). The cluster patterns reveal grouping behavior consistent with previous observations.[6] The most unusual amino acids, Gly and Pro, almost always separate from the rest, while amino acids Thr, Ser, Asp, Asn, Cys, and to a lesser extent His and Ala all cluster in predictable patterns. The rest of the amino acids usually group according to hydrophobicity, indicating that their influence on backbone conformation derives primarily from non-local considerations.
- While the overall information gain from local sequence is small (less than 10% of the initial entropy in specifying the [ϕ,ψ] angles), there are some local sequences that have significantly narrower structural distributions than others. We find that there are strongly-coding trimer fragments that have an entropy of 287–300 cnats, which is at least 20% less than the overall uncertainty of 387 cnats. On the other hand, there are weakly-coding trimers that have entropy levels at 440–453 cnats, or at least 14% above the overall uncertainty. This spread is an expression of the influence of the local sequence on conformational propensities. The local protein folding code is concerned not only with the absolute values of the information brought about by local sequence but also with the relative values of the entropies of these structurally informative fragments, as well as the detailed form of their [ϕ,ψ] propensities.

The ultimate result of this study is a set of sequence-specific local backbone structural distributions optimized with respect to structural information content. These distributions, besides revealing aspects of local coding, are available for a number of applications, including use as a guide to bias simulations in structure prediction efforts.

ACKNOWLEDGMENTS

The authors acknowledge Dr. Hank Chien for his assistance in organizing the protein data set used in this work. We are also grateful for the support of A.D.S. by the Robert Wood Johnson Pharmaceutical Research Institute and by Smith Kline Beecham Pharmaceutical Corporation.

REFERENCES

1. Baldwin RL, Rose GD. Is protein folding hierarchic? I. Local structure and peptide folding. TIBS 1999;24:26–33.
2. Sippl MJ. Calculation of conformational ensembles from potentials of mean force. J Mol Biol 1990;213:859–883.
3. Bahar I, Kaplan M, Jernigan RL. Short-range conformational energies, secondary structure propensities, and recognition of correct sequence-structure matches. Prot Struct Funct Genet 1997;29:292–308.
4. Bystroff C, Baker D. Prediction of local structure in proteins using a library of sequence-structure motifs. J Mol Biol 1998;281:565–577.
5. Bystroff C, Thorsson V, Baker D. HMMSTR: a hidden Markov

