Protein Sequence Randomness and Sequence/Structure Correlations

Runa S. Rahman* and S. Rackovsky‡
*Department of Physics and Astronomy, University of Rochester, and the ‡Department of Biomathematical Sciences, Mt. Sinai School of Medicine, New York, NY 10029-6579 USA

ABSTRACT We investigated protein sequence/structure correlation by constructing a space of protein sequences, based on methods developed previously for constructing a space of protein structures. The space is constructed by using a representation of the amino acids as vectors of 10 property factors that encode almost all of their physical properties. Each sequence is represented by a distribution of overlapping sequence fragments. A distance between any two sequences can be calculated. By attaching a weight to each factor, intersequence distances can be varied. We optimize the correlation between corresponding distances in the sequence and structure spaces. The optimal correlation between the sequence and structure spaces is significantly better than that which results from correlating randomly generated sequences, having the overall composition of the data base, with the structure space. However, sets of randomly generated sequences, each of which approximates the composition of the real sequence it replaces, produce correlations with the structure space that are as good as that observed for the actual protein sequences. A connection is proposed with previous studies of the protein folding code. It is shown that the most important property factors for the correlation of the sequence and structure spaces are related to helix/bend preference, side chain bulk, and β-structure preference.

INTRODUCTION

The problem of the comparison of biopolymer sequences is central to an understanding of evolutionary and functional relationships and, as a result, has received considerable attention. The most widely used methods for sequence comparison are based on the concept of alignment. In this approach, two or more sequences to be compared are juxtaposed in such a manner that corresponding monomer units in the various molecules or fragments are identified, and some measure is then computed of the degree to which pairwise similarity of sequences is observed. Similarity of monomer units is established by means of a predefined scale, which can measure actual sequence identity, or functional identity as quantified by similarity of physicochemical properties, frequency of mutational replacement, or some other relevant measure. Penalties must be included to account for the presence of either insertions or deletions in one sequence with respect to another, and various functions have been developed to do this. This method has proven useful in numerous areas, and an extensive literature exists on various aspects of the approach (Needleman and Wunsch, 1970; Altschul and Lipman, 1990; Barton and Sternberg, 1990; Karlin and Altschul, 1990; Altschul, 1991; Leung et al., 1991; Niefind and Schomburg, 1991; Schuler et al., 1991; Saqi and Sternberg, 1991).

The alignment method, despite its power and utility, has certain limitations:

- It is best applied to sequences that are fairly closely related. The detection of distant relatedness presents difficulties, although recent progress in this area has been reported (Lawrence et al., 1993).
- It becomes difficult to align sequences as the difference in their lengths increases.
- The penalty functions for insertions and deletions, which are an essential feature of the method, are generally defined in an arbitrary fashion.
- Simultaneous alignment of multiple sequences is not straightforward.

For these reasons, an alternative approach has been developed (Blaisdell, 1986; Blaisdell, 1989a; Blaisdell, 1989b; Blaisdell, 1991; Pietrokovski et al., 1990), based on a "linguistic" view of biopolymer sequences. In this approach, one is not interested in matching long stretches of sequence. Rather, one compares the distributions of "words," or sequence fragments, of a predefined length in the sequences of interest. As these distributions can be normalized by sequence length, it is possible to compare sequences of arbitrarily different length on a uniform basis. Furthermore, it is not necessary to identify pairwise correspondences between equivalent monomer units. Therefore, sequence similarity can be detected without searching for homology. Finally, the presence of insertions and deletions is automatically accounted for by alterations in the sequence fragment distributions. Therefore, it is not necessary to invoke arbitrarily defined penalty functions.

We have previously (Rackovsky, 1990) applied this approach to the comparison and classification of protein structures. In the present work, we develop methods for the analysis of protein sequences in terms of sequence fragment

Received for publication 6 September 1994 and in final form 5 January 1995.

Address reprint requests to Dr. S. Rackovsky, Department of Biomathematical Sciences, Mt. Sinai School of Medicine, One Gustave L. Levy Place, Box 1023, New York, NY 100259-6579. Tel.: 212-241-4868; Fax: 212-860-4630; E-mail: shelly@msvax.mssm.edu.

0006-3495/95/04/1531/09 $2.00

distributions, which are constructed in a manner that reflects the physical characteristics of the amino acids. Our approach is structure driven. We seek to find that representation of protein sequences that results in the best possible correlation between protein sequence differences and protein structural differences. We shall be interested in determining precisely how good that correlation can be, and what the implications are for general questions of protein folding. This approach differs from that taken in recent work on the threading problem, in which that structure is sought that gives the best match to a specific sequence (Sippl and Weitckus 1992; Godzik and Skolnick 1992; Maiorov and Crippen 1992; Bryant and Lawrence 1993).

MATERIALS AND METHODS

Before discussing details of the methods used herein, it is appropriate to give a brief summary of the approach we shall take.

We begin by representing the amino acids in terms of their physical properties, using a statistically derived representation that ensures that almost all of the properties are accounted for. The amino acids are clustered in property space, making it possible to transform actual amino acid sequences into reduced sequences that improve the statistics of sequence matching while retaining the physical basis of that matching. These reduced sequences are used, together with methods that were previously developed for the study of protein structures (Rackovsky, 1990), to calculate pairwise distances between protein sequences in a large database, and thus to define the structure of a sequence space. The proteins in this data base are those that were used in the previous structural studies. Weights are then attached to the factors that define the amino acid representation, and those weights are varied to produce the optimal correlation between distances in the sequence space that we have constructed and the corresponding distances in the previously defined (Rackovsky, 1990) structure space for the same proteins.

