Proc. Natl. Acad. Sci. USA
Vol. 90, pp. 644–648, January 1993
Biophysics

On the nature of the protein folding code

(structure prediction/entropy/sequence fragment types)

S. RACKOVSKY

Department of Biophysics, School of Medicine and Dentistry, University of Rochester, Rochester, NY 14642

Communicated by H. A. Scheraga, October 5, 1992

ABSTRACT **This paper investigates quantitatively the characteristics of the local folding code. The overlapping four-residue fragments which make up the amino acid sequences of 114 proteins are divided into classes on the basis of the physical properties of their constituent amino acids. The distribution of structural types associated with each class of sequence fragment is determined and compared with an ensemble of random structural distributions of the same size selected from the actual protein structures. A criterion is proposed, based on the relative entropies of the two types of distribution, and on a hypothesis as to the characteristics of fragments which code for local structure, that makes it possible to identify those four-residue sequence elements which encode specific time-averaged structure. It is determined that, by this criterion, only 60–70% of the four-residue fragments encode specific structures. It is suggested that the remaining sequence fragments intrinsically encode susceptibility to conformational alteration under the influence of long-range interactions and that this susceptibility is required for correct folding of the molecule. This feature introduces an inherent indeterminacy into the local folding code. The implications of this observation for the prediction of protein structure by various methods are briefly discussed.**

It has been known since the classic experiments of Anfinsen *et al.* (1) that the protein folding process is controlled by the amino acid sequence of the molecule. This observation has led to great interest in the folding process and in the problem of predicting protein structure from amino acid sequence. Efforts to carry out such predictions can be divided into two broad classes: energy-based calculations and code-based prediction schemes.

Energy-based methods in their pure form make no prior assumptions about the coding properties of the amino acids but rather attempt to locate the global minimum in the free-energy surface of the protein molecule, which, it is reasonably assumed, will correspond to the native conformation of the molecule.

Code-based methods assume, either implicitly or explicitly, that the protein folding code must be a 1:1 correspondence between amino acid and single-residue structure, or, in the worst case, an *m:n* correspondence between some small number *m* of amino acids and a limited number *n* of local structural possibilities. This picture parallels the observed coding properties of the nucleotides which carry the information necessary for manufacturing the protein sequence. In that case, a simple 3:1 code connects DNA sequence with protein sequence. Attempts to delineate such a relationship for protein folding have resulted in a number of secondary-structure prediction schemes (2–7), as well as pattern-recognition (8, 9) and neural-net (10–12) algorithms.

It is clear that a folding code, in some general sense, must exist. After all, proteins *do* fold without the aid of external agents. Nevertheless, the problem of extracting that code has proven extraordinarily refractory. Any proposed code must address two observations: (*i*) None of the code-based prediction schemes put forward to date give better than 60–70% agreement with experimental observation (13). (*ii*) It is not uncommon for a particular sequence fragment to be associated with more than one structure (14).

It has been suggested (15, 16) that these discrepancies arise from long-range interactions, in which the conformation of a given sequence fragment is influenced by the spatial proximity of residues which are distant from it along the chain. This viewpoint raises a question about the nature of the local folding code. Rather than assuming an intrinsic structural preference for every sequence element, one may ask whether certain sequence fragments have conformational energy surfaces tailored to allow conformational diversity within folded proteins. Such sequence elements would introduce indeterminism into a local folding code (though not, of course, into protein structure), and their existence would have significant implications for code-based structure prediction.

The existence of such noncoding elements would be entirely consistent with the results of folding simulations by Skolnick and Kolinski (17), as well as with the general chain-nucleation mechanism proposed by various workers (18) and with the suggestion of Chan and Dill (19, 20) that ordered backbone structures form as a result of chain confinement. These models suggest that folding is a process in which small fragments of the initially fluctuating, unfolded chain assume conformations which are relatively long-lived, and these act as nuclei which cause the conformational stabilization of other regions of the molecule, until the final, native state is reached by a cascade mechanism.

One may also ask whether the encoding unit in the protein folding code is, in fact, the single amino acid. The monomer unit is not the encoding unit in DNA, and the possibility that this is also the case for proteins cannot be overlooked.

This paper investigates the local coding properties of protein sequences, expressed either as single-amino-acid sequences or as dipeptide sequences, building upon methods which were previously developed for the classification of protein structures (21, 22). It is suggested, on the basis of these investigations, that the protein folding code is indeed not a simple local structural code, but rather more general in nature.

