Influence of the local amino acid sequence upon the zones of the torsional angles φ and ψ adopted by residues in proteins

ARTICLE *in* BIOCHEMISTRY · MARCH 1991

Impact Factor: 3.02 · DOI: 10.1021/bi00220a019 · Source: PubMed

Influence of the Local Amino Acid Sequence upon the Zones of the Torsional Angles φ and ψ Adopted by Residues in Proteins[†]

Jean-Francois Gibrat, Barry Robson,[‡] and Jean Garnier*

Unité d'Ingénierie des Protéines, Bâtiment des Biotechnologies, INRA, Jouy-en-Josas 78352 Cedex, France

Received April 26, 1990; Revised Manuscript Received August 31, 1990

ABSTRACT: A set of parameters is derived to express the influence of the local amino acid sequence on the torsional angles φ and ψ adopted by each residue in a protein. The formalism used, which is based on information theory, evaluates the probability for a given residue to be in a particular zone of the Ramachandran map. Comparisons with crystallographic structures suggest that the method can extract almost all of the available information from the local sequence and show that the local sequence carries only, on average, about 65% of the information necessary for specifying the conformation of a given residue in a protein. The rest is specified by long-range interactions that are specific for each protein fold. The parameters derived here provide a more detailed description of the prediction than other methods in allowing the allocation of the torsional angles for residues having an aperiodic structure and are intended to be used for directing the conformational search in a subsequent simulation of the three-dimensional structure. This method should also predict segments of the polypeptide chain that are the most stable and thus less sensitive to long-range interactions.

One of the most important problems of molecular biology is the elucidation of the relationship between the amino acid sequence and the three-dimensional structure of proteins. One step toward this goal would be realized if one could make an accurate prediction of the secondary structure. Many algorithms based on different principles have been proposed. An extensive review of which has been published by Fasman (1989). However, a survey of the most recent algorithms (Garnier & Levin, 1990) shows a tendency for the accuracy of the predictions to be limited to around 65% of residues correctly predicted if three states are used for the secondary structure (α helix, β sheet, and coil or aperiodic structure). It is interesting that several of these methods are of the "neural net" type. There is a strong relation between neural net approaches and the information theory approach used here, except that the information theory approach "preselects" the cross-interactions between residues and neglects those interactions expected to carry zero information. This renders the information theory computationally much more efficient.

Secondary structure prediction algorithms are based on short-range interactions; i.e., they use the information drawn from the local amino acid sequence. This is usually about 10 residues or so on each side of the residue to be predicted. It is generally admitted that the conformation of a residue in a protein depends also on long-range interactions, i.e., interactions between residues far along the sequence but brought close to the residue in question during the folding process. Nonetheless the relative magnitudes of the short-range and long-range interactions in influencing the conformation of a residue are not known precisely, nor is the reason for the lack of knowledge clear. If the accuracy of the prediction is limited to about 65%, is it because the short-range interactions contribute, on average, only this amount of information to the conformation of a residue? Alternatively, is it simply that the short-range interactions carry more information but the prediction methods, so far, are unable to extract it? These two possibilities are not, of course, mutually exclusive.

Deeper understanding would be valuable in view of the increasing usefulness of secondary structure prediction. Apart from the fact that there is relatively little else a genetic engineer can do when obtaining a new sequence (except for homology scanning in which comparison of predicted secondary structure can also play a role!) there are several specific applications. Examples include (a) predicting antigenic determinants whose structures could be mimicked by peptides (Pfaff et al., 1982) and (b) predicting the secondary structure of the signal sequence for protein exportation (Garnier et al., 1980, Emr et al., 1983), (c) use in sequence alignment (Levin & Garnier, 1988), or rational site-directed mutagenesis. Furthermore, the secondary structure predition can be the first step of a subsequent modeling of the three-dimensional structure by the combination of regular secondary structures (α helix and β sheet) into super-secondary structures (Taylor & Thornton, 1983) or by an energy minimization (Robson et al., 1985) and can also be used to weight the conformational search throughout the simulation (Robson & Pain, 1973). For these latter modeling cases, one needs to translate the secondary structure of a residue into a set of Cartesian coordinates (external coordinates) or, more conveniently, into a pair of φ and ψ angles (internal coordinates). It is straightforward for α-helical and β-sheet secondary structures to obtain approximate values of φ and ψ angles but for the residues predicted as coil this is not possible.

In this work, we wanted first to derive a set of parameters predicting those zones, defined by torsional angles φ and ψ on the Ramachandran map, which are populated by residues in proteins. The method is based on the information theory, but here we present a new aspect with the introduction of the fraction of residues in a given state rather than use of the information value. The fraction represents an estimate of the probability for a residue to be in one of the zones of the

[†] This research was financed in part by a grant from the Biotechnology Action Program of the European Economic Community (BAP-0148F).

* To whom correspondence should be addressed.

[‡] Present address: Proteus Biotechnology Ltd., Proteus House, 48 Stockport Road, Marple SK6 6AB, UK, and Department of Biochemistry and Molecular Biology, University of Manchester, Manchester M13 9PT, UK.

0006-2960/91/0430-1578$02.50/0

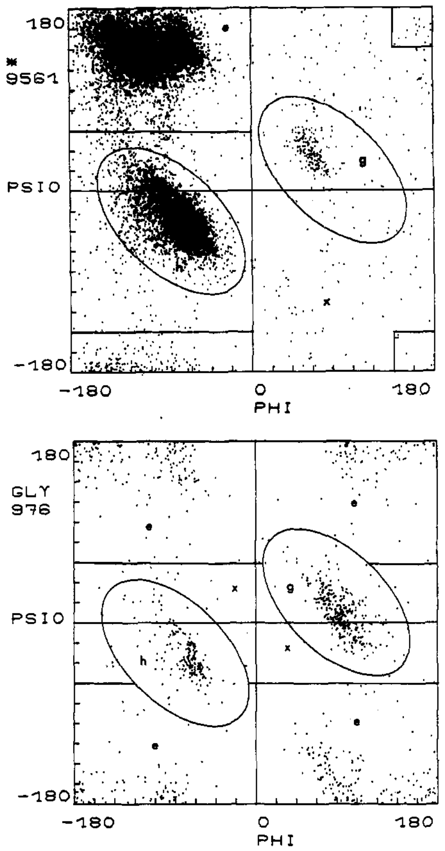

FIGURE 1: Distribution of the φ and ψ angle pairs. This figure shows the distribution of the φ and ψ angle pairs in the Ramachandran map from the proteins listed in Table I. The four zones, h, e, g, and x, are represented for all the amino acid residues except Gly (top) and for Gly (bottom) corresponding to 9561 and 976 residues, respectively.

Ramachandran map given a local sequence.

The second point we wanted to study, using the results of the predictions, is whether the method is able to extract from the data base all or at least most of the information included in the local amino acid sequence. If this is the case, we can assess the respective influence of the short-range and long-range interactions on the conformation adopted by a given residue in the protein. Such matters are difficult to resolve in view of the interelation (in practice) between the information in the system studied and that accessible to the observer, but we believe that the following is a useful first attempt that will stimulate further thoughts and study.

METHOD

(*a*) *Development of the Data Base.* We used 61 proteins (66 polypeptide chains) from the Brookhaven Protein Data Bank of Bernstein et al. (1977) with a resolution of less than 2.8 Å and a crystallographic factor R less than or equal to 0.25. These limits represent a compromise between the accuracy of the data and the amount of data available. We computed, using the atomic coordinates, the φ and ψ angles for each residue. The points corresponding to pairs of φ and ψ angles were then plotted on a Ramachandran map for each of the 20 residue types. We defined for the residues a number of conformational zones, trying to fulfil two contradictory conditions: to obtain the smallest zones in the Ramachandran map but still keep the maximum number of residues in them. As long as the two conditions are respected, results are insensitive to the shape of the zones. The most convenient definition of the zones are as follows (see also Figure 1):

(i) The zone denoted h includes the α-helix conformation. This zone is an ellipse centered at (−80°,−30°), with a major axis a = 90° and a minor axis b = 50°. The major axis is tilted 135° with respect to the ψ axis.

(ii) The zone denoted g embodies the left-handed helix. This zone is an ellipse centered at (80°,20°), with a = 90° and b = 50°. The major axis is tilted 135° with respect to the φ axis.