Sequence representation

The first problem that must be addressed is the representation of protein sequences. Ideally, a representation of sequence should not consist simply of a list of amino acid names but should rather encapsulate attributes of the amino acids in such a way that the metric for sequence comparison measures an actual physical difference between sequences. The selection of appropriate amino acid properties for this purpose, however, is subject to certain difficulties. First, there are many property sets that have been reported for the naturally occurring amino acids. The selection from this list of a representative set is necessarily a subjective exercise. Second, any given set of properties is likely to contain members that, having been developed without reference to one another, are statistically correlated to some degree. This leads to a sequence representation in which certain attributes of the amino acids are excessively weighted, whereas others are insufficiently stressed.

To avoid these problems, we make use of results developed by Kidera et al. (1985a, b). They collected almost all of the property sets available for the amino acids and performed a factor analysis on them. They were able to show that most of these property sets can be represented by a set of 10 factors and further demonstrated that these factors account for 86% of the variance of all of the properties included in the data base. It is therefore a good approximation to represent the physical properties of the 20 amino acids with the 10 property factors.

We summarize briefly the physical significance of the 10 factors. The first 4 are correlated with particular physical properties: factor 1, helix or bend structure preference related; factor 2, bulk related; factor 3, β-structure preference related; and factor 4, hydrophobicity related. The remaining 6 factors are linear combinations of several properties. The property with which each has the highest correlation is given: factor 5, normalized frequency of double bend; factor 6, average value of average composition or average value of partial specific volume; factor 7, average relative fractional occurrence of flat extended structure; factor 8, normalized frequence of α-region; factor 9, pK-C; and factor 10, surrounding hydrophobicity in β-structure.

Thus, we may represent each of the 20 naturally occurring amino acids by a vector of 10 property factors. We wish, as mentioned above, to attach weights to each of the property factors, with a view to determining the role of each of the properties in correlating the sequence distances that we will calculate with structure distances already available. An amino acid X thus becomes

$$X = (a_1x_1, a_2x_2, \cdots, a_{10}x_{10}), \tag{1}$$

where the x_i are the property factors and the a_i are the corresponding weights.

Amino acid clustering

One of the problems that one faces in comparing sequences in a data base of finite size is a scarcity of matches, which leads to statistical difficulties. To circumvent this problem, we make use of the amino acid representation of Eq. 1 to carry out a clustering of the amino acids. This clustering will enable us to rewrite actual amino acid sequences as reduced sequences of amino acid clusters, the members of which have common physical properties. This in turn improves the statistics of sequence matching in a physically meaningful way and makes it possible to identify sequence similarities that would not be evident otherwise.

It is natural to think of the property factor representation as defining a 10-dimensional Euclidean space, in which each amino acid corresponds to a point with coordinates given by the 10 components of its property factor vector. One can then define a distance function between two amino acids, X and Y, as the Euclidean distance between their corresponding points:

$$\Delta(\mathrm{X}, \mathrm{Y}) = \left[\sum_{i=1}^{10} a_i^2(x_i - y_i)^2\right]^{1/2}. \tag{2}$$

This distance function has the properties one would expect, in the sense that two amino acids that have similar physical characteristics will be close together in the space, and the distance between them will increase as they become less alike.

Once the distance function is defined, it is a simple matter to carry out a cluster analysis of the 20 points in property space for any given set of values of the property weights. In this work, the K-means clustering algorithm (Späth, 1980) is used. The number N_C of clusters to be identified is selected in advance, and the algorithm finds the optimal distribution of the amino acids among the clusters. In most of what follows, we will use $N_C = 4$, a value chosen as a compromise between the preservation of detail and greater data compression.

This approach, which is based directly on amino acid properties, represents an alternative to methods using substitution matrices derived from multiple alignments (Henikoff and Henikoff 1992) or contact frequencies (Miyazawa and Jernigan 1993).

Reduced amino acid sequences

Once the membership of the N_C clusters is identified, the actual amino acid sequences of the proteins in question can be rewritten in terms of the numbers of the clusters in which the amino acids fall. Thus, for example, consider the sequence fragment AlaLysTrp. Suppose that the distribution of amino acids among the clusters was such that Ala and Lys occur in cluster 2, whereas Trp occurs in cluster 3. (Note that the numbering of the amino acid clusters is arbitrary.) The actual sequence would then be mapped into the reduced sequence 223.

Construction of sequence fragment distributions

Once the reduced amino acid sequences of the proteins of interest have been determined, we are in a position to develop methodology for the comparison

of the sequences. As suggested in the Introduction, we do not intend to carry out a linear comparison of sequence strings of the type involved in alignment algorithms. In fact, we shall not represent the sequences as linear strings at all. Rather, we shall characterize a sequence by the distribution in that sequence of sequence fragments, or "words", of a length chosen in advance. The distribution is constructed by moving a window of length L along the sequence and counting the number of occurrences of each observed L-residue fragment. This is equivalent to constructing a histogram of sequence fragment counts. The histogram is then normalized by the total number of sequence fragments in the protein under study.

To make this idea more precise, suppose we choose $L = 4$, so that we describe the protein sequence in terms of 4-residue fragments. (Most of the results of this work will be developed for $L = 4$, a length scale that includes multiple residues but allows the generation of adequate population statistics.) Then the sequence of protein P is described by a distribution that we will call **P**. This distribution is mathematically equivalent to an array with four indices. The element P_{ijkl} is equal to the number of occurrences of the sequence fragment $ijkl$ (where i, j, k, and l represent the numbers of the clusters in which the actual amino acids of the sequence occur, as described above), divided by the total number of sequence fragments of length 4 in P:

$$P_{ijkl} = N_{ijkl}/\Sigma_{\mathbf{P}}(N_{mnpq}). \tag{3}$$

The distribution **P** is a unique fingerprint for the sequence of the protein P and will be used in the actual comparison of sequences. The normalization of **P** makes it possible to compare sequences of completely different lengths on an equal footing, as one is now comparing not the absolute number of occurrences of different sequence fragments but rather the fractional occurrence of the various sequence fragments in the two sequences.