METHODS

The approach is readily summarized. The sequence fragments of a large set of proteins are divided into a number of classes on the basis of physically reasonable criteria. The structural properties of the members of each class of sequence units are examined and compared with the properties of randomly generated sets of fragments. This comparison will be the basis for conclusions as to the coding properties of protein sequences.

The sequences and conformations of 4-C^{α} backbone fragments are examined here. These were chosen because they

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.

are the shortest fragments in the virtual-bond backbone which can be said to be folded—i.e., the shortest fragments which contain nonplanar information. Furthermore, the conformation of these fragments is controlled by two nearest-neighbor pairs of (ϕ, ψ) angles. Therefore, the information developed herein is completely distinct from that resulting from more traditional studies of amino acid conformations.

Sequence/Structure Data Base. The set of protein structures and associated sequences used in this work is the same as that used in previous investigations (22), with the exception of 9 proteins for which no detailed sequence (or only a partial sequence) was available. [The omitted proteins (in the Brookhaven Protein Data Bank identification code) are 2BCL, 156B, 155C, 1PGI, 2YHX, 1KGA, 1PEP, 2PGK, 2TNC.] This data base was constructed to be well representative of the entire set of proteins for which an acceptable structure and sequence are both available. It contains information on 20,004 amino acids in 114 proteins.

Representation of Amino Acids. The first technical question to be addressed is the method adopted for the representation of protein sequences. In previous work (22) the use of a generalized bond matrix representation was demonstrated for protein structure fragments of arbitrary size. The methods which were set forth there can also be used for the representation of protein sequences, once the parameters which describe particular amino acids are chosen.

We desire to represent amino acids by their physical properties, rather than simply as a list of names. Instead of concentrating on a particular property, or small group of properties, chosen in a manner which is not entirely objective, we would like to find a method for representing amino acids in terms of *all* their physical characteristics. This can be done by using the results of Kidera *et al.* (23, 24), who carried out a factor analysis of essentially all the physical properties which have been attributed to the 20 amino acids. They demonstrated that these properties are representable by 10 property factors, which account for 86% of the variance of all the physical properties. Their approach also eliminates concerns arising from possible correlations between different, independently derived property sets. Therefore, an amino acid X is represented as a vector,

$$\mathbf{X} = (x_1, x_2, \ldots, x_{10}), \qquad [1]$$

of its 10 property factors x_i.

With this representation, one can define a 10-dimensional Euclidean factor space in which each amino acid corresponds to a point whose coordinates are given by the 10 components of the amino acid vector. One can further define a distance between two amino acids X and Y as the Euclidean distance between their representative points:

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

This function has the desirable property that it increases as the difference between the physical characteristics of the two amino acids increases, so that the mathematical distance accurately reflects the physical differences between any two amino acids.

Grouping of Amino Acids. The definition of this distance function enables one to divide the amino acids into classes with similar physical characteristics. The procedure is straightforward. The number of occurrences of each amino acid in the sequence/structure data base is plotted as a function of the position of the amino acid in the factor space. The maxima on the resulting surface (which correspond to those amino acids which are present in greater number than others which have similar characteristics) are identified, and each of the remaining amino acids is assigned to that maximum to which it is closest. This divides the entire set of 20 amino acids into a limited number of groups, the members of which have similar physical properties. The actual amino acid sequence of any protein can then be rewritten as a reduced sequence by using the number of the group to which each specific amino acid belongs. This procedure has the advantage of mitigating statistical problems likely to arise in the subsequent analysis due to the difficulty of matching the less common sequence fragments. Furthermore, this sequence reduction is done in a manner which takes into account the physical properties of the amino acids in a quantitative fashion.

In practice, the 20 amino acids are divided into five groups by this procedure. (The distribution of the amino acids among these groups is given in Table 1.) There are therefore 625 possible four-residue sequence fragment types. Of these, 473 are actually observed to occur in the data base.