(iii) The zone denoted e includes the extended region. This zone is defined by $-180° < \varphi < 180°$ and $\psi < -60°$ or $\psi > 60°$ for Gly and by $\varphi < 0°$ and $\psi > 60°$ or $\psi < -140°$, $\varphi > 140°$ and $\psi < 140°$, or $\varphi < -140°$ and $\psi < -140°$ for the rest of the residues.

(iv) The residual regions on the map are denoted x and may be considered a "catch-all" state for unusual, distorted, and sometimes even experimentally erroneous residue conformations.

With these definitions, most of the residues are located in two zones, e and h. One has to notice that, in the area around ψ = 50°, the distinction between the two zones is somewhat arbitrary.

Each residue in the data base was then assigned a zone according to the values of its φ and ψ pair and the above definitions.

(*b*) *Method for Determining the Information Parameters.* The influence exerted by the local amino acid sequence on the conformation of a given residue is conveniently expressed by the formalism of the information theory (Robson & Pain, 1971; Robson, 1974). (By local sequence we mean hereafter the residue in question and the eight preceding N-terminal residues plus the eight following C-terminal residues.) The basic equations calculating the fraction or probability of a residue being in a given zone are the following.

The basic definition we use for the information that the general event Y carries about the occurrence of the general event X follows Fano (1961):

$$\mathbf{I}(X;Y) = \ln\,[\mathbf{P}(X|Y)/\mathbf{P}(X)] \qquad (1)$$

where $\mathbf{P}(X)$ is the probability of event X, and $\mathbf{P}(X|Y)$ is the conditional probability of X, i.e., the probability of event X knowing that event Y has occurred. $\mathbf{I}(X;Y)$ may be defined as the statistical constraint between the two events X and Y. In the present method, we use $\mathbf{I}(X{:}\bar{X};Y) = \mathbf{I}(X;Y) - \mathbf{I}(\bar{X};Y)$, where $\bar{X}$ is the complementary event of X.

In practice we need to evaluate the value of the following information: $\mathbf{I}(Sj=Z{:}\bar{Z};\ R_{j-8},\ldots,R_{j+8})$. This is the influence of the local sequence R_{j-8}, ..., R_{j+8} (even Y) on the conformations Z and $\bar{Z}$ of the residue at position j (events X and $\bar{X}$). In the present case, Z can be one of the four zones in the Ramachandran map: h, e, g, x. Equation 1 becomes

$$\mathbf{I}(S_j=Z{:}\bar{Z};\ R_{j-8},\ldots,R_{j+8}) = \ln\,[\mathbf{P}(S_j=Z,\ R_{j-8},\ldots,R_{j+8})/\mathbf{P}(S_j=\bar{Z},\ R_{j-8},\ldots,R_{j+8})] + \ln\,[\mathbf{P}(S_j=\bar{Z})/\mathbf{P}(S_j=Z)] \qquad (2)$$

Equation 2 can be expressed in term of frequencies as

$$\mathbf{I}(S_j=Z{:}\bar{Z};\ R_{j-8},\ldots,R_{j+8}) = \ln\,[f(S_j=Z,\ R_{j-8},\ldots R_{j+8})/f(S_j=\bar{Z},\ R_{j-8},\ldots,R_{j+8})] + \ln\,[f(S_j=\bar{Z})/f(S_j=Z)] \qquad (3)$$

The ratio $f(S_j=Z)/f(S_j=\bar{Z})$ is a constant for the data base and should be representative of the ratio between state Z and state $\bar{Z}$ in proteins, for a sufficiently large data base. We thus note

$f(S_j=Z)/f(S_j=\bar{Z}) = K_{zz}$ and $\mathbf{I}(S_j=Z{:}\bar{Z};\ R_{j-8},\ldots,R_{j+8}) = I_{zz}$

then

$$f(S_j=Z,R_{j-8},\ldots,R_{j+8})/f(S_j=\bar{Z},\ R_{j-8},\ldots,R_{j+8}) = K_{zz}e^{I_{zz}} \qquad (4)$$

This expression can be considered as a partition coefficient between states Z and $\bar{Z}$ for the residue at j under the influence of the local sequence R_{j-8}, ..., R_{j+8}. We call fz the fraction of residues in state Z when the local sequence is $R_{j-8}, ..., R_{j+8}$. That is

$$\mathrm{fz} = f(S_j{=}Z,\ R_{j-8},...,R_{j+8})/F_{\mathrm{tot}} \tag{5}$$

where

$$F_{\mathrm{tot}} = f(S_j{=}Z,\ R_{j-8},...,R_{j+8}) + f(S_j{=}\bar{Z},\ R_{j-8},...,R_{j+8})$$

Equation 4 then becomes

$$\mathrm{fz}/(1 - \mathrm{fz}) = K_{z\bar{z}}e^{I_{z\bar{z}}} \tag{6}$$

and thus

$$\mathrm{fz} = K_{z\bar{z}}e^{I_{z\bar{z}}}/(1 + K_{z\bar{z}}e^{I_{z\bar{z}}}) \tag{7}$$

We can therefore determine the fraction for the four states h, e, g, and x. This fraction is the probability that, given a local sequence, the central residue will be in one of the specified zones h, e, g, or x.

If the information values are consistent, then the sum of the fractions, fz, for the four zones should add up to 1 for a given residue. If one wants to assign a zone to a residue, the best choice consists of the zone having the highest fraction. For example if the different fractions are 0.55 for h, 0.35 for e, 0.07 for g, and 0.03 for x, then the choice for the predicted zone is h (see an example in Figure 2). Nevertheless we know that this assignment (representing the effect of residues independent of long-range effects, see below) will be right only in 55% of all the observations.

The difficulty arises with the determination of the information $\mathbf{I}(S_j{=}Z{:}\bar{Z};\ R_{j-8},...,R_{j+8})$. Equation 3 requires data about events such as $R_{j-8}, ..., R_{j+8}$ involving 17 residues. There are 20^{17}, roughly 10^{22}, such local sequences. As a consequence the value of the above information has to be approximated. The problem is, how can we deal with events like the ones above? All the prediction methods have to include this kind of approximation (even if the formalism is different).

The different possibilities for approximating eq 3 have been the subject of a previous paper (Gibrat et al., 1987). For the sake of completeness we give here the result [the reader interested can consult Gibrat et al. (1987)] and Appendix 1:

$$\mathbf{I}(S_j{=}Z{:}\bar{Z};\ R_{j-8},...,R_{j+8}) = \mathbf{I}(S_j{=}Z{:}\bar{Z};\ R_j) + \sum_{m=-8}^{+8} \mathbf{I}(S_j{=}Z{:}\bar{Z};\ R_{j+m}|R_j) \tag{8}$$

The constituent terms on the right-hand side represent specific contributions. $\mathbf{I}(S_j{=}Z{:}\bar{Z};\ R_j)$ is called self-information, that is, the information the residue at j bears about its own conformation. $\mathbf{I}(S_j{=}Z{:}\bar{Z};\ R_{j+m}|R_j)$ is called pair information; this is the information borne by a residue at $j + m$ (m ranging from −8 to +8 and $m \neq 0$) about the conformation of the residue at j, knowing that this residue is of a given type. This pair information is conditional information and thus has a definition that is distinct from that for Fano's "self-information" (see Appendix 1).

In practice, there are inadequate data to estimate all interactions and the complex event of eq 3 involving 17 residues has been approximated in eq 8 by events involving at most pairs of residues. This inadequacy of data applies, of course, to all methods, including neural nets.