It should be remarked that Van Heel (1991) has also developed a sequence space approach to sequence analysis. His method, however, is based on actual sequence counts (for dipeptides) rather than on the use of amino acid properties.

Sequence distance function

It remains to define a distance function for the comparison of the sequence fragment distributions we have just defined. Several definitions are possible. We have chosen to use a simple Euclidean distance function. The distance between the sequences of two proteins P and Q is given by the function

$$D(\mathbf{P},\mathbf{Q}) = \left[\sum_{i,j,k,l} (P_{ijkl} - Q_{ijkl})^2\right]^{1/2}. \tag{4}$$

Here the sum is over all the indices of the distributions **P** and **Q**.

Structure distance matrix

We are interested in optimizing the correlation between the sequence distances that we calculate by using the methods as described above and structure distances calculated for the same molecules. These distances were calculated previously (Rackovsky, 1990) with methods analogous to those we have outlined here.

Protein data base

The set of proteins used in the present work is the same as that used in our previous classification of protein structures (Rackovsky, 1990), with the exception of 9 proteins with sequences that were not adequately known at the time the structures were deposited. This results in a set of 114 proteins for which it is possible to compare sequence and structure distances. The 9 proteins omitted, in the Brookhaven Protein Data Bank identification code, are 2BCL, 156B, 155C, 1PGI, 2YHX, 1KGA, 1PEP, 2PGK, and 2TNC.

Optimization of factor weights

The crux of the present work is the search for the set of factor weights that gives the best correlation between the sequence distances and the corresponding structure distances. Let the structural distance between proteins P and Q be Δ(P,Q). Furthermore, for convenience, let us define the squares of the property factor weights of Eq. 1 by

$$a_i^2 = w_i. \tag{5}$$

We shall be interested in minimizing the correlation function

$$f(\{w_i\}) = \frac{2}{N(N-1)\sigma_D\sigma_\Delta} \sum_{\substack{P,Q \\ P<Q}} (D(P,Q) - \bar{D})(\Delta(P,Q) - \bar{\Delta}) \tag{6}$$

with respect to the set of weights w_i. Here,

$$\sigma_D = \left\{\frac{2}{N(N-1)} \sum_{\substack{P,Q \\ P<Q}} [(D(P,Q)) - \bar{D}]^2\right\}^{1/2}, \tag{7}$$

with a similar definition for σ_Δ, the overbar denotes the average value of a quantity, and N is the number of proteins.

Note that f depends on the set of weights through the D(P, Q) and that the elements of the structure distance matrix Δ(P, Q) remain fixed. As the property weights are varied, the amino acid clustering changes. This in turn changes the reduced sequences of the proteins, and thus the sequence distances D change. The value of f therefore varies.

The absolute values of the w_i are not important. What matters is the relative magnitudes of the w_i. Because of this fact, it is sufficient to search over the nine-dimensional hyperplane

$$\sum_{i=1}^{10} w_i = 1. \tag{8}$$

It should be remarked that continuous variation of the w_i does not lead to continuous variation of the amino acid clustering. Therefore, D(P, Q) and f change discontinuously as the weights change. The value of D(P, Q) remains constant over intervals of variation of the w_i. This consideration suggests that minimization of the objective function subject to the constraint of Eq. 8 can be carried out by searching on a grid in the weight space. The searches detailed herein were carried out on a grid with intervals $w_i = 0.1$. (Exploration of the weight space in the neighborhood of the weighting optima indicates that this interval is adequate to locate the optimal weighting. In some cases, a small neighborhood of a weighting optimum on the grid may give correlation coefficients equal to those observed at the optimum that we report. However, these neighborhoods seem to be small, and this possibility does not affect the conclusions that we put forth below.)

The procedure for minimization of the objective function is summarized in Fig. 1.

RESULTS AND DISCUSSION

We wish to address two questions in this investigation. We would like to know how closely the sequence distances can be brought into correlation with the structure distances, and we would like to know whether the best correlation that can be produced between the two distance matrices is statistically significant. We shall consider the correlation between sequence and structure distance matrices to be statistically significant if it is better than that which can be produced by correlating the structure distance matrix with a set of sequence distances arising from randomly generated sequences. Results of our calculations are summarized in Table 1. A number of points are evident from the data therein.

The best correlation achieved between the sequence distance matrix D and the structure distance matrix Δ is not very high ($f = 0.451$). The range of variation is also not large. It will be seen that $0.379 \leq f \leq 0.451$, with the value of f a function of the length scale chosen for D, the length scale

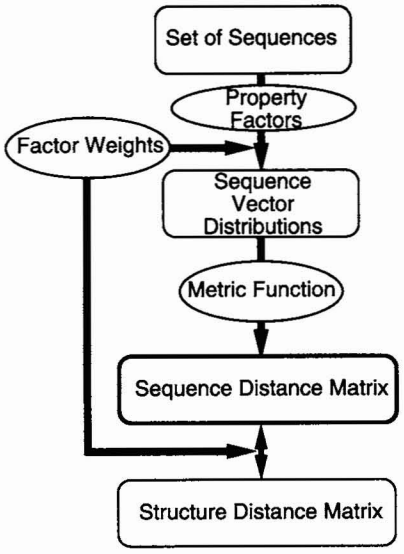

FIGURE 1 A schematic diagram of the optimization loop used to determine the optimal weights for correlating sequence-sequence and structure-structure distances.

chosen for Δ, and the number of amino acid clusters used to generate the reduced sequence.

