On the existence and implications of an inverse folding code in proteins

(inverse folding problem/structure–sequence correlation/sequence coding)

S. RACKOVSKY
Department of Biomathematical Sciences, Mt. Sinai School of Medicine, One Gustave L. Levy Place, Box 1023, New York, NY 10029-6574

Communicated by Harold A. Scheraga, Cornell University, Ithaca, NY, March 27, 1995

ABSTRACT The existence of a code relating the set of possible sequences at a given position in a protein backbone to the local structure at that location is investigated. It is shown that only 73% of 4-Cα structure fragments in a sample of 114 protein structures exhibit a preference for a particular set of sequences. The remaining structures can accommodate essentially any sequence. The structures that encode specific sequence distributions include the classical "secondary" structures, with the notable exception of planar (β) bends. It is suggested that this has implications as to the mechanism of folding in proteins with extensive sheet/barrel structure. The possible role of structures that do not encode specific sequences as mutation hot spots is noted.

In previous work (1), the characteristics of the code that connects protein sequence with local folding have been investigated. It was demonstrated that a feature of that code is the fact that some sequence fragments have no intrinsic preference for a particular set of conformations. It was suggested that this characteristic of the code is a requirement for successful protein folding, since it guarantees that there exist sequence fragments that assume their native conformations in a particular protein under the influence of interactions external to themselves, allowing for correct assembly of the molecule. This characteristic of the code also explains the failure of secondary structure prediction algorithms to exceed ～70% accuracy (2), since it is found that the fraction of 4-residue sequence fragments that actually encode a preference for specific structures is 69%.

The methods developed in that work can be applied in an essentially identical manner to study the inverse problem—which structural features are characterized by an intrinsic preference for a particular set of sequences? A closely related problem, known as the inverse folding problem, has received considerable attention. The standard formulation of that problem is as follows: What sequence are compatible with a given protein structure (3–6)? The traditional approach to this question involves a consideration of packing relationships and long-range contacts. Short-range interactions are generally ignored by these algorithms, which attempt to predict the acceptability of specified sequences based on the assumption that main chain structure remains fixed as sequence is altered. Recent experimental evidence (7) suggests that this approach is not adequate. The present work, while conceptually related to the inverse folding problem, proceeds from a different viewpoint. I inquire into the attributes of the distributions of sequence fragments associated with particular types of structure fragment. I thus focus on local properties, rather than on long-range interactions, which are effectively averaged over in the data. I ask whether a local sequence code exists, analogous to the local folding code, that restricts the sequences which can be built into a given short structural fragment, and, if so, whether all structural fragments exhibit sequence coding behavior. I discuss the implications of such a code for protein folding.

METHODS

Analysis is based on the same experimental data used in previous work (1), a set of 114 protein structures selected to represent the entire set of structures determined by X-ray diffraction. The coding properties of 4-residue fragments, the shortest segments of the virtual-bond backbone that contain nonplanar information, are discussed. As in previous work (1), the sequences of these fragments will be represented in reduced form, using a sequence "alphabet" constructed by clustering the 20 amino acids according to their physical properties. Their structures are represented by using the generalized bond matrix representation (8) in terms of three parameters—virtual bond lengths, angles, and dihedral angles. The structural space available to a 4-Cα fragment is partitioned into subregions by dividing the ranges of these structural parameters as described (8).

The strategy pursued is precisely analogous to that used to investigate the direct folding code (1). The 4-Cα structure types occurring in the protein data base are enumerated at fairly high resolution (1, 8). In general, a given structure type will occur more than once at positions in proteins characterized by various sequences. Thus, any given structure type will be associated with a distribution of sequences. These sequence distributions are constructed for all structure types and each is compared to a set of randomly generated sequence distributions of the same size. I consider that a given structure type encodes a specific sequence preference if its associated sequence distribution is narrower than the average width of the set of randomly generated distributions. Those structure types whose associated sequence distributions are as broad as, or broader than, the average width of the set of random distributions of the same size are considered not to encode a sequence preference.