Results and Discussion

(*a*) *Analysis of the Data Base.* With the definition of the zones given above, we obtain (for the 10937 residues of the data base) 4986 h, 5102 e, 519 g, and 330 x. The first and last residues of each protein are excluded.

```
a    TRYPSIN INHIBITOR

              h    e    g    x     PRED   OBS
   1 R     0.45 0.47 0.05 0.03     Z      Z
   2 P     0.36 0.58 0.00 0.02     e      e
   3 D     0.48 0.41 0.04 0.06     h      h
   4 F     0.35 0.59 0.01 0.01     e      h
   5 C     0.30 0.60 0.01 0.07     e      h
   6 L     0.22 0.66 0.02 0.02     e      h
   7 E     0.10 0.85 0.00 0.05     e      e
   8 P     0.10 0.84 0.00 0.03     e      e
   9 P     0.31 0.43 0.00 0.20     e      e
  10 Y     0.34 0.46 0.04 0.14     e      e
  11 T     0.34 0.51 0.04 0.18     e      h
  12 G     0.03 0.86 0.02 0.06     e      e
  13 P     0.38 0.55 0.00 0.06     e      h
  14 C     0.36 0.55 0.01 0.02     e      e
  15 K     0.39 0.48 0.11 0.01     e      h
  16 A     0.29 0.56 0.33 0.01     e      e
  17 R     0.33 0.59 0.03 0.01     e      e
  18 I     0.21 0.70 0.00 0.01     e      e
  19 I     0.46 0.48 0.00 0.00     e      e
  20 R     0.40 0.57 0.00 0.01     e      e
  21 Y     0.46 0.49 0.01 0.05     e      e
  22 F     0.41 0.48 0.02 0.01     e      e
  23 Y     0.40 0.54 0.01 0.02     e      e
  24 N     0.50 0.38 0.02 0.06     h      e
  25 A     0.52 0.34 0.06 0.00     h      h
  26 K     0.56 0.36 0.04 0.00     h      h
  27 A     0.65 0.26 0.16 0.00     h      h
  28 G     0.20 0.38 0.39 0.02     g      g
  29 L     0.39 0.53 0.01 0.02     e      e
  30 C     0.33 0.60 0.03 0.01     e      e
  31 Q     0.41 0.55 0.00 0.01     e      e
  32 T     0.27 0.67 0.01 0.04     e      e
  33 F     0.31 0.61 0.00 0.04     e      e
  34 V     0.22 0.65 0.10 0.03     e      e
  35 Y     0.31 0.55 0.02 0.08     e      e
  36 G     0.20 0.60 0.05 0.01     e      h
  37 G     0.07 0.64 0.27 0.00     e      g
  38 C     0.32 0.57 0.03 0.02     e      e
  39 R     0.44 0.40 0.75 0.01     g      g
  40 A     0.29 0.61 0.06 0.01     e      e
  41 K     0.38 0.50 0.02 0.03     e      e
  42 R     0.43 0.47 0.01 0.01     e      h
  43 N     0.34 0.35 0.10 0.15     e      e
  44 N     0.30 0.42 0.14 0.06     e      e
  45 F     0.30 0.68 0.00 0.00     e      e
  46 K     0.47 0.45 0.01 0.04     h      h
  47 S     0.34 0.58 0.00 0.02     e      e
  48 A     0.58 0.34 0.00 0.03     h      h
  49 E     0.66 0.25 0.04 0.01     h      h
  50 D     0.53 0.31 0.04 0.01     h      h
  51 C     0.56 0.39 0.02 0.01     h      h
  52 M     0.52 0.41 0.04 0.00     h      h
  53 R     0.45 0.52 0.00 0.02     e      h
  54 T     0.39 0.52 0.02 0.05     e      h
  55 C     0.30 0.58 0.05 0.06     e      h
  56 G     0.25 0.35 0.45 0.08     g      h
  57 G     0.09 0.64 0.28 0.01     e      e
  58 A     0.45 0.47 0.05 0.03     Z      Z
```

FIGURE 2: Computer output of the prediction of bovine trypsin inhibitor for zones h, e, g, and x with the observed conformation added for comparison. Z refers to the first and the last residues, which lack one torsional angle.

It is important to distinguish between the secondary structure of a residue (H, E, or C) and its zone (or state); h, e, g, or x. We adopt for the secondary structure definitions H, α helix; E, β strand; and C aperiodic structure, those given by Kabsch and Sander (1983) based on hydrogen bond patterns after reduction to three states from Levin and Garnier (1988). This allows a homogeneous definition of the secondary structure in the data base. As mentioned previously, zone h embodies the α-helical region and zone e includes the extended region but these two zones *do not* necessarily correspond directly to the regular secondary structures. For example, a series of e's may not correspond to a β strand but may correspond to an aperiodic structure if the φ and ψ angles are near the limits of the zone. However, a region having an aperiodic secondary structure is usually constituted of a mixture of residues located in different zones, h, e, g, and x.

A little more than 3% of the residues belong to zone x. Some of these residues appear near the border of the e or h zones. It is likely that the location of some of them is due to small errors in the coordinates. Errors vary with authors and with the extent to which the different parts of the electron density map are resolved. Uncertainties of ± 20° on dihedral angles are common. Certainly a number of residues are in the region of the Ramachandran map that is not favorable energetically (for example, 74 residues in the quadrant $\varphi > 0°$ and $\psi < 0°$). These residues may belong to the less well-refined proteins in the data base or parts of the protein that are not very well resolved by X-ray diffraction. However, Moult and James (1986) mentioned the case of an asparagine from Streptomyces griseus protease A, SGPA, that lies in a very clear area in the electron density map and has the values $\varphi = 81°$ and $\psi = -75°$. It must be accepted there are real but "exotic" cases.

It was necessary to define zone g because it is by no means an exotic case. More than one-third of the Gly residues and 12% of the Asn residues are observed in this zone. These two residues represent 70% of the residues in zone g. Moreover this zone is characteristic of certain turns according to Rose et al. (1985) (β turns: type I, residues $i + 1$ and $i + 2$; type II, residue $i + 2$; type III′, residues $i + 1$ and $i + 2$), and thus it is interesting to predict this area of the Ramachandran map. Ab initio calculations for Gly dipeptide [see Robson and Garnier (1986)] show the same energy minimum for zones g and h. However, zone g contains twice as many Gly residues as zone h. This underlines the particular role of Gly in the proteins due to the lack of a side chain. With the marked exception of Asn (12%), this zone is rather sparsely populated by the other residues (0% for Trp to 5% for His).

The majority of the residues (92%) are distributed between zone h and zone e (4986 h and 5102 e). We compared the secondary structure and the zone in the Ramachandran map for the residues (data not shown). The residues in the α helix mostly fall in zone h, as was expected; however, a few of them are in zones e or g. These latter residues appear at the end of the helices; they are no longer in zone h but they are able to give appropriate hydrogen bonds with the rest of the helix. The same phenomenon occurs with the residues of the sheet, which are mostly found in zone e except for a few at the ends of the β sheets. Aperiodic structure (C) appears to be a mixture of residues in zones h, e, and g. Zone e is more populated than zone h for residues with an aperiodic structure. For a dipeptide, zone e is energetically more favored than zone h (Zimmerman et al., 1977). However, in the data base there are approximately the same number of residues in the two zones. This can be considered as a consequence of the cooperative effects between residues in the helical conformation.

Jones and Thirup (1986) have shown that it is possible to build the retinol binding protein (RBP) starting from a number of segments of same end-to-end distance belonging to various proteins. For example, they found fragments in 23 different proteins that are similar (with a root-mean-square distance (rms) less than or equal to 0.5 Å) to a turn appearing several times in the RBP. This turn is a pentapeptide with a Gly at position $i + 3$. If one considers that residues other than Gly have mainly two possibilities (zones h or e) and that Gly has three possibilities (h, e, and g), then there are for the pentapeptide $(2^4)3 = 48$ different "conformation classes". It is therefore understandable to find in the data base a number of pentapeptides belonging to the same conformation class and having a small value fo the rms with the test pentapeptide.

It is interesting to see if there are some restrictions for the residues defined as coil, i.e., if some of the above mentioned conformation classes are not found. For that we considered only the residues observed in aperiodic secondary structures (C). A window (varying from four to seven residues) is moved by one residue along the sequence. For example, consider the following segment of protein:

```
A R G V G K L S T P G V   sequence
C C C C C C C C C C C C   observed secondary structure (K &S)
h e e g h h e h e e e e   obs. zones in the Ramachandran map
>-----<
  >-----<
    >-----<
```

In this example, nine tetrapeptides are defined with a window of four residues. We consider only the peptides containing h and e (here five peptides). As a consequence, a class is determined by a sequence of four letters (e or h); i.e., there are 16 possible classes for a tetrapeptide.

It turns out that each class is observed, but some are more represented than others (data not shown). Peptides with a majority of e's are the most abundant. This is particularly the case for those peptides containing only e's, which always constitute the most populated class. This was expected because the e's represent 53% (32% for the h's) of the residues having an aperiodic secondary structure. The less populated classes are those with an alternation of e's and h's. This is due to the fact that the periodic secondary structures exhibit cooperativity arising largely from hydrogen bonding, which does not operate when an alternation of e's and h's occurs.