In view of the low values of f that we observe, it is of particular interest to determine the statistical significance of the structure space/sequence space correlation. In this connection, we first examine the correlation coefficients denoted in Table 1 by the descriptor RAND A. These values of f arise from sets of 114 sequences that are generated randomly, with each random sequence the same length as the actual sequence it replaces and amino acid composition selected according to a probability distribution that mimics the overall composition of the actual 114-protein sample. The time requirements of these computations are such that only 10 random sequence sets were generated. Despite the limited size of this sample, it is clear that the f values arising from actual protein sequences are substantially higher than those arising from the RAND A procedure. The value of f arising from the actual protein sequences, for the parameter set used in the random generation (four-residue structure and sequence length scales and four amino acid clusters), is 0.438. This is 4.4 standard deviations greater than the average of the 10 random correlation coefficients ($\langle f_{\text{RAND A}} \rangle = 0.372$, $s_{\text{RAND A}} = 0.015$, where s is the sample standard deviation). We conclude that the structure/sequence correlation produced by the optimization process is statistically significantly different from those arising from the RAND A sequences.

It is appropriate to ask whether amino acid clustering and the use of reduced sequences have a significant effect on the sequence distance matrix. To investigate this point, a correlation coefficient was determined between the sequence distance matrix with no amino acid clustering and that with four amino acid clusters and equal factor weighting (Table 2). The correlation coefficient is very high ($f = 0.905$),

TABLE 1 Correlation coefficients and factor weights and optimized sequence distance matrices

Length scale		Amino acid clusters	Sequences	Correlation coefficient	Weights*
Structure	Sequence				
4	4	4	Actual	0.438	2330000200
4	5	4	Actual	0.418	2330000200
5	5	4	Actual	0.451	2330000200
4	4	2	Actual	0.379	3201002002
4	4	3	Actual	0.433	2220100210
4	4	4	Actual	0.438	2330000200
4	4	5	Actual	0.421	0010110520
4	4	6	Actual	0.398	1010010331
4	4	4	RAND A‡	0.377	0010010116
4	4	4	RAND A	0.360	0001223011
4	4	4	RAND A	0.342	0002050201
4	4	4	RAND A	0.388	0300101014
4	4	4	RAND A	0.377	1002500200
4	4	4	RAND A	0.382	0011124010
4	4	4	RAND A	0.372	0200100241
4	4	4	RAND A	0.386	0101011033
4	4	4	RAND A	0.381	0002312020
4	4	4	RAND A	0.356	0102201310
4	4	4	RAND B§	0.458	2141001100
4	4	4	RAND B	0.452	3121100200
4	4	4	RAND B	0.395	0030010033
4	4	4	RAND B	0.448	2210101300
4	4	4	RAND B	0.425	4200110020
4	4	4	RAND B	0.433	2200102111
4	4	4	RAND B	0.371	1230010201
4	4	4	RAND B	0.400	2210110201
4	4	4	RAND B	0.461	2230000012
4	4	4	RAND B	0.415	2121100300

*This is a shorthand notation for the squared weights. Each of the 10 integers equals 10 times the corresponding squared weight of Eq. 5. Thus, 2 denotes $w = 0.2$.

‡RAND A, 114 randomly generated amino acid sequences, each of the same length as the actual sequence it replaces but with amino acid composition generated by using a probability distribution based on the overall amino acid composition of the 114-sequence data base.

§RAND B, 114 random amino acid sequences, each of the same length and with the same average amino acid composition as the actual sequence it replaces.

which demonstrates clearly that the data compression resulting from amino acid clustering retains sequence similarity information. We have also calculated correlation coefficients between the optimized distance matrix arising from sequence reduction with four amino acid clusters and the distance matrices arising from actual amino acid sequences and from equally weighted (unoptimized) reduced sequences. The correlation coefficients (Table 2) are 0.752 and 0.722, respectively. The similarity of these results lends additional support to the use of reduced sequences in sequence comparison studies. (It will also be seen from Table 2 that the correlation coefficient between the unoptimized distance matrix arising from the reduced sequences, i.e., the starting condition for the optimization, and the actual structure matrix is 0.323, well below those arising from randomly generated sequences.)

We now compare the optimal sequence distance matrix with those arising from a different set of random sequences. In these sequences, denoted in Table 1 by the descriptor RAND B, the composition of each of the 114 random

TABLE 2 Some other correlation coefficients

Types of distance matrix		
Matrix 1	Matrix 2	$f_{Distance}$*
Sequence; length scale 4; 4-amino-acid clusters; equal weights	Sequence; length scale 4; no amino acid clustering	0.905
Sequence; length scale 4; 4-amino-acid clusters; equal weights	Structure; length scale 4	0.323
Sequence; length scale 4; 4-amino-acid clusters; equal weights	Sequence; length scale 4; 4-amino-acid clusters; optimized weights	0.722
Sequence; length scale 4; no amino acid clustering	Sequence; length scale 4; 4-amino-acid clusters; optimized weights	0.752

*See text.