The width of the sequence distribution associated with a given structural type is measured by its entropy. The structure type under discussion is specified (for a 4-Cα fragment) by six indices, which specify the particular interval in its range into which each structural parameter falls. Thus, S(i, j, k, l, m, n) is the sequence entropy associated with a structure in which bond lengths 1, 2, and 3 fall into intervals i, j, and k, respectively, bond angles 1 and 2 fall into intervals l and m, and the dihedral angle occurs in interval n. This entropy is defined as

\[ S(i, j, k, l, m, n) = - \sum_{q=1}^{N} p_q \ln(p_q), \]  

[1]

where the shorthand index q runs over the sequence fragment types represented in the distribution, p_q is the fractional occurrence of fragment type q among all fragment types associated with the structure fragment (i, j, k, l, m, n), and N

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. §1734 solely to indicate this fact.
is the total number of sequence fragment types associated with 
(i, j, k, l, m, n). It is easily shown that

\[ O \leq S(i, j, k, l, m, n) \leq \ln(N), \]  

that the minimum value is assumed when the distribution is as sharp as possible, and that the maximum value is assumed when the distribution is uniform and featureless. The entropy therefore provides a useful measure of the sharpness or specificity of the sequence distribution associated with the structure parameters (i, j, k, l, m, n).

As noted above, I wish to compare the width of the sequence distribution associated with given structural parameters with those of randomly generated sequence distributions of the same size. The number of fragments in the data base characterized by structural indices (i, j, k, l, m, n) is denoted as \( N_{ijklnm} \). For each of a large set of physically relevant values of \( N_{ijklnm} \), 10,000 distributions were generated, each containing \( N_{ijklnm} \) sequence fragments chosen randomly without replacement from the protein data base. The entropy was calculated for each, and the average entropy of the 10,000 distributions was determined.

The condition formulated above defining the encoding of a sequence preference by a specified type of structure fragment can be written quantitatively in terms of an entropy difference function as follows:

\[ \delta(i, j, k, l, m, n) = \bar{S}_{R}(N_{ijklnm}) - S(i, j, k, l, m, n), \]  

where \( \bar{S}_{R}(N) \) is the average entropy of an ensemble of randomly generated distributions of sequence fragments, each containing \( N \) fragments. A positive value of \( \delta \) indicates that the average entropy of the ensemble of randomly generated distributions is greater than that observed for the actual distribution of sequence fragments associated with the structure fragment of interest and therefore corresponds to the condition for encoding of a structure preference. A zero or negative value of \( \delta \) corresponds to the lack of a preference for a specific set of sequences.

The average entropy, \( \bar{S}_{R}(N) \), is found to be well fit, over a wide range of physically relevant values of \( N \), by a polynomial in \( \log N \):

\[
\bar{S}_{R}(N) = 0.036 + 2.005[\log_{10}(N)] + 0.268[\log_{10}(N)]^2 \\
+ 0.276[\log_{10}(N)]^3 + 0.039[\log_{10}(N)]^4, \tag{4}
\]

with \( R^2 + 0.99985 \). This formula allows determination of \( \delta \) (Eq. 3) for any value of \( N_{ijklnm} \).

RESULTS AND DISCUSSION

It is found that 72% of structure fragments exhibit a preference for a particular set of sequence fragments by the coding criterion set forth in the preceding section. This result establishes the existence of an inverse protein folding code that is local in nature and demonstrates that it is similar in character to the direct folding code previously discussed (1). It implies that \( \sim 30\% \) of 4-Cα structure fragments are able to tolerate essentially any sequence without prejudice to local folding preference.