(*b*) *Analysis of the Predictions.* Table I shows the results of the prediction for the 66 polypeptide chains of the data base. The prediction is done with a number M [added frequencies: see Appendix 1 and Gibrat et al. (1987)] equal to 200. The distribution of predicted residues is 5097 h, 5647 e, 188 g, and only 5 x. Note that the protein to be predicted is each time removed from the data base and the information values used are recalculated. The percentage of correctly predicted residues is thus more truly representative of the expected percentage of an unknown protein. The percentage of correctly predicted residues obtained for the whole data base is 62%. If the protein is not removed from the data base the result for the prediction is 68.4%. Clearly, any algorithm based on the data base including the protein of interest can be misleadingly good. Some proteins are poorly predicted (the percentage is about the same as the one we would obtain with a random prediction: 42.5%), such as 2ABX, 1CRN, 1FDX, and 2SNS (Table I). For some of them this can be related to the large percentage of residues appearing in zone x (which is very much underpredicted here, five residues only): 44% for 2ABX; 15% for 2SNS. It is more difficult to explain the results obtained for 1CRN and 1FDX, which contain less residues in zone x.

There are also less residues that are predicted in zones g and x than are observed (respectively 188/519 and 5/330). It is possible that the local sequence bears no information at all about zone x. So in most cases the fraction of residues in zone x tends to 0 and only a few residues are predicted in this zone. Nevertheless there is also a technical difficulty. The residues in zones g and x are the less abundant and therefore the frequencies used for the calculation of the information parameters are small. In such a case the probabilities estimated from these frequencies are often wrong. This fact causes the accuracy of the prediction to decrease appreciably. In order to avoid this as much as possible we used the hash function of Robson (1974). It can be formally proven that this hash function weights out automatically the smaller frequencies in the calculation of the information values; i.e., rather than including contributions that are likely wrong, the method prefers to damp down these contributions. A side effect of this technique is that many contributions to the zones g and x vanish and the local sequence appears to bear less information about these two zones.

Table I: Results of the Prediction[a]

protein	Brookhaven file name	% of residue corr pred
acid proteinase (*Penicilium* J)	2APP	64.2
actinidin	2ACT	56.9
agglutinin (wheat germ)	3WGA	63.1
alcohol dehydrogenase (apo)	4ADH	48.1
α-lytic protease	2ALP	68.4
aspartate carbamoyltransferase (chain 1)	4ATC	52.9
aspartate carbamoyltransferase (chain 2)	4ATC	52.3
azurin (*Alcaligenes denitrificans*)	1AZA	53.5
α-bungarotoxin	2ABX	37.5
Ca-binding parvalbumin	3CPV	70.8
Ca-binding protein (intestinal)	3ICB	86.3
carbonic anhydrase B (human)	2CAB	62.6
carboxypeptidase A	5CPA	61.3
catalase (beef liver)	8CAT	64.7
α-chymotrypsin A (chain 1)	5CHA	53.5
α-chymotrypsin A (chain 2)	5CHA	67.4
citrate synthase (pig)	2CTS	67.1
crambin	1CRN	38.6
γ-II crystallin (calf)	1GCR	67.4
cytochrome *c* (rice)	1CCR	69.7
cytochrome *c* (prime)	2CCY	83.2
cytochrome *c* peroxidase (yeast)	2CYP	59.5
cytochrome c_2 (reduced)	3C2C	69.1
cytochrome c_3 (*Desulfovibrio vulgaris*)	2CDV	57.1
cytochrome *c*-551 (oxidized)	351C	72.5
dihydrofolate reductase (*Lactobacillus casei*)	3DFR	51.9
elastase (procine)	2EST	64.3
erabutoxin B (sea snake)	2EBX	78.3
erythrocruonin (reduced deoxy)	1ECD	71.6
ferredoxin (*Peptococcus aerogenes*)	1FDX	42.3
ferredoxin (*Spirulina platensis*)	3FXC	49.0
flavodoxin (*Clostridium* MP, oxidized)	3FXN	69.9
glutathione peroxidase (bovine)	1GP1	58.2
hemerythrin (met)	1HMQ	78.4
hemoglobin (human, deoxy; chain 1)	2HHB	82.0
hemoglobin (human, deoxy; chain 2)	2HHB	78.5
hemoglobin V (cyano, met, lamprey)	2LHB	73.5
high potential iron protein	1HIP	51.8
IGG FAB (κ) MCPC603 light chain	1MCP	62.8
IGG FAB (κ) MCPC603 heavy chain	1MCP	65.5
kallikrein (porcine; chain 1)	2PKA	53.8
kallikrein (porcine; chain 2)	2PKA	62.0
lactate dehydrogenase	4LDH	59.6
leghemoglobin	1LH1	80.1
lysozyme (bacteriophage T4)	2LZM	69.1
lysozyme (human)	1LZ1	53.1
melittin	1MLT	66.7
myoglobin (sperm whale, met)	1MBN	83.4
scorpion neurotoxin (variant)	1SN3	65.1
ovomucoid third domain (quail)	1OVO	61.1
papain D	1PPD	59.0
phospholipase A_2 (bovine)	1BP2	55.4
plastocyanin	1PCY	56.7
prealbumin (human plasma)	2PAB	55.4
proteinase A (*Streptomyces griseus*)	2SGA	65.4
proteinase II (rat mast cell)	3RP2	56.8
ribonuclease A	1RN3	59.8
rubredoxin	5RXN	55.8
staphylococcal nuclease	2SNS	41.7
subtilisin BPN prime	1SBT	59.0
superoxide dismutase	2SOD	59.7
thermolysin	3TLN	59.6
trypsin (orthorhombic)	1TPO	61.1
trypsin inhibitor (bovine)	4PTI	62.5
virus (satellite tobacco necrosis)	2STV	61.5
virus coat protein (southern bean mosaic)	4SBV	55.0
results for the whole data base		62.0

[a] The results above are obtained by removing each time the protein to be predicted from the data base used to calculate the information parameters as in Gibrat et al. (1987).

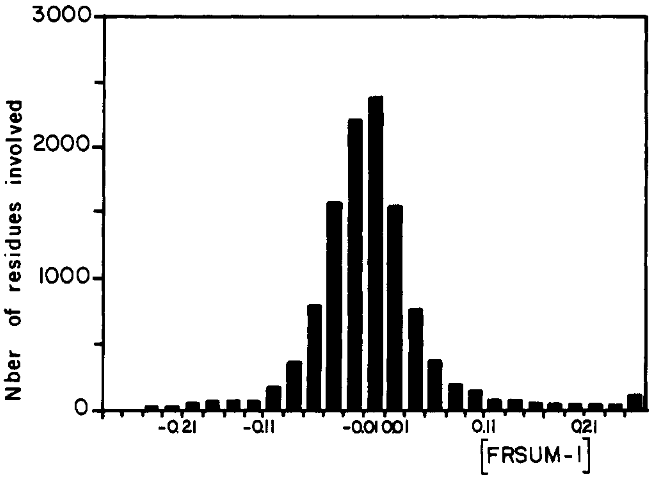

FIGURE 3: Deviation from 1 of the sum of the residue fractions. This histogram shows the partition of the residues as a function of the difference FRSUM-1. FRSUM is the sum of the residue fractions for the four states computed from the information values. The distribution is not symmetrical. There are 5392 residues for which FRSUM is less than 1 - 0.01 and 3404 residues for which FRSUM is greater than 1 + 0.01. There are 2388 residues for which the difference FRSUM-1 lies in the interval]-0.01, 0.01[. See text for the explanation of the distribution asymmetry.

It is interesting to see how the prediction of the Ramachandran zones is related to the secondary structure. According to the definition of the secondary structures given by Kabsch and Sander (1983) and reduced to three states, there are 3226 H, 2329 E, and 5382 C in the data base. The percentage of residues that are correctly predicted in term of Ramachandran zones for the three above secondary structures, %CPs, is given by the following expression, %CPs = Nr/Ns where Ns represents the number of residues that have the secondary structure S [Kabsch and Sander (1983) definition] and Nr represents the residues having the same secondary structure S that are also correctly predicted by the Ramachandran zone prediction method. For example, there are 5382 aperiodic residues, 2811 of which are correctly predicted in terms of Ramachandran zones: h, e, or g, that is, 52%. With this definition, we found that 73% of the residues defined as H, 70% of the residues defined as E, and only 52% of the residues defined as C are correctly assigned by the prediction of zones. The local sequence thus bears more constraints on the α helix and β sheet conformations than on the aperiodic structure.