Clustering of Dipeptide Fragments. The approach detailed above can be generalized to give the distance between sequence fragments of any length. In addition to distances between single amino acids, we shall be interested in the distance between dipeptide fragments. The distance between two such fragments WX and YZ can be defined as

$$\Delta(WX, YZ) = [\Delta^2(W, Y) + \Delta^2(X, Z)]^{1/2}. \qquad [3]$$

This function makes it possible to group dipeptide fragments in the same way that we grouped single amino acids. In this representation, a sequence is regarded as a set of overlapping dipeptide fragments. As before, we count the number of occurrences of each dipeptide fragment in the data base and plot that number as a function of the position of the dipeptide in the factor space defined by Eq. 3. The population maxima (corresponding to those dipeptides which are present in greater number than others with similar characteristics) are identified, and other dipeptides are assigned to the maximum to which they are closest (i.e., to which they are most similar in their physical characteristics). In this case, it is found that there are three maxima, corresponding to the dipeptides Gly-Ser, Ser-Gly, and Ala-Ala. The assignment of the 400 dipeptides to these three groups is specified in Table 2. For the purposes of the present work, note that the sequence of a protein can be rewritten as a reduced *dipeptide* sequence in terms of these three groups, in a manner exactly analogous to the reduction of the single-amino-acid sequence described above. There are 27 possible dipeptide sequence fragment types, all of which occur in the data base.

Determination of Structural Entropies. Associated with each sequence in the data base is an x-ray structure. This allows examination of the structural characteristics of the reduced sequences resulting from the grouping procedures described above. For each four-residue sequence fragment type in the data base, the associated distribution of 4-C^α structure types is examined. Thus, for example, one might examine the range of structures associated with the single-residue reduced-sequence fragment 1543 (in which the numbers label amino acids in the groups detailed in Table 1). The protein structures are described by using the generalized

Table 1. Division of the 20 amino acids into groups

Group 1	Group 2	Group 3	Group 4	Group 5
Gly	**Ala** Val	**Pro**	**Trp**	**Lys**
	Ser Asn		Tyr	Gln
	Thr Phe			His
	Ile Glu			Arg
	Asp Cys			Met
	Leu			

The member of each group at which the population maximum falls is indicated in bold type.

Table 2. Division of the 400 dipeptides into groups

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

The first amino acid in each dipeptide is the row index, the second the column index. The table entries are the group numbers. The entries for the three dipeptides which are population maxima (see text) are underlined.

bond matrix representation (22), in terms of virtual bond lengths, angles, and dihedral angles. The structural space associated with a 4-C^{α} fragment is divided into subregions, by subdividing the ranges of the three structural parameters—virtual bond length, virtual bond angle, and virtual bond dihedral angle—in a manner described previously (22). With that subdivision, there are 17,496 theoretically possible subregions of the structural space, of which 587 are actually found to be occupied.

To examine the coding characteristics of four-residue sequence fragments, we shall compare the observed distribution of structures associated with each type of sequence fragment to a large ensemble of random structural distributions of the same number of fragments, generated from the actual structural data base. It is postulated that those sequence fragments whose associated structural distributions are as broad as, or broader than, the corresponding random structural distributions do not code for local structure, while those sequence fragments with structural distributions narrower than the corresponding randomly generated distributions do carry structural coding information.

To carry out this program, we need to describe the characteristics of the structural distribution associated with each sequence-fragment type. This is conveniently accomplished by using the entropy of the distribution. (Entropy has been used as a descriptor of various distributions associated with proteins. See, for example, ref. 25 and references therein.) The structural entropy $S(i, j, k, l)$ associated with the sequence fragment $ijkl$ is defined as

$$S(i, j, k, l) = -\sum_{m=1}^{M} p_m \ln(p_m), \qquad [4]$$

where m runs over the structural subregions (22) which are represented in the distribution associated with $ijkl$, p_m is the fractional occupation of subregion m, and M is the total number of occupied structural subregions in the distribution. It is readily shown (26) that

$$0 \leq S(i, j, k, l) \leq \ln M. \qquad [5]$$

The minimum value is assumed when the distribution is as narrow as possible (i.e., when only one subregion is occupied), and the maximum value is assumed when the distribution is completely uniform and featureless. Therefore, the entropy provides a measure of the sharpness of the distribution, which corresponds physically to the structural specificity exhibited by fragments whose sequence is $ijkl$. The comparison of the sharpness of the observed and random distributions can thus be accomplished by comparing the entropies of the distributions.

One may also view a 4-C^{α} sequence fragment as a sequence of three overlapping dipeptide fragments, $\alpha\beta\gamma$. (We denote dipeptide groups by Greek letters, to distinguish them from the single-amino-acid groups.) One can perform precisely the same entropy calculation on the reduced dipeptide sequence. For reasons noted below, we will be interested in those results as well as the entropies arising from the single-amino-acid sequence representation.