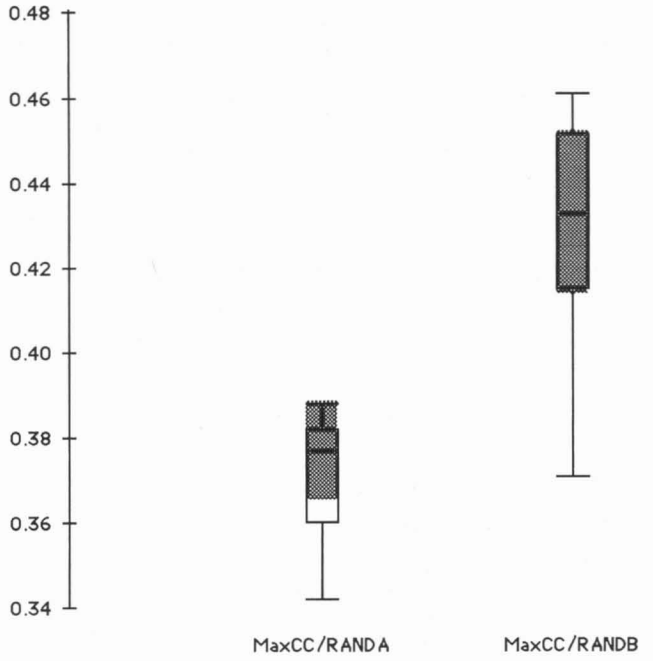

FIGURE 2 Side-by-side box plots demonstrating that the maximal correlation coefficients between sequence matrices arising from random sequences, generated by two different methods (see text and Table 1), and the structure distance matrix, differ in a statistically significant manner. In each box plot, the central horizontal line denotes the median correlation coefficient, the upper and lower horizontals of the box delimit the middle half of the data (between the 25th and 75th percentiles), and the extension bars denote the outer limits of the data. The shaded region demarcates the 95% confidence interval for each set of data. Note that these confidence intervals do not overlap.

sequences mimics that of the actual sequence that it replaces. It will be observed that the correlation coefficients are, for the most part, substantially higher than in the RAND A cases and that the correlation coefficient between the actual protein sequence distances and the structure distances is within the range observed for these randomly generated sequences [$\langle f_{\text{RAND B}} \rangle = 0.426$, $s_{\text{RAND B}} = 0.030$].

We next demonstrate that the difference between the two sets of randomly generated conformations is statistically significant. In Fig. 2 we show side-by-side box plots of the data for the two sets of randomly generated conformations. It will be seen that the ranges of the correlation coefficients in the two cases are almost completely disjoint and that the 95% confidence intervals for the two cases are completely separate, indicating that the two sets of data represent statistically different behavior.

The sharp contrast between the two cases, one in which randomly ordered sequences are generated that approximate the overall composition of the data base and one in which randomly ordered sequences are generated that approximate the specific composition of each protein in the data base, suggests that much of the structural information encoded in the protein is dictated by the composition of the molecule rather than the specific sequence of residues. Hints of this remarkable fact have been noted before. Several workers (Nakashima et al., 1986; Klein and DeLisi, 1986; Klein and Somorjai, 1988) have demonstrated that amino acid composition can be used to predict the structural class (α, β, α/β, $\alpha+\beta$) of a protein. Muskal and Kim (1992) showed that accurate estimates of the ordered structure (α, β) content of proteins can be deduced from amino acid composition by using a tandem neural network technique.

This viewpoint is further supported by recent results of White and Jacobs (1993), who showed that actual protein sequences are slightly different from random sequences when represented by amino acid names but statistically indistinguishable from random sequences when viewed in terms of the lengthwise distributions of selected amino acid properties. The present work (in which amino acids are represented by their properties) demonstrates that actual sequences and random sequences with the same composition are also indistinguishable with respect to their correlation with actual three-dimensional structure.

It is appropriate to demonstrate that the real and random sequences considered herein are, like those considered by White and Jacobs (1993), indistinguishable in a statistical sense. We have used several tests to investigate this point. In the first, we examined sequence alignment similarity. We first considered all pairs of RAND B sequence sets. Corresponding random sequences in the members of each pair were aligned, and the number of amino acid matches was counted. The number of such matches in each comparison (of two sets each containing 114 random sequences) was determined and divided by the total number of residues in each set to give the fraction of matches. This exercise was repeated for alignments of the 114 actual sequences with the corresponding random sequences in each of the RAND B sets. We find that the average fraction of matches between two randomly generated sequences, $\langle \phi(\text{rand-rand}) \rangle = 0.07$, with standard deviation $\sigma(\text{rand-rand}) = 0.0019$. The average deviation

between the real-random comparisons and the average of the random-random comparisons,

$$Y = \langle[\phi(\text{real-rand})\text{-}\langle\phi(\text{rand-rand})\rangle]/\sigma(\text{rand-rand})\rangle, \quad (9)$$

over the 10 random sets of RAND B sequences, is $Y = 0.189$. Thus the average deviation between the real-random matchings and the random-random matchings is considerably less than σ(rand-rand), and there is no distinction between the real and RAND B sequences on the basis of sequence matching.

This test relied on a relatively small set of data, provided by the 10 RAND B sequences. We therefore carried out an additional test, in which it was possible to generate larger numbers of random sequences. For each of the 114 actual protein sequences in the data base, 100 random sequences were generated with the RAND B algorithm. Distances between all pairs of random sequences within the set of 100 were calculated by using Eq. 4 and full (rather than reduced) amino acid sequences. Distances were also calculated between the actual protein sequences and all 100 surrogate random sequences. The average real-random distance was compared with the average random-random distance. With the length scale of the sequence space $L = 4$, it was found that, for 92% of the sequences, the average real-random distance was within $\pm\sigma$(rand-rand) of the average random-random distance. Furthermore, 77% of the real-random distances were less than the average random-random distance. When $L = 2$, results were essentially the same. It was found that 87% of the real-random distances were within $\pm\sigma$(rand-rand) of the average random-random distance, and 89% of the real-random distances were less than the average random-random distance.