(*c*) *Tests for the Approximations Made.* As stated under Method, if the calculation of the information values is consistent, then the sum of the derived fractions for the four zones for each residue must be equal to 1. Figure 3 presents a histogram that shows the difference between the sum of the calculated fractions and the theoretical value of 1. The distribution is not symmetrical, there are more residues for which the sum of the fractions is less than 1 - 0.01 (5392 residues) than residues for which the sum of the fractions is greater than 1 + 0.01 (3404 residues). This is a consequence of the problem mentioned in paragraph b with the zones x and g; i.e., some information for these zones has been lost. However, there are only 849 residues in zones g and x, which is not enough to explain the asymmetry of the distribution. Similarly, some smaller amount of information about zones e and h is also lost during the calculation. Nevertheless it is pleasing that there are 9126 residues (83%) for which the sum of the fractions is equal to 1 ± 0.05, indicating that the computation of the information values is self-consistent with only small discrepancies, which can reasonably be attributed to experimental uncertainties.

Equation 4 gives the "partition coefficient" between zones Z and $\bar{Z}$ for the residue at position *j* under the influence of

Table II: Correctly Predicted Residues as a Function of the Residue Fraction[a]

<table>
<tr><th></th><th colspan="18">fraction interval</th></tr>
<tr><th></th><th colspan="2">0.2</th><th colspan="2">0.3</th><th colspan="2">0.4</th><th colspan="2">0.5</th><th colspan="2">0.6</th><th colspan="2">0.7</th><th colspan="2">0.8</th><th colspan="2">0.9</th><th colspan="2">1.0</th></tr>
<tr><td>% of h corr pred</td><td></td><td colspan="2">b</td><td colspan="2">36</td><td colspan="2">51</td><td colspan="2">56</td><td colspan="2">67</td><td colspan="2">73</td><td colspan="2">80</td><td colspan="2">93</td><td></td></tr>
<tr><td>% of e corr pred</td><td></td><td colspan="2">b</td><td colspan="2">41</td><td colspan="2">43</td><td colspan="2">55</td><td colspan="2">63</td><td colspan="2">74</td><td colspan="2">84</td><td colspan="2">90</td><td></td></tr>
<tr><td>% of g corr pred</td><td></td><td colspan="2">b</td><td colspan="2">32</td><td colspan="2">35</td><td colspan="2">53</td><td colspan="2">54</td><td colspan="2">61</td><td colspan="2">b</td><td colspan="2">c</td><td></td></tr>
<tr><td>% of res corr pred</td><td></td><td colspan="2">b</td><td colspan="2">38</td><td colspan="2">48</td><td colspan="2">55</td><td colspan="2">64</td><td colspan="2">73</td><td colspan="2">82</td><td colspan="2">91</td><td></td></tr>
<tr><td>no. of res involved</td><td></td><td colspan="2">6</td><td colspan="2">194</td><td colspan="2">1570</td><td colspan="2">3628</td><td colspan="2">2804</td><td colspan="2">1854</td><td colspan="2">855</td><td colspan="2">130</td><td></td></tr>
</table>

[a] Let us consider a residue with the following fraction: 0.54 for h, 0.27 for e, 0.17 for g, and 0.02 for x. This residue is predicted h and lies in the fraction interval 0.5–0.6 (column 4 in the table). There is, theoretically, a 54% chance to predict correctly this residue in zone h. The figures in the rows give the percentage of residues that being predicted in a zone with a given fraction are effectively observed in this zone in the data base. [b] The number of residues involved is too small for the percentage to be meaningful. [c] No residues are predicted for this state with this fraction.

the local sequence R_{j-8}, ..., R_{j+8}. For example, if $\mathbf{I}(S_j=h:\bar{h}; R_{j-8}, ..., R_{j+8}) = 200$ cnats ($K_{h\bar{h}} = 0.838$ for the data base), the partition coefficient is equal to 6.1, and thus the fraction of residues at j that are in zone h is 86%. If one predicts the residue to be in state h, there should be, as a consequence, a 14% chance of being wrong. One has to keep in mind that the information value expresses the balance between a zone and its complementary zone and not an absolute preference for the zone in question. For example, the information value for h should not simply be the highest but should be as high as 450 cnats for having a 99% chance for a residue in state h.

An important aspect of this study is that the fractions of residues in a given state allow us to estimate if the information values are correctly calculated, that is, if the method is able to extract all the information contained in the local amino acid sequence. For each residue in the data base, we computed the information values and the fractions for each state. Then we compared them with the observed data to check if a similar proportion of correctly predicted residues was observed. For example, let us consider all the residues for which the highest fraction is for state h and lies between 0.5 and 0.6. These residues are predicted h and there is on average 55% chance that they are effectively observed in this zone. Such figures are shown in Table II. This table shows that the fractions of correctly predicted residues agree reasonably well with the calculated fractions. Zone g shows a tendency to be less well predicted than the calculated fractions; this is likely related to the problem mentioned in section b.

These two results (namely, the fact that, for most of the residues, the fractions for the four states add up to 1 and, more important, that if the residues are predicted with a given fraction in a given state they are also observed in this state with the same relative frequency in the data base) show that basically the fractions and information values are calculated correctly (see also Appendix 2). This is pleasing in view of the fundamental assumptions and approximations made. Therefore the method is able to extract all the information of the local sequence and a figure of 65% represents the average influence of the short-range interactions on the conformation adopted by residues in the protein. We can fix the upper limit of the average of the short-range interaction influence to the upper limit found by other methods (Biou et al., 1988; Qian & Sejnowski, 1988) of 65% instead of the figure of 62% we found.

This helps tackle the question, is it possible to improve the accuracy of the prediction? Since the method is able to extract the information contained in the local sequence to reproduce correctly the actual distribution of residues in the data base, a limited extension of the local sequence information should not improve the accuracy. This is effectively observed. The use of greater windows (±10 and ±12) does not increase the accuracy (data not shown). This has been confirmed by other methods based on different algorithms (Levin & Garnier, 1988; Holley & Karplus, 1989). In theory long-range interactions might help. In general practice, however, it appears that residues outside the window bring no more information but a slight "noise", which perturbs the prediction accuracy. With a smaller data base of 26 proteins, Robson and Suzuki (1976) early observed that the directional information tends to zero for residues more than 8 residues from the central residue. If we can extrapolate from this extension of data bases from 26 to 61 proteins, one may expect that further extension of that order in the size of the data base will not bring significant improvements in prediction. This is not to deny that identification of a specific domain folding motif by homology can be beneficial (Levin & Garnier, 1988). In such a case, however, the secondary structure prediction is somewhat superfluous (except sequence alignments) and certainly less central in role.

Conclusion

Starting with the observation that there are, roughly speaking, only two conformations (two zones in the Ramachandran diagram) allowed for the branched residues, we developed a set of parameters expressing the influence of the local amino acid sequence on these zones. From the information values, we determined the fractions for each state. With these fractions we showed that the approximations made allow us to extract all the information from the local sequence. This leads us to propose that the local sequence (short-range interactions) contributes *on average* up to 65% of the conformation of the residues in proteins. The rest of the contributions come from long-range interactions along the sequence, i.e., between residues brought close by the folding of the polypeptide chain. These types of interactions are too specific to be taken into account by the information theory, at least for any reasonably sized data base. Other methods based on peptide similarity (Levin & Garnier, 1988) can introduce these effects for homologous proteins and raised the accuracy of prediction for those proteins to about 90%.

Table II shows that 10% of the residues have their conformation almost exclusively determined by the local sequence (the fraction is greater than 0.80). It is tempting and by no means heretical to speculate that these residues act as seeds for the nucleation sites during the folding. In this respect, it is interesting to notice that the regular secondary structures (α helix and β sheet) are better predicted in term of Ramachandran zones than the aperiodic structure. This is consistent with the picture of these regular secondary structures acting as a frame for the protein three-dimensional structure, whereas the aperiodic structure is mainly found in the loops connecting these regular structures.