model for local sequence-structure correlations in proteins. J Mol Biol 2000;301:173–190.
6. Solis AD, Rackovsky S. Optimized representations and maximal information in proteins. Prot Struct Funct Genet 2000;38:149–164.
7. Rooman MJ, Wodak SJ. Identification of predictive sequence motifs is liimited by protein structure database size. Nature 1988;335:45–49.
8. Gibrat J-F, Robson B, Garnier J. Influence of the local amino acid sequence upon the zones of the torsional angles ϕ and ψ adopted by residues in proteins. Biochemistry 1991;30:1578–1586.
9. Rooman MJ, Kocher J-PA, Wodak SJ. Extracting information on folding from the amino acid sequence: accurate predictions for protein regions with preferred conformation in the abence of tertiary interactions. Biochemistry 1992;31:10226–10238.
10. Munoz V, Serrano L. Intrinsic secondary structure propensities of the amino acids, using statistical ϕ-ψ matrices: comparison with experimental scales. Prot Struct Funct Genet 1994;20:301–311.
11. Kolinski A, Milik M, Rycombel J, Skolnick J. A reduced model of short range interactions in polypeptide chains. J Chem Phys 1995;103:4312–4323.
12. Swindells MB, MacArthur MW, Thornton JM. Intrinsic ϕ,ψ propensities of amino acids, derived from the coil regions of known structures. Nat Struct Biol 1995;2:596–603.
13. Miyazawa S, Jernigan RL. An empirical energy potential with a reference state for protein fold and sequence recognition. Prot Struct Funct Genet 1999;36:357–369.
14. Shannon CE. A mathematical theory of communication. Bell Syst Tech J 1948;27:379–423.
15. Goldman S. Information Theory. New York: Prentice-Hall; 1953.
16. Cover TM, Thomas JA. Elements of information theory. New York: John Wiley & Sons; 1991.
17. Griffiths-Jones SR, Sharman GJ, Maynard AJ, Searly MS. Modulation of intrinsic ϕ,ψ propensities of amino acids by neighbouring residues in the coil regions of protein structures: nmr analysis and dissection of a β-hairpin peptide. J Mol Biol 1998;284:1597–1609.
18. Sippl MJ. Knowledge-based potentials for proteins. Curr Opin Struct Biol 1995;5:229–235.
19. Sun S. Reduced representation model of protein structure prediction: Statistical potential and genetic algorithms. Prot Sci 1993;2:762–785.
20. Hendlich M, Lackner P, Weitckus S, Floeckner H, Froschauer R, Gottsbacher K, Casari G, Sippl MJ. Identification of native protein folds amongst a large number of incorrect models. J Mol Biol 1990;216:167–180.
21. Bryant SH & Lawrence CE. An empirical energy function for threading protein sequence through the folding motif. Prot Struct Funct Genet 1993;16:92–112.
22. Abagyan R, Totrov M. Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. J Mol Biol 1994;235:983–1002.
23. Lee B, Kurochkina N, Kang HS. Protein folding by a biased Monte Carlo procedure in the dihedral angle space. FASEB J 1996;10:119–125.
24. Gibbs N, Clarke AR, Sessions RB. Ab Initio protein structure prediction using physicochemical potentials and a simplified off-lattice model. Prot Struct Funct Genet 2001;43:186–202.
25. Rackovsky S. On the nature of the protein folding code. Proc Natl Acad Sci 1993;90:644–648.
26. 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.
27. Hobohm U, Scharf M, Schneider R, Sander C. Selection of representative protein data sets. Prot Sci 1992;1:409–417.
28. Hobohm U, Sander C. Enlarged representative set of protein structures. Prot Sci 1994;3:522–524.
29. Rackovsky S, Goldstein DA. Differential geometry and protein conformation. v. medium-range conformational influence of the individual amino acids. Biopolymers 1987;26:1163–1187.

APPENDIX A
Estimation and Computer Calculation of Various Informatic Quantities

The average information gain (Eq.9) can be expressed as

$$I(C|S) = H(C) - H(C|S) \tag{A1}$$

where $H(C)$ is the entropy of the universe of conformations found in the data set and $H(C|S)$ is the average entropy of the local sequence–structure system. The quantity $H(C)$ is easily computed by summarizing all the backbone conformations in the data set into a frequency histogram and applying Eq.(5). The quantity $H(C|S)$ is computed as follows: we note that

$$H(C|S) = \sum_i H(C|S = s_i)p(S = s_i) \tag{A2}$$

where

$$H(C|S = s_i) = E[-ln\ p(C = c_m|S = s_i)] \tag{A3}$$

$p(S = s_i)$ is the probability of the local sequence $S = s_i$ occurring in the data set, and $p(C = c_m|S = s_i)$ is the probability of having conformation $C = c_m$ when the local sequence is $S = s_i$. The entropy of structures given by a specific sequence $S = s_i$ is the expected value of the self-information $-ln\ p(C = c_m|S = s_i)$. This is a straightforward calculation: one simply classifies all occurrences of all sequence fragments, noting their conformations, and then uses the probability estimate $p(C = c_m|S = s_i)$, derived by the hybrid distribution method, to determine individual self-informations. The average of these self-informations is our estimate for the entropy $H(C|S = s_i)$. One then computes each of the local-sequence-dependent entropies, and then taking their weighted average (as in Eq.A2) to arrive at $H(C|S)$.

A computationally convenient form for this entropy quantity $H(C|S)$ is

$$H(C|S) = -1/n_{tot} \sum_j ln\ p(C = c_j|S = s_j) \tag{A4}$$

where the index j goes through all n_{tot} local sequence fragments in the data set, noting the local sequence $S = s_j$ and its associated conformation $C = c_j$. The calculation of $H(C|S)$ for a particular amino acid cluster configuration thus becomes straightforward.