To determine whether this apparent similarity between real and randomly generated distributions represents actual similarity, we applied Student's t-test (Press et al., 1992). We first compared the variances of the real-random and random-random distance distributions in the same manner by using the F-test (Press et al., 1992). It was found that the variances of real-random distances were not significantly different from those of random-random distances at the 5% level for 64 proteins and at the 1% level for 86 proteins. For each of the 114 proteins, the average real-random distance was then compared with the average random-random distance by using the unequal variance t-test. It was found that the two averages were not significantly different for 33 proteins at the 5% confidence level and for 49 proteins at the 1% confidence level. We next compared each of the 114 ensembles of random sequences to one additional random sequence, generated with the RAND B algorithm. We will refer to the 114 additional random sequences as test sequences. For the test sequences, the variances were not significantly different from random-random variances at the 5% level in 64 cases and at the 1% level in 93 cases. The t-test was applied to a comparison of the average random-random distances with the average test-random distances. It was found that the two averages were not significantly different at the 5% level in 33 cases and at the 1% level in 45 cases. Thus, the comparison of the actual protein sequences with the ensemble of RAND B sequences gave results indistinguishable from those obtained when comparing randomly generated sequences to the RAND B sequences.

These results provide additional support for the suggestion that there is no statistically significant difference between actual protein sequences and random sequences generated with the same composition when the sequences are represented in terms of amino acid properties.

Having established this, we return to a consideration of our results. We first ask whether simple amino acid composition alone, in the absence of sequence information, is sufficient to produce the results we observe. The flexibility of the comparison method makes this computation straightforward. By choosing length scale $L = 1$ for the sequence distribution, which corresponds to the use of only composition information, we find a correlation between sequence and structure space $f_1 = 0.36$ when the amino acids are grouped into four clusters and 0.27 with no clustering. This is well below the range of values observed for $f_{\text{RAND B}}$. It therefore seems clear that composition alone is not sufficient to explain the similarity between RAND B sequences and actual protein sequences with respect to their correlation with structural similarity.

It should be noted, however, that the imposition of a composition constraint on randomly generated sequences automatically imposes corresponding constraints on the linear arrangement of the random sequences. It would seem that the signals that connect amino acid properties to the architecture of the folded polypeptide chain are built into the random sequences in the same way that they are in the actual sequences that they mimic. The elucidation of the nature of those signals is a task of major importance.

In this connection we note recent studies of the local code that has been postulated to govern protein folding (Rackovsky, 1993). In that work it was demonstrated that only ~60–70% of the 4-α-carbon sequence fragments in a protein encode for time-averaged structure. It was suggested, on the basis of sequence/structure correlations, that the remaining sequence fragments encode for conformational flexibility, the ability to adopt alternative conformations under the influence of long-range interactions. (We shall refer to these two types of sequence fragment as coding and noncoding, respectively, denoting their behavior with respect to time-averaged structure). It was additionally suggested that this flexibility is a specific feature of the folding code and is necessary for the proper folding of the molecule. (These observations also enable us to understand the low value of the best correlation coefficient between the sequence and structure spaces. If only 60–70% of the sequence encodes time-averaged structural information, considerable indeterminacy is introduced into the local folding code, and a high value of f is not to be expected.)

The approximate combinatorial independence of the sequence-structure correlation suggests that the size and number of coding and noncoding regions in a protein sequence may be major determinants of structure. The coding

character of a fragment is probably less sensitive to sequence randomization than other sequence characteristics. It has been noted by Lattman and Rose (1993) that neither efficient packing nor folding thermodynamics seems to determine the actual architecture of a folded protein. Rather, they suggest that a distributed control mechanism, with necessary information spread throughout the sequence, determines the shape of the folded protein. We suggest that the interplay between coding and noncoding sequence fragments may provide a realization of this mechanism. In future work, the nature of this interplay will be investigated. Questions of interest, for example, include the distribution of coding and noncoding segments along the sequence of different types of structures and the relationship between these factors and actual folding mechanisms.

Lau and Dill (1990) and Shakhnovich and Gutin (1990) have suggested that it is highly probable that a randomly synthesized polypeptide chain will have a preferred fold. This suggestion has been confirmed experimentally by Davidson and Sauer (1994). Our results reinforce this observation and raise the intriguing speculation that some subgroup of permutations of the sequence of a given protein might actually have folds approximating that of the naturally occurring sequence. Not all permuted sequences would be thermodynamically stable in the folded state (Lattman and Rose, 1993), but the information that determines architectural preference might be retained. One may reason that some permutations would be excluded from a particular fold due to the accumulation of unfavorable packing interactions (Zhang and Eisenberg, 1994). On the other hand, it has been shown (Behe et al., 1991) that efficient packing alone does not determine conformation. Recently, in fact, Shakhnovich and Gutin (1993) have used Monte Carlo techniques in sequence space, with the constraint that amino acid composition remain fixed, to search for stable, rapidly folding sequences of model proteins. In their technique, the interresidue interaction assumes a specified form, and sequence is varied. Our work, which includes the optimization of factor weighting, is equivalent to the variation of both sequence and interresidue interactions. It is therefore of particular interest to examine the actual values of the relative weightings that produce optimal correlation between the sequence and structure spaces.

It will be seen in Table 1 that the relative weights for almost all of the optimal correlations between actual sequence distances and the structure distances place strong emphasis on the first three factors. (The members of the four-amino-acid clusters arising from the optimal weighting are given in Table 3.) These factors relate to helix/bend preference, side chain bulk, and β-structure preference, respectively (Kidera et al., 1985a). Factor 8, which is related to preference for the helix region of the Ramachandran map, is also weighted. The fourth factor, which relates to the hydrophobicity properties of the amino acid, is slightly weighted in only one of the optimal correlations with actual sequence distances. This pattern is also followed in many of the RAND B correlations between randomly generated sequences with actual protein composition and the structure distances. Thus, relatively minor adjustments of factor weightings are sufficient to optimize the correlation between the structure space and the spaces of RAND B sequences.