The set of parameters developed here can be used, mainly, for two purposes. The first one is the determination of peptides that are intended to be used as synthetic vaccines. The choice of such peptides is a complicated task. There are various aspects to consider. One of them is the following. Once a peptide, i.e., a sequential part of the target protein, has been chosen, what is the likelihood that this isolated peptide in solution will retain the three-dimensional structure it has in the protein? If the conformation of most of the residues in this peptide are not influenced by the local sequence, i.e., if there is no dominant fraction for a given state, then it is very unlikely that the peptide will have the same three-dimensional structure in solution and in the protein. The parameters allow one to screen the peptides that are better candidates as synthetic vaccines from this point of view and synthesize only the ones (a) that are likely to be epitopic and (b) for which the conformation of a majority of residues is strongly influenced by the local amino acid sequence.

Another use of the set of parameters is for piloting the conformational space search in simulations of the three-dimensional structure. Robson and Pain (1973) tried this with energy minimization by SIMPLEX and COMPLEX optimization methods. More recently, Gibrat and Garnier have used a "simulated annealing" type algorithm. The generation of a new conformation must be consistent with the fractions determined from the information values. Therefore the residues for which the fraction for a given zone is as high as 0.8, say, will act as nucleation sites (because they are generated 80% of the time in the same zone). Those for which there is no preferred zone will allow more flexibility in the folding process. Although the conformational space that remains to be explored is still very large, representation of this as an algorithm allows us to disregard large portions of it that are not probable.

The set of parameters as well as the program are available on request.

ACKNOWLEDGMENTS

We thank J. M. Levin for his critical reading of the manuscript.

APPENDIX 1

This appendix reproduces the information theory argument for the sake of completeness. Also it aims at explaining how the approximations involved in the calculation of the information values arise and what exactly their meaning is (see eq A15 below). Furthermore, the second part represents an explanation of the dummy frequencies.

We need to calculate the following expression:

$$\mathbf{I}(S_j{=}Z{:}\bar{Z};\ R_{j-8},\dots,R_{j+8}) = \ln\,[\mathbf{P}(S_j{=}Z,\ R_{j-8},\dots,R_{j+8})/\mathbf{P}(S_j{=}\bar{Z},\ R_{j-8},\dots,R_{j+8})] + \ln\,[\mathbf{P}(S_j{=}\bar{Z})/\mathbf{P}(S_j{=}Z)] \tag{A1}$$

Equation A1 can be expressed in term of frequencies as

$$\mathbf{I}(S_j{=}Z{:}\bar{Z};\ R_{j-8},\dots,R_{j+8}) = \ln\,[f(S_j{=}Z,\ R_{j-8},\dots,R_{j+8})/f(S_j{=}\bar{Z},\ R_{j-8},\dots,R_{j+8})] + \ln\,[f(S_j{=}\bar{Z})/f(S_j{=}Z)] \tag{A2}$$

It is clear that terms like $f(S_j{=}Z,\ R_{j-8},\dots,R_{j+8})$ involving 17 residues and a state are impossible to determine directly from the data base. The data base we have contains about 11 000 residues and therefore it allows us to use only frequencies such as $f(S_j{=}Z,\ R_j,R_k)$ involving two residues and a state. There are 20 × 20 × 4 such terms and thus the average frequency is about 8. The correct expression for the information (A1) has to be approximated. Let us consider a simpler case with only three residues. We have to determine the following information value: $\mathbf{I}(S_j{=}Z{:}\bar{Z};\ R_{j-1}R_jR_{j+1})$. The event $R_{j-1}R_jR_{j+1}$ is the intersection of three more general events, i.e., $[R_{j-1}\bigcirc\bigcirc] \cap [\bigcirc R_j \bigcirc] \cap [\bigcirc\bigcirc R_{j+1}]$, where ○ means the occurrence of a residue whatever its type. Obviously, from this definition the type and the relative position of the residues are important. In the following we will refer, for the sake of simplicity, to an event such as $[R_{j-1}\bigcirc\bigcirc]$ with the short notation R_{j-1}.

The information conveyed by the event $R_{j-1}R_jR_{j+1}$ about the occurrence of the event $S_j = Z$ is given by

$$\mathbf{I}(S_j{=}Z;\ R_{j-1}R_jR_{j+1}) = \ln\,[\mathbf{P}(S_j{=}Z|R_{j-1}R_jR_{j+1})/\mathbf{P}(S_j{=}Z)] \tag{A3}$$

We can expand (A3) as

$$\mathbf{I}(S_j{=}Z;\ R_{j-1}R_jR_{j+1}) = \ln\,[\mathbf{P}(S_j{=}Z|R_{j-1}R_jR_{j+1})/\mathbf{P}(S_j{=}Z|R_jR_{j+1})] + \ln\,[\mathbf{P}(S_j{=}Z|R_jR_{j+1})/\mathbf{P}(S_j{=}Z|R_j)] + \ln\,[\mathbf{P}(S_j{=}Z|R_j)/\mathbf{P}(S_j{=}Z)] \tag{A4}$$

Each term of eq A4 can be renamed as

$$\mathbf{I}(S_j{=}Z;\ R_{j-1}R_jR_{j+1}) = \mathbf{I}(S_j{=}Z;\ R_{j-1}|R_jR_{j+1}) + \mathbf{I}(S_j{=}Z;\ R_{j+1}|R_j) + \mathbf{I}(S_j{=}Z;\ R_j) \tag{A5}$$

The last term is simply the information borne by event R_j about the event $S_j = Z$. The two first terms are called conditional information. These conditional informations are constituted by the logarithm of the ratio of two conditional probabilities and therefore it does not have strictly speaking the form of information as previously described. If we develop these two terms, we have

$$\mathbf{I}(S_j{=}Z;\ R_{j+1}|R_j) = \ln\,[\mathbf{P}(S_j{=}Z,\ R_jR_{j+1})/\mathbf{P}(R_jR_{j+1})] + \ln\,[\mathbf{P}(R_j)/\mathbf{P}(S_j{=}Z,\ R_j)] \tag{A6}$$

$$\mathbf{I}(S_j{=}Z;\ R_{j-1}|R_jR_{j+1}) = \ln\,[\mathbf{P}(S_j{=}Z,\ R_{j-1}R_jR_{j+1})/\mathbf{P}(R_{j-1}R_jR_{j+1})] + \ln\,[\mathbf{P}(R_jR_{j+1})/\mathbf{P}(S_j{=}Z,\ R_jR_{j+1})] \tag{A7}$$

If we express now the difference of the information bearded by the local sequence about state Z and state $\bar{Z}$ (complementary state of Z), we have

$$\mathbf{I}(S_j{=}Z{:}\bar{Z};\ R_{j+1}|R_j) = \ln\,[\mathbf{P}(S_j{=}Z,\ R_jR_{j+1})/\mathbf{P}(S_j{=}\bar{Z},\ R_jR_{j+1})] + \ln\,[\mathbf{P}(S_j{=}\bar{Z},\ R_j)/\mathbf{P}(S_j{=}Z,\ R_j)] \tag{A8}$$

$$\mathbf{I}(S_j{=}Z{:}\bar{Z};\ R_{j-1}|R_jR_{j+1}) = \ln\,[\mathbf{P}(S_j{=}Z,\ R_{j-1}R_jR_{j+1})/\mathbf{P}(S_j{=}\bar{Z},\ R_{j-1}R_jR_{j+1})] + \ln\,[\mathbf{P}(S_j{=}\bar{Z},\ R_jR_{j+1})/\mathbf{P}(S_j{=}Z,\ R_jR_{j+1})] \tag{A9}$$

In eq A8 we use at most frequencies involving two residues and a state. The data base can provide such frequencies. On the contrary we cannot have access directly to the frequencies needed for calculating eq A9. We need a way to approximate this conditional information. For that we rewrite eq A9 using conditional probabilities as

$$\mathbf{I}(S_j{=}Z{:}\bar{Z};\ R_{j-1}|R_jR_{j+1}) = \ln\,[\mathbf{P}(S_j{=}Z,\ R_{j-1}R_j|R_{j+1})/\mathbf{P}(S_j{=}\bar{Z},\ R_{j-1}R_j|R_{j+1})] + \ln\,[\mathbf{P}(S_j{=}\bar{Z},\ R_j|R_{j+1})/\mathbf{P}(S_j{=}Z,\ R_j|R_{j+1})] \tag{A10}$$

$$\mathbf{I}(S_j{=}Z{:}\bar{Z};\ R_{j-1}|R_jR_{j+1}) = \ln\,[\mathbf{P}(S_j{=}Z,\ R_{j-1}R_j|R_{j+1})/\mathbf{P}(S_j{=}Z,\ R_j|R_{j+1})] + \ln\,[\mathbf{P}(S_j{=}\bar{Z},\ R_j|R_{j+1})/\mathbf{P}(S_j{=}\bar{Z},\ R_{j-1}R_j|R_{j+1})] \tag{A11}$$