Calculation of Entropies of Randomly Generated Distributions. As noted above, we wish to compare the value of $S(i, j, k, l)$, the structural entropy of the observed distribution of N_{ijkl} sequence fragments of reduced sequence type $ijkl$, with the entropies associated with an ensemble of randomly generated structural distributions of N_{ijkl} fragments. These entropies can be determined by direct simulation. This was done by generating 10,000 distributions, each containing N_{ijkl} structural fragments chosen at random, without replacement, from the observed total distribution of structural fragments for the 114 proteins in the data base. The entropies were determined as a function of N_{ijkl}, for a range of physically relevant values of N_{ijkl}.

To quantitatively formulate the condition for structural coding, we define the difference function

$$\delta(i, j, k, l) = S_R(N_{ijkl}) - S(i, j, k, l), \qquad [6]$$

where $S_R(N_{ijkl})$ is the average entropy of an ensemble of randomly generated distributions, each containing N_{ijkl} fragments. When this difference function is positive, the observed structural entropy associated with the sequence indices $ijkl$ is less than the average entropy of the ensemble of randomly generated distributions. Thus, $\delta > 0$ corresponds to a quantitative formulation of the condition which was set forth above for a sequence fragment type to code for structure. This seems the least restrictive definition of local coding which can be written in terms of the distribution entropies,

and therefore gives a reasonable upper bound on the degree of local coding which occurs.

The function $S_R(N_{ijkl})$ is well fit over the entire range of N_{ijkl}, $1 < N_{ijkl} \leq 7022$, by a polynomial in $\log_{10}N_{ijkl}$:

$$S_R(N_{ijkl}) = -0.005929 + 2.1304(\log_{10}N_{ijkl}) + 0.0019242(\log_{10}N_{ijkl})^2 - 0.16054(\log_{10}N_{ijkl})^3 + 0.022651(\log_{10}N_{ijkl})^4, \quad [7]$$

with $r^2 = 0.9998$. This fit enables us to determine the difference function for any value of N_{ijkl}.

It should be noted that, although the remarks in this section have been cast in the notation associated with the single-residue representation of sequences (using indices *ijkl*), the entire formulation is equally true for the dipeptide sequence representation. Observe particularly that Eq. 7 is true when N_{ijkl} is replaced by $N_{\alpha\beta\gamma}$, since the information it carries is purely structural and does not depend on sequence representation. The use of both the single-amino-acid and dipeptide representations constitutes a check on the validity of our results, since this is equivalent to dividing the same data into classes in two different ways and performing the subsequent calculations twice independently.

RESULTS

The methods outlined in the preceding section have been applied to the sequence/structure data base. Using the condition of Eq. 6, in the single-residue representation, 69% of the sequence fragments code for structure. In the dipeptide representation, 60% code for structure. The remaining 30–40% do not. It seems likely, as suggested above, that those sequence fragments which do not code for local structure have conformational energy surfaces which are strongly perturbed by neighboring backbone segments. As a result, their time-averaged structure in the folded protein is governed by nonlocal interactions. This introduces an element of indeterminacy into the local folding code. This indeterminacy is an inherent feature of the code and is necessary for the proper self-assembly of the molecule.

This is not merely a semantic issue. This observation has significant implications for attempts to predict protein structure from sequence. It suggests that approaches based on traditional codes—secondary structure predictions, pattern recognition, etc.—are intrinsically limited in their accuracy. Energy-based approaches have the potential to achieve greater accuracy, because the lack of any *a priori* assumptions as to the properties of various residues makes it possible to adjust the conformations of structurally noncoding residues to accommodate the preferences of the structurally coding sequence segments. Indeed, Segawa and Richards (27), working with protein x-ray structures, have suggested a method for identifying flexible regions in proteins. Their viewpoint is fully consistent with the picture of coding proposed herein.

The degree of coding which is observed is remarkably consistent with the maximum accuracy exhibited by various secondary-structure prediction algorithms (13). It should be remembered that this work analyzes the sequence/structure relationship on a different length scale than those methods (four-residue vs. single residue), so that precise correspondence is not to be expected. This makes the observed agreement all the more striking.

The present work goes beyond previous discussions of long-range interactions (15, 16) in suggesting that the conformational properties which lead to the observed effects of long-range contacts are inherent in the local folding code. Susceptibility to conformational rearrangement as a result of those contacts must be encoded in specific sequence elements, in the same way that preferences for time-averaged conformation are. The present approach enables one to identify the particular sequence elements in which this susceptibility is manifest.

