Discrete representations of the protein C_α chain

Xavier F de la Cruz[1], Michael W Mahoney[2] and Byungkook Lee

Background: When a large number of protein conformations are generated and screened, as in protein structure prediction studies, it is often advantageous to change the conformation in units of four consecutive residues at a time. The internal geometry of a chain of four consecutive C_α atoms is completely described by means of the three angles $\theta 1$, τ, and $\theta 2$, where τ is the virtual torsion angle defined by the four atoms and $\theta 1$ and $\theta 2$ are the virtual bond angles flanking the torsion angle on either side. In this paper, we examine the quality of the protein structures that can be obtained when they are represented by means of a set of discrete values for these angles (discrete states).

Results: Different models were produced by selecting various different discrete states. The performance of these models was tested by rebuilding the C_α chains of 139 high-resolution nonhomologous protein structures using the build-up procedure of Park and Levitt. We find that the discrete state models introduce distortions at three levels, which can be measured by means of the 'context-free', 'in-context', and the overall root-mean-square deviation of the C_α coordinates (crms), respectively, and we find that these different levels of distortions are interrelated. As found by Park and Levitt, the overall crms decreases smoothly for most models with the complexity of the model. However, the decrease is significantly faster with our models than observed by Park and Levitt with their models. We also find that it is possible to choose models that perform considerably worse than expected from this smooth dependence on complexity.

Conclusions: Of our models, the most suitable for use in initial protein folding studies appears to be model S8, in which the effective number of states available for a given residue quartet is 6.5. This model builds helices, β-strands, and coil/loop structures with approximately equal quality and gives the overall crms value of 1.9 Å on average with relatively little variation among the different proteins tried.

Addresses: Laboratory of Molecular Biology, Division of Basic Sciences, National Cancer Institute, National Institutes of Health, Building 37, Room 4B15, 37 Convent Drive, MSC 4255, Bethesda, MD 20892-4255, USA. [1]Present address: Biomolecular Structure and Modeling Unit, Department of Biochemistry and Molecular Biology, University College, London WC1E 6BT, UK. [2]Present address: Department of Physics, Yale University, New Haven, CT 06520, USA.

Correspondence: Byungkook Lee
E-mail: bkl@helix.nih.gov

Key words: alpha carbon chain, discrete representation, virtual angle

Received: 06 May 1997
Accepted: 12 May 1997

Published: 25 Jun 1997
Electronic identifier: 1359-0278-002-00223

Folding & Design 25 Jun 1997, 2:223–234

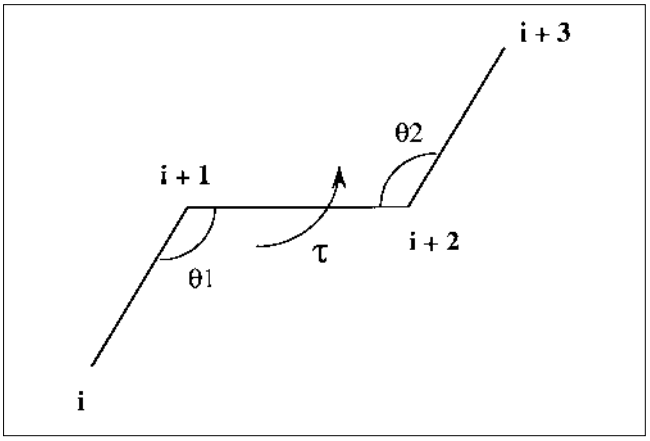
Introduction

One of the major problems faced when trying to predict the three-dimensional structure of protein molecules is the enormous size of their conformational space [1,2]. One way to reduce this vast conformational space is to discretize it by allowing only a small number of states for each residue. A common strategy is to place the protein chain on a lattice [1–4]. Another, perhaps less popular, procedure is to discretize the dihedral angles of the structure. For example, Rooman *et al.* [5] discretized the Ramachandran map using six or seven discrete states to represent the heavily populated areas of the map. This approach gave promising results in modeling small peptides [5] and has been used in combination with genetic algorithms for protein structure prediction studies [6,7].

An obvious problem with these discrete state models is that they introduce distortions to the protein structure. The quality of the protein backbone structure that the models produce is measured in this article mainly by means of the root-mean-square deviation of the C_α coordinates (crms). It varies according to the type and number of conformational states (complexity of the model) allowed for each residue or a combination of the residues. For lattice models, for example, the accuracy ranges from 5.39 Å crms for a simple lattice with $\sqrt{3}$ states per lattice point to 0.90 Å crms for the complex 55-state lattice model [8]. The choice of a particular model is therefore an exercise in compromise between simple models that will reduce the search space and more complex ones that will increase the fidelity of the structure produced.

In the present article, we study the quality of the structures that can be obtained by discretizing a particular type of representation that we use for the structure prediction studies. In this representation, the conformation of a protein is described by the C_α chain only and the conformational search variables are the set of three virtual bond and torsion angles that will completely define the local geometry of four consecutive C_α atoms. For a set of four consecutive residues (i to i+3), the

Figure 1