We not consider the following approximations

$$\mathbf{P}(S_j{=}Z,\ R_{j-1}R_j|R_{j+1})/\mathbf{P}(S_j{=}Z,R_j|R_{j+1}) \simeq \mathbf{P}(S_j{=}Z,\ R_{j-1}R_j)/\mathbf{P}(S_j{=}Z,\ R_j) \tag{A12}$$

$$\mathbf{P}(S_j{=}\bar{Z},\ R_j|R_{j+1})/\mathbf{P}(S_j{=}\bar{Z},R_{j-1}R_j|R_{j+1}) \simeq \mathbf{P}(S_j{=}\bar{Z},\ R_j)/\mathbf{P}(S_j{=}\bar{Z},\ R_{j-1}R_j) \tag{A13}$$

Although $\mathbf{P}(S_j{=}Z,\ R_{j-1}R_j|R_{j+1}) \neq \mathbf{P}(S_j{=}Z,\ R_{j-1}R_j)$ and $\mathbf{P}(S_j{=}Z,\ R_j|R_{j+1}) \neq \mathbf{P}(S_j{=}Z,\ R_j)$, we assume that their ratios are close.

If we define the event $[S_j{=}Z, R_j] = W$, the first term of expression A12 becomes

$$\mathbf{P}(WR_{j-1}|R_{j+1})/\mathbf{P}(W|R_{j+1}) = \mathbf{P}(WR_{j-1}R_{j+1})/\mathbf{P}(WR_{j+1}) = \mathbf{P}(R_{j-1}R_{j+1}|W)/\mathbf{P}(R_{j+1}|W) \qquad \text{(A14)}$$

With the same notation the second term of eq A12 is

$$\mathbf{P}(S_j{=}Z, R_{j-1}R_j)/\mathbf{P}(S_j{=}Z, R_j) = \mathbf{P}(WR_{j-1})/\mathbf{P}(W) = \mathbf{P}(R_{j-1}|W) \qquad \text{(A15)}$$

If the two conditional events $R_{j-1}|W$ and $R_{j+1}|W$ are independent, the two expressions (A14) and (A15) are equal and therefore the approximation that has been done in eq A12 consists of assuming that the conditional events $R_{j-1}|W$ and $R_{j+1}|W$ are independent. The same holds for eq A13. In other words, the influence of the residue R_{j-1} on the event $S_j = Z$ depends on the type of residue at j but is independent on the type of the residue R_{j+1}.

Then eq A10 becomes

$$\mathbf{I}(S_j{=}Z{:}\bar{Z}; R_{j-1}|R_jR_{j+1}) \simeq \ln [\mathbf{P}(S_j{=}Z, R_{j-1}R_j)/\mathbf{P}(S_j{=}\bar{Z}, R_{j-1}R_j)] + \ln [\mathbf{P}(S_j{=}\bar{Z}, R_j)/\mathbf{P}(S_j{=}Z, R_j)] \qquad \text{(A16)}$$

$$\mathbf{I}(S_j{=}Z{:}\bar{Z}; R_{j-1}|R_jR_{j+1}) \simeq \mathbf{I}(S_j{=}Z{:}\bar{Z}; R_{j-1}|R_j) \qquad \text{(A17)}$$

Equation A17 is calculated by replacing triplets by pairs with respect to event R_j. Because our goal is to determine the influence of the local sequence upon the state of the residue at j, it is natural to consider that event R_j is peculiar with respect to the other events (the occurrence of a given residue at another position).

By extension to a local sequence of eight amino acid residues on each side of R_j, the exact expression is

$$\mathbf{I}(S_j{=}Z{:}\bar{Z}; R_{j-8},...,R_{j+8}) = \mathbf{I}(S_j{=}Z{:}\bar{Z}; R_j) + \mathbf{I}(S_j{=}Z{:}\bar{Z}; R_{j+1}|R_j) + \mathbf{I}(S_j{=}Z{:}\bar{Z}; R_{j-1}|R_jR_{j+1}) + ... + \mathbf{I}(S_j{=}Z{:}\bar{Z}; R_{j-8}|R_{j-7},...,R_j,...,R_{j+8}) \qquad \text{(A18)}$$

Each conditional information term gives the contribution of a peculiar event (the occurrence of a given residue type at a given position in the local sequence) knowing that a number (increasing) of events have already occurred. By application of the approximations of eqs A12 and A13 to all the conditional information terms involving more than two residues the following expression is obtained:

$$\mathbf{I}(S_j{=}Z{:}\bar{Z}; R_{j-8},...,R_{j+8}) = \mathbf{I}(S_j{=}Z{:}\bar{Z}; R_j) + \sum_{m=-8}^{+8} \mathbf{I}(S_j{=}Z{:}\bar{Z}; R_{j+m}|R_j) \qquad \text{(A19)}$$

Dummy Frequencies. As mentioned previously the average number of observations for frequencies involving two residues and a state is about eight. Unfortunately, for zones like g or x the number of observations can be very small: one or two or even for some pairs there is no observation at all. If there is no observation, we obtain conditional information values that are infinite. If the number of observation is too small, then the relative frequencies can be very different from the true probabilities. For example, we can observe in the data base a ratio of relative frequencies of 0.5, whereas the true ratio should be 0.33. This is statistical variation. The difference between this two ratios in term of the information value is 40 cnats, which can be important for a single interaction. This kind of problem is the cause of errors in the calculation of information values and consequently the result of the prediction decreases appreciably. In order to palliate this problem we use two different techniques (Gibrat et al., 1987). First we use the hash function of Robson (1974), which weights out automatically the smaller frequencies in the calculation of the formation values; i.e., rather than including contributions that are likely wrong, the method prefers to ignore these contributions. A side effect of this technique is that many contributions to the zones g and x vanish and the local sequence appears to bear less information about these two zones. Another technique consists in the addition of "dummy observations", f_d. Let us consider the following information value:

$$\mathbf{I}(S_j{=}Z{:}\bar{Z}; R_kR_j) = \mathbf{I}(S_j{=}Z{:}\bar{Z}; R_k|R_j) + \mathbf{I}(S_j{=}Z{:}\bar{Z}; R_j) \qquad \text{(A20)}$$

An approximation similar to eq A17 can be made, replacing conditional pair information by information involving only one residue (directional information):[1]

$$\mathbf{I}(S_j{=}Z{:}\bar{Z}; R_kR_j) \simeq \mathbf{I}(S_j{=}Z{:}\bar{Z}; R_k) + \mathbf{I}(S_j{=}Z{:}\bar{Z}; R_j) \qquad \text{(A21)}$$

The development of (A21) in term of frequencies gives

$$\ln [f_\mathrm{d}(S_j{=}Z; R_kR_j)/f_\mathrm{d}(S_j{=}\bar{Z}; R_kR_j)] \simeq \ln [f(S_j{=}Z; R_j)/f(S_j{=}\bar{Z}; R_j)] + \ln [f(S_j{=}Z; R_k)/f(S_j{=}\bar{Z}; R_k)] + \ln [f(S_j{=}\bar{Z})/f(S_j{=}Z)] \qquad \text{(A22)}$$

We therefore obtain an approximation of the ratio of pair frequencies: $f_\mathrm{d}(S_j{=}Z; R_kR_j)/f_\mathrm{d}(S_j{=}\bar{Z}; R_kR_j)$ by using frequencies involving only one residue.