TABLE 3 Amino acid clusters arising from optimal weighting

Cluster number	Cluster members
1	Ile, Arg, Val, Tyr
2	Gly, Asn
3	Cys, Pro, Ser, Thr
4	Ala, Asp, Glu, Phe, His, Lys, Leu, Met, Gln, Trp

The relative unimportance of hydrophobicity is consistent with recent results of Chan and Dill (1989a, b), Gregoret and Cohen (1991), and Hao et al. (1992), which demonstrate that ordered backbone structures can be formed as a result of spatial confinement of a stiff polypeptide chain. The picture of protein folding that emerges from that work and the present observations is one in which ordered structures form initially as a result of either local structural coding properties or hydrophobic collapse and the attendant chain confinement. However, the specific local structure that forms is not determined by the hydrophobicity property factor, which, by the nature of the factor analysis, is independent of conformational preference information and relates only to the general tendency of an amino acid to be on the inside or outside of the protein. Rather, local structure is determined by those property factors that are related to the local conformational properties of the chain. As a result, the optimal weightings for correlating the sequence and structure spaces take little account of the hydrophobicity factors associated with the amino acids.

SUMMARY

In this work, we have developed methods for the mathematical representation of protein sequences, which result in the description of a set of sequences as a collection of points in a hyperspace. We then investigated to what extent it is possible to correlate the structure of this sequence space with that of an independently developed hyperspace representing the relationships between protein structures. The following points were demonstrated:

- It is possible to represent protein sequences in a physically reasonable way in terms of the properties of the naturally occurring amino acids.
- With the use of an appropriately constructed metric function, it is possible to determine the relationship between sequences. This function has the property that very similar sequences are shown to be near one another and that the distance between sequences increases as the sequences become less similar. The distance function has the additional properties that it is independent of the molecular weights of the proteins being compared and that insertions and deletions are automatically corrected for, without the use of artificially constructed penalty functions.

- By attaching weights to the property factors used to represent the individual amino acids, and varying those weights to produce optimal correlation with the sequence space, an optimal correlation between the sequence and structure spaces can be developed.
- Randomly generated sequences with compositions approximating those of the actual proteins in the data base can produce correlations with the structure space as good as that produced by the actual protein sequences. This extends work of other investigators, who have shown that actual protein sequences are indistinguishable from random amino acid sequences with respect to the lengthwise distribution of properties, by demonstrating indistinguishability with respect to structure correlations.
- This observation leads to an interpretation of the folding process in terms of previously demonstrated characteristics of the folding code. In particular, it is suggested that an important determinant of the folded structure of a sequence is the linear relationship between sequence fragments that encode for time-averaged structure and those that do not but rather encode for structural flexibility under the influence of long-range interactions.
- It is shown that the factors most responsible for the optimal correlation between the sequence and structure spaces are helix/bend preference, side chain bulk, and β-structure preference. Hydrophobicity is not a significant factor in this correlation. It is pointed out that this is consistent with recent studies that suggest that ordered structures can be formed as a result of spacial confinement of the chain.
- These results establish a connection between random sequences and actual protein architectures and are fully consistent with suggestions by a number of workers that folding is not a unique property of the sequences of biologically occurring proteins but rather a general characteristic of amino acid copolymers.

We thank Profs. Robert S. Knox and Thomas Foster for many helpful discussions.

This work was supported by the Office of Naval Research under grant N00014–91-J-1943.

REFERENCES

Altschul, S. F. 1991. Amino acid substitution matrices from an information theoretic perspective. *J. Mol. Biol.* 219:555–565.

Altschul, S. F., and D. J. Lipman. 1990. Protein database searches for multiple alignments. *Proc. Natl. Acad. Sci. USA* 87:5509–5513.

Barton, G. J., and M. J. E. Sternberg. 1990. Flexible protein sequence patterns: a sensitive method to detect weak structural similarities. *J. Mol. Biol.* 212:389–402.

Behe, M. J., E. E. Lattman, and G. D. Rose. 1991. The protein folding problem: the native fold determines packing, but does packing determine the native fold? *Proc. Natl. Acad. Sci. USA* 88:4195–4199.

Blaisdell, B. E. 1986. A measure of the similarity of sets of sequences not requiring sequence alignment. *Proc. Natl. Acad. Sci. USA* 83:5155–5159.

Blaisdell, B. E. 1989a. Effectiveness of measures requiring and not requiring prior sequence alignment for estimating the dissimilarity of natural sequences. *J. Mol. Evolution* 29:526–537.

Blaisdell, B. E. 1989b. Average values of a dissimilarity measure not requiring sequence alignment are twice the averages of conventional mismatch counts requiring sequence alignment for a computer-generated model system. *J. Mol. Evol.* 29:538–547.

Blaisdell, B. E. 1991. Average values of a dissimilarity measure not requiring sequence alignment are twice the averages of conventional mismatch counts requiring sequence alignment for a variety of computer-generated model systems. *J. Mol. Evol.* 32:521–528.

Bryant, S. H., and C. E. Lawrence. 1993. An empirical energy function for threading protein sequence through the folding motif. *Proteins Struct. Funct. Genet.* 16:92–112.

Chan, H. S., and K. A. Dill. 1989a. Intrachain loops in polymers: effects of excluded volume. *J. Chem. Phys.* 90:492–509.