It has been suggested that another reason for the observed limitations on prediction accuracy is that the available protein structure data base is not large enough to provide accurate prediction (15, 16). It is, however, not possible to know whether, if many more structures were available, the observed fragment structure distributions would be substantially different. In the present work, each actual structure distribution is compared with a large ensemble of randomly generated distributions of the same size, drawn from the same data base. The effect of this procedure is to automatically correct for the size of the data base by providing answers based on the statistical expectation for the available data.

The approach here outlined is capable of determining not only the degree to which protein sequence fragments code for three-dimensional structure but also which fragment types code for structure and which structures they encode. The details of these relationships, as well as the relationship between the single-amino-acid and dipeptide encoding properties, and the distribution of coding and noncoding regions in proteins, will be discussed elsewhere. It will also be of interest to compare the specific conformational properties of peptide fragments which code for local structure with those of peptides which do not.

I am grateful to Prof. William Simon for most helpful discussions on curve fitting and to Runa Rahman and Prof. Robert S. Knox for stimulating conversations. I thank Profs. W. A. Bernhard, B. M. Goldstein, D. A. Goldstein, T. Gunter, and George Némethy for commenting on the manuscript, the anonymous referees for thoughtful suggestions, and the Office of Naval Research for support of this work under Grant N00014-91-J-1943.

1. Anfinsen, C. B., Haber, E., Sela, M. & White, F. H., Jr. (1961) *Proc. Natl. Acad. Sci. USA* **47**, 1309–1314.
2. Chou, P. Y. & Fasman, G. D. (1978) *Adv. Enzymol.* **47**, 45–148.
3. Pain, R. H. & Robson, B. (1970) *Nature (London)* **227**, 62–63.
4. Garnier, J. & Robson, B. (1989) in *Prediction of Protein Structure and the Principles of Protein Conformation*, ed. Fasman, G. D. (Plenum, New York), pp. 417–465.
5. Levin, J. M., Robson, B. & Garnier, J. (1986) *FEBS Lett.* **205**, 303–308.
6. Nishikawa, K. & Ooi, T. (1986) *Biochim. Biophys. Acta* **871**, 45–54.
7. Sweet, R. M. (1986) *Biopolymers* **25**, 1565–1577.
8. Schiffer, M. & Edmundson, A. B. (1967) *Biophys. J.* **7**, 121–135.
9. Biou, V., Gibrat, J. F., Levin, J. M., Robson, B. & Garnier, J. (1988) *J. Protein Eng.* **2**, 185–191.
10. Qian, N. & Sejnowski, T. J. (1988) *J. Mol. Biol.* **202**, 865–884.
11. Bohr, H., Bohr, J., Brunak, S., Cotterill, R. M. J., Lautrup, B., Norskov, L., Olsen, O. H. & Petersen, S. B. (1988) *FEBS Lett.* **241**, 223–228.
12. Holley, H. W. & Karplus, M. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 152–156.
13. Garnier, J. & Levin, J. M. (1991) *Comp. Appl. Biosci.* **7**, 133–142.
14. Kabsch, W. & Sander, C. (1984) *Proc. Natl. Acad. Sci. USA* **81**, 1075–1078.
15. Rooman, M. J. & Wodak, S. J. (1991) *Proteins Struct. Funct. Genet.* **9**, 69–78.
16. Rooman, M. J. & Wodak, S. (1988) *Nature (London)* **355**, 45–49.
17. Skolnick, J. & Kolinski, A. (1990) *Science* **250**, 1121–1125.
18. Kim, P. & Baldwin, R. (1982) *Annu. Rev. Biochem.* **51**, 459–489.
19. Chan, H. S. & Dill, K. A. (1989) *J. Chem. Phys.* **90**, 492–509.

20. Chan, H. S. & Dill, K. A. (1989) *Macromolecules* **22**, 4559–4573.
21. Rackovsky, S. & Goldstein, D. A. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 777–781.
22. Rackovsky, S. (1990) *Proteins Struct. Funct. Genet.* **7**, 378–402.
23. Kidera, A., Konishi, Y., Oka, M., Ooi, T. & Scheraga, H. A. (1985) *J. Protein Chem.* **4**, 23–54.
24. Kidera, A., Konishi, Y., Ooi, T. & Scheraga, H. A. (1985) *J. Protein Chem.* **4**, 265–297.
25. Shenkin, P. S., Erman, B. & Mastrandrea, L. D. (1991) *Proteins Struct. Funct. Genet.* **11**, 297–313.
26. Brillouin, L. (1962) *Science and Information Theory* (Academic, New York).
27. Segawa, S.-I. & Richards, F. M. (1988) *Biopolymers* **27**, 23–40.