The estimate for the probability is applied in the computer program as

$$p(C = c_j|S = s_j) = [\pi(C = c_j|B)\gamma + \pi(C = c_j|S = s_j) \times n(S = s_j) - 1] / [\gamma + n(S = s_j) - 1] \tag{A5}$$

This equation is slightly different from Eq.(4) (i.e., 1 is subtracted from the numerator and denominator). This is necessary so that that the particular observation at j is not used to estimate the probability $p(C = c_j|S = s_i)$ used in Eq.(A4). Such a correction assures that the calculations of probabilities are not inaccurately biased by rarely occurring fragments. Information gain can then be computed using the equations derived in Theory and Methodology.

APPENDIX B
Understanding the Magnitude of the Basic Informatic Unit

The "nat" is the unit of entropy and information when the natural logarithm is used in calculating the Shannon entropy (Eq.5) (and alternatively, the "bit" is the unit

when logarithm 2 is used). A "cnat," or centinat, is 1/100 of 1 nat.

The concept of equivalent states can help us understand the magnitudes of these units. For a discrete structural distribution P with an entropy value of $H(P)$, we can always find an informatically equivalent uniform distribution E with n number of equiprobable states. For a uniform distribution, each state has probability

$$p = 1/n \tag{A6}$$

with Shannon entropy (Eq.5) of

$$H(E) = ln\ n. \tag{A7}$$

The distribution P is said to be informatically equivalent to this uniform distribution if their entropies are equal

$$H(P) = H(E) = ln\ n. \tag{A8}$$

Therefore, distribution P is informatically equivalent to n equiprobable states, where

$$n = e^{H(P)}. \tag{A9}$$

For instance, a structural distribution P with an entropy value of 3.87 nats is informatically equivalent to a uniform distribution with 47.94 or approximately 48 equiprobable states. Therefore, information necessary to resolve P is equivalent to the information necessary to be able to choose the correct state out of 48 possible states.

In order to understand the magnitudes involved in information gain, we can calculate the equivalent number of states that are "lost" or resolved per amino acid position. A gain in information is a loss in entropy, which can be understood as an average reduction of the number of states one has to choose from in the process of determining the structure of a protein fragment. For instance, if we observe an information gain I_g in using a distribution Q instead of a more entropic distribution P,

$$I_g = H(P) - H(Q), \tag{A10}$$

the reduction in the number of states is

$$n_{red} = e^{H(P)} - e^{H(Q)}. \tag{A11}$$

The fraction of states reduced by using the distribution Q instead of P is

$$f_{red} = [e^{H(P)} - e^{H(Q)}]/e^{H(P)} = 1 - e^{H(Q)-¿ \le (P)} = 1 - \exp(-I_g) \tag{A12}$$

An information gain of 1 nat is equivalent to eliminating 63% of the original equiprobable states per amino acid position, while an information gain of 1 cnat is equivalent to eliminating 1%. In this paper, we showed that the initial uncertainty in the [ϕ,ψ] space (discretized in an 18×18 grid) is 3.87 nats, which is equivalent to 48 states. An information gain of around 30 cnats is equivalent to reducing this number by 25.9%, or 12.4 states. This equation can also show the difference in the number of equiprobable states between any two partitions. For this calculation, the *difference* in information gains between the two partitions is used in Eq.(A12) instead of the information gain $-I_g$.

We note that the magnitude of the differences in information gain values among different partitions varies widely. For instance, let us examine the partitions in Table IIIA. While some partitions are clearly preferable than others (e.g., $\Omega^4_{b=2}$ = {4.20.8.4}, with 29.61 cnats, is better than $\Omega^2_{b=2}$ = {16.20}, with only 26.17 cnats), other differences are less severe (e.g., $\Omega^2_{b=2}$ = {14.20}, with 26.16 cnats vs. $\Omega^2_{b=2}$ = {16.20}, with 26.17 cnats). It is important to remember that these informatic values are exponential quantities, and therefore slight differences in information gain per position accumulate exponentially in the length of the sequence, and therefore may prove to be significant. While a rigorous statistical treatment is needed to determine precisely the significance of very slight differences, the major goal of maximizing information will still ultimately prefer those partitions with higher information gain. Therefore, choosing $\Omega^2_{b=2}$ = {16.20} over $\Omega^2_{b=2}$ = {14.20} remains a prudent move.