Chan, H. S. and K. A. Dill. 1989b. Compact polymers. *Macromolecules* 22:4559–4573.

Davidson, A. R., and R. T. Sauer. 1994. Folded proteins occur frequently in libraries of random amino acid sequences. *Proc. Natl. Acad. Sci. USA* 91:2146–2150.

Godzik, A., and J. Skolnick. 1992. Sequence structure matching in globular proteins: application to supersecondary and tertiary structure prediction. *Proc. Natl. Acad. Sci. USA* 89:12098–12102.

Gregoret, L. M., and F. E. Cohen. 1991. Protein folding: effect of packing density on chain conformation. *J. Mol. Biol.* 219:109–122.

Hao, M.-H., S. Rackovsky, A. Liwo, M. R. Pincus, and H. A. Scheraga. 1992. Effects of compact volume and chain stiffness on the conformations of native proteins. *Proc. Natl. Acad. Sci. USA* 89:6614–6618.

Henikoff, S., and J. G. Henikoff. 1992. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* 89:10915–10919.

Karlin, S., and S. F. Altschul. 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. USA* 87:2264–2268.

Kidera, A., Y. Konishi, M. Oka, T. Ooi, and H. A. Scheraga. 1985a. Statistical analysis of the physical properties of the 20 naturally occurring amino acids. *J. Protein Chem.* 4:23–54.

Kidera, A., Y. Konishi, T. Ooi, and H. A. Scheraga. 1985b. Relation between sequence similarity and structural similarity in proteins: role of important properties of amino acids. *J. Protein Chem.* 4:265–297.

Klein, P., and C. DeLisi. 1986. Prediction of protein structural class from the amino acid sequence. *Biopolymers* 25:1659–1672.

Klein, P., and R. L. Somorjai. 1988. Nonlinear methods for discrimination and their application to classification of protein structures. *J. Theor. Biol.* 130:461–468.

Lattman, E. E., and G. D. Rose. 1993. Protein folding- what's the question? *Proc. Natl. Acad. Sci. USA* 90:439–441.

Lau, K. F., and K. A. Dill. 1990. Theory for protein mutability and biogenesis. *Proc. Natl. Acad. Sci. USA* 87:638–642.

Lawrence, C. E., S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wooton. 1993. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* 262:208–214.

Leung, M.-Y., B. E. Blaisdell, C. Burge, and S. Karlin. 1991. An efficient algorithm for identifying matches with errors in multiple long molecular sequences. *J. Mol. Biol.* 221:1367–1378.

Maiorov, V. N., and G. M. Crippen. 1992. Contact potential that recognizes the correct folding of globular proteins. *J. Mol. Biol.* 277:876–888.

Miyazawa, S., and R. L. Jernigan. 1993. A new substitution matrix for protein sequence searches based on contact frequencies in protein structures. *Protein Eng.* 6:267–278.

Muskal, S. M., and S.-H. Kim. 1992. Predicting protein secondary structure content: a tandem neural network approach. *J. Mol. Biol.* 225: 713–727.

Nakashima, H., K. Nishikawa, and T. Ooi. 1986. The folding type of a protein is relevant to the amino acid composition. *J. Biochem.* 99: 153–162.

Needleman, S. B., and C. D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48:443–453.

Niefind, K., and D. Schomburg. 1991. Amino acid similarity coefficients for protein modeling and sequence alignment derived from main-chain folding angles. *J. Mol. Biol.* 219:481–497.

Pietrokovski, S., J. Hirshon, and E. N. Trifonov. 1990. Linguistic measure of taxonomic and functional relatedness of nucleotide sequences. *J. Biomol. Struct. & Dyn.* 7:1251–1268.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. 1992. Numerical Recipes in FORTRAN: The Art of Scientific Computing, 2nd ed. Cambridge University Press, Cambridge, UK. 609–613.

Rackovsky, S. 1990. Quantitative organization of the known protein x-ray structures. I. Methods and short-length-scale results. *Proteins Struct. Funct. Genet.* 7:378–402.

Rackovsky, S. 1993. On the nature of the protein folding code. *Proc. Natl. Acad. Sci. USA* 90:644–648.

Saqi, M. A. S., and M. J. E. Sternberg. 1991. A simple method to generate non-trivial alternate alignments of protein sequences. *J. Mol. Biol.* 219: 727–732.

Schuler, G. D., S. F. Altschul, and D. J. Lipman. 1991. A workbench for multiple alignment construction and analysis. *Proteins Struct. Funct. Genet.* 9:180–190.

Shakhnovich, E. I., and A. M. Gutin. 1990. Implications of thermodynamics of protein folding for evolution of primary sequences. *Nature* 346:773–775.

Shakhnovich, E. I., and A. M. Gutin. 1993. Engineering of stable and fast-folding sequences of model proteins. *Proc. Natl. Acad. Sci. USA* 90: 7195–7199.

Sippl, M. J., and S. Weitckus. 1992. Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a data base of known protein conformations. *Proteins Struct. Funct. Genet.* 13: 258–271.

Späth, H. 1980. Cluster analysis algorithms for data reduction and classification of objects. Ellis Horwood, Chichester, UK.

Van Heel, M. 1991. A new family of powerful multivariate statistical sequence analysis techniques. *J. Mol. Biol.* 220:877–887.

White, S. H., and R. E. Jacobs. 1993. The evolution of proteins from random amino acid sequences. I. Evidence from the lengthwise distribution of amino acids in modern protein sequences. *J. Mol. Evolution* 36:79–95.

Zhang, K. Y. J., and D. Eisenberg. 1994. The three-dimensional profile method using residue preference as a continuous function of residue environment. *Protein Sci.* 3:687–695.