If we impose $f_\mathrm{d}(S_j{=}Z; R_kR_j) + f_\mathrm{d}(S_j{=}\bar{Z}; R_kR_j) = M$, we can calculate each term of this ratio from eq A22 with the observed single frequencies. These dummy observations are added to the real observations, f_o, in the data base to give the total f_t frequencies and

$$f_\mathrm{t}(S_j{=}Z; R_kR_j)/f_\mathrm{t}(S_j{=}\bar{Z}; R_kR_j) = [f_\mathrm{o}(S_j{=}Z; R_kR_j) + f_\mathrm{d}(S_j{=}Z; R_kR_j)]/[f_\mathrm{o}(S_j{=}\bar{Z}; R_kR_j) + f_\mathrm{d}(S_j{=}\bar{Z}; R_kR_j)] \qquad \text{(A23)}$$

Now instead of the observed frequencies, we use in the calculations the frequencies given by expression A23. The idea behind this is that the ratio of dummy observations gives a reasonable approximation of the true ratio. If there are many observed frequencies, the result of the correction with the dummy observations is only small. On the other hand if the observed frequencies are small (0, 1, or 2 observations), then the ratio gives a very bad approximation of the true ratio, and therefore we obtain a better approximation if we add dummy observations. The global result should be an improvement of the information values measured, for example, by the number of residues correctly predicted by the method. The number M of dummy observations is thus determined experimentally in order to optimize the accuracy of the prediction.

Appendix 2

At a request from a referee to be more explicit about the relative influence of short- and long-range interactions, let us define the short-range interactions on a residue as **I**(sr) and the long-range interactions as **I**(lr|sr). For a given residue in a given protein, the conformation of that residue will be determined by the sum **I**(sr) + **I**(lr|sr). For a given conformational state, let us consider all the residues in the data base submitted to the same short-range interactions **Ia**(sr) *irrespective* of the proteins that they belong to. Suppose that for all these residues the *average* of the long-range interactions, **Ia**(lr|sr), is zero. Then the fraction (probability) of these residues in that conformational state calculated from **Ia**(sr)

[1] In Gibrat et al. (1987), we mentioned that $\mathbf{I}(S_j{=}Z;\bar{Z}; R_kR_j) = \mathbf{I}(S_j{=}Z;\bar{Z}; R_k) + \mathbf{I}(S_j{=}Z;\bar{Z}; R_j)$ if the two events R_k and R_j are independent. This is a mistake since as we ourselves have noted the correct expansion involves conditional information, $\mathbf{I}(S_j{=}Z;\bar{Z}; R_k|R_j)$, which is, in theory, different from the normal information term. We can only use the approximation given by expression A21.

should be, if **Ia**(sr) is correctly calculated, the fraction (probability) observed for these residues in the data base *irrespective* of the protein. And so for all calculated fractions and the four predicted conformations, this is effectively observed. Besides an unlikely compensation between **Ia**(sr) and **Ia**(lr|sr) for all the calculated fractions, **Ia**(lr|sr) being zero implies that there is no correlation between the long-range interactions, acting as "noise", and the residues having the same value of **Ia**(sr) in the data base. This should be evident for unrelated proteins, but the data base contains a certain number of homologous proteins that could bring such a correlation. However, in previous simulation, Gibrat et al. (1987) found that the *average* accuracy of the prediction was not significantly modified by the presence of homologous proteins in the data base. This is probably because they are small in number and percentage of identity, and they are overweighted by nonhomologous proteins in the data base. On the other hand, further extension of the short-range interactions over ± eight amino acids does not improve the prediction, suggesting that the long-range interactions really act as noise. Then the average accuracy of prediction, 65%, reflects correctly only the short-range interactions.

Registry No. Bovine trypsin inhibitor, 12407-79-3.

References

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., & Tasumi, M. (1977) *J. Mol. Biol. 112*, 535–542.

Biou, V., Gibrat, J.-F., Levin, J. M., Robson, B., & Garnier, J. (1988) *Protein Eng. 2*, 185–191.

Emr, S. D., & Silhavy, T. J. (1983) *Proc. Natl. Acad. Sci. U.S.A. 80*, 4599–4603.

Fano, R. (1961) *Transmission of Information*, Wiley, New York.

Fasman, G. D. (1989) in *Prediction of Protein Structure and the Principles of Protein Conformation* (Fasman, G. D., Ed.) Chapter 6, pp 193–301, Plenum Press, New York.

Garnier, J., Gaye, P., Mercier, J.-C., & Robson, B. (1980) *Biochimie 62*, 231–239.

Garnier, J., & Levin, J. M. (1990) *CABIOS* (in press).

Gibrat, J.-F., Garnier, J., & Robson, B. (1987) *J. Mol. Biol. 198*, 425–443.

Holley, L. H., & Karplus, M. (1989) *Proc. Natl. Acad. Sci. U.S.A. 86*, 152–156.

Jones, T. A., & Thirup, S. (1986) *EMBO J. 5*, 819–822.

Kabsch, W., & Sander, C. (1983) *Biopolymers 22*, 2577–2637.

Levin, J. M., & Garnier, J. (1988) *Biochim. Biophys. Acta 955*, 283–295.

Moult, J., & James, M. N. G. (1986) *Proteins 1*, 146–163.

Pfaff, E., Mussgay, M. N., Böhm, H. O., Schulz, G. E., & Schaller, H. (1982) *EMBO J. 1*, 869–874.

Qian, N., & Sejnowski, T. J. (1988) *J. Mol. Biol. 202*, 865–884.

Robson, B., & Pain, R. H. (1971) *J. Mol. Biol. 58*, 237–259.

Robson, B., & Pain, R. H. (1973) in *The Fifth Jerusalem Symposium on Quantum Chemistry and Biochemistry* (Pullman, & Pullman, Eds.) Academic Press, New York.

Robson, B. (1974) *Biochem. J. 141*, 853–857.

Robson, B., & Suzuki, E. (1976) *J. Mol. Biol. 107*, 327–356.

Robson, B., Platt, E., Finn, P. W., Millard, P., Gibrat, J. F., & Garnier, J. (1985) *Int. J. Peptide Protein Res. 25*, 1–8.

Robson, B., & Garnier J. (1986) *Introduction to Proteins and Protein Engineering*, Elsevier, Amsterdam.

Rose, G. D., Gierasch, L. M., & Smith, J. A. (1985) *Adv. Protein Chem. 37*, 1–109.

Taylor, W. R., & Thornton, J. M. (1983) *Nature 305*, 540–542.

Zimmermann, S. S., Pottle, M. S., Nemethy, G., & Scheraga, H. A. (1977) *Macromolecules 10*, 1–9.

GroE Facilitates Refolding of Citrate Synthase by Suppressing Aggregation

Johannes Buchner,*,‡ Marion Schmidt,‡ Miriam Fuchs,‡ Rainer Jaenicke,‡ Rainer Rudolph,§ Franz X. Schmid,‖ and Thomas Kiefhaber*,‖

Institut für Biophysik und Physikalische Biochemie, Universität Regensburg, Postfach, D-8400 Regensburg, FRG, Boehringer Mannheim GmbH, Forschungszentrum Penzberg, Nonnenwald 2, D-8122 Penzberg, FRG, and Laboratorium für Biochemie, Universität Bayreuth, D-8580 Bayreuth, FRG

Received August 16, 1990; Revised Manuscript Received October 17, 1990

ABSTRACT: The molecular chaperone GroE facilitates correct protein folding in vivo and in vitro. The mode of action of GroE was investigated by using refolding of citrate synthase as a model system. In vitro denaturation of this dimeric protein is almost irreversible, since the refolding polypeptide chains aggregate rapidly, as shown directly by a strong, concentration-dependent increase in light scattering. The yields of reactivated citrate synthase were strongly increased upon addition of GroE and MgATP. GroE inhibits aggregation reactions that compete with correct protein folding, as indicated by specific suppression of light scattering. GroEL rapidly forms a complex with unfolded or partially folded citrate synthase molecules. In this complex the refolding protein is protected from aggregation. Addition of GroES and ATP hydrolysis is required to release the polypeptide chain bound to GroEL and to allow further folding to its final, active state.

Correct in vivo folding and assembly of newly formed polypeptide chains appears to be dependent on the presence of several cellular proteins. These "molecular chaperones" (Laskey et al. 1978; Ellis, 1987, 1990), which belong to the group of heat-shock proteins, can interact with nonnative or partially folded polypeptide chains in an ATP-dependent manner (Ellis & van der Vies, 1988; Roy et al., 1988; Bochkareva et al., 1988; Ostermann et al., 1989). GroEL (also called cpn60[1]) from *Escherichia coli* is a prominent member

[*] Corresponding authors.
[‡] Universität Regensburg.
[§] Boehringer Mannheim.
[‖] Universität Bayreuth.

0006-2960/91/0430-1586$02.50/0 © 1991 American Chemical Society