It is of interest to ask what structures are characterized by either high or low degrees of sequence coding. A convenient tool for addressing this question is the entropy difference \( \delta \) (Eq. 3) measured as a function of the virtual bond dihedral angle characteristic of 4-Cα units. (It can be shown by inspection of the data that bond lengths and bond angles are not revealing variables in this respect.) Fig. 1 shows a scatterplot of \( \delta \) as a function of dihedral angle. It will be seen that both positive and negative values of \( \delta \) are found for all values of the dihedral angle, arising from structures with differing values of the other structural parameters—particularly the bond angles. This plot, however, gives no information as to the relative

![Fig. 1. Distribution of the entropy difference (Eq. 3) as a function of backbone dihedral angle of 4-Cα units. The range of the dihedral angle, 0 ≤ γ ≤ 360°, is divided into 20° intervals, labeled by n. The interval n contains values of γ falling in the range of (n − 1) × 20° ≤ γ < (n) × 20°.](image-url)
populations of the different species. Rather, it can be thought of as enumerating the different coding states available to 4-C seen fragments with dihedral angles in specified intervals. These states will be of different energies, and a useful empirical indicator of these energies is their relative populations. Fig. 2 shows a plot of the population-weighted average of \( \delta \),

\[
\delta_n = \frac{\sum_{i,j,k,l,m} N_{ijklm} \delta(i, j, k, l, m, n)}{\sum_{i,j,k,l,m} N_{ijklm}}
\]

as a function of dihedral angle interval. It can be seen that there are three peaks in \( \delta_n(n) \). It should be remembered that, by nature of the bond dihedral angle, a cyclic boundary condition exists in \( n \), so that \( \delta_n(m + 18) = \delta_n(m) \). These peaks occur at values of \( n \) that correspond to three of the classical secondary structures. The peak at \( 17 \leq n \leq 3 \) includes left- and right-handed and flat extended structures (9, 10). The large peak at \( 6 \leq n \leq 8 \) arises from right-handed helix/bend structures, and the small peak at \( n = 12 \) is from left-handed bend fragments.

This correspondence would seem to be physically significant and to reflect the rules these structures play in the folding process. In regions where \( \delta_n(n) > 0 \), the majority of structure fragments have \( \delta(i, j, k, l, m, n) > 0 \) (3). It is well known that the population distribution of amino acid properties mirrors the energetics of those properties (11). We therefore conclude that, in regions where \( \delta_n(n) > 0 \), structures with \( \delta > 0 \) are of lower energy than others. But these structures have a preference for a particular set of sequences. This implies that they require a particular set of interactions for their formation. In regions where \( \delta_n(n) \leq 0 \), the majority of structures—the low-energy structures—have \( \delta \approx 0 \) and show no sequence preference. It follows that these structures do not form because of particular internal interactions but rather arise under the influence of interactions external to themselves—long-range interactions.

Reference to Fig. 2 reveals that near-planar bends behave in a manner that is qualitatively different from other ordered backbone structures. In the region \( 9 \leq n \leq 11 \), \( \delta_n(n) < 0 \). This implies that these bends, which are generally associated with sheet/barrel structures, are formed as a result of long-range interactions and that the formation of planar turns is probably not an initiating step in the formation of sheet/barrel structures. On the other hand, the twisted helix/bend structures characteristic of helical proteins are likely to form at an early stage in folding and act as initiators of the folding process. Dyson et al. have observed behavior consistent with this viewpoint in studies of myoglobin (12) and plastocyanin (13).

The presence in folded proteins of structure fragments that do not encode particular sequences necessarily occur and participate in a role in sequence evolution. At these positions, mutations can occur without violating constraints imposed by local folding. In positions in the structure where sequence constraints are encoded, certain mutations are possible only at a cost in local conformational energy or when accompanied by local conformational rearrangement. At the same time, it must be remembered that long-range ("packing") constraints also play a significant role in determining the acceptability of a mutation. A comparison of the local inverse folding code with the characteristics of aligned sequences within families of homologous proteins of known structure, or with the results of cassette mutagenesis experiments (14), should give useful information as to how mutations may occur and the relative weights that must be attached to short- and long-range constraints in this regard.

In this work, I have demonstrated the existence of a local inverse folding code in proteins and suggested that the character of that code carries implications as to the nature of the folding process in helical and sheet/barrel structures. I have also suggested a connection between the inverse local folding code and protein evolution. The implications of these findings remain to be explored.

I would like to thank Profs. Craig Benham and Istvan Sugar for helpful comments. I am grateful for support of this work by the Office of Naval Research under Grant N00014-91-J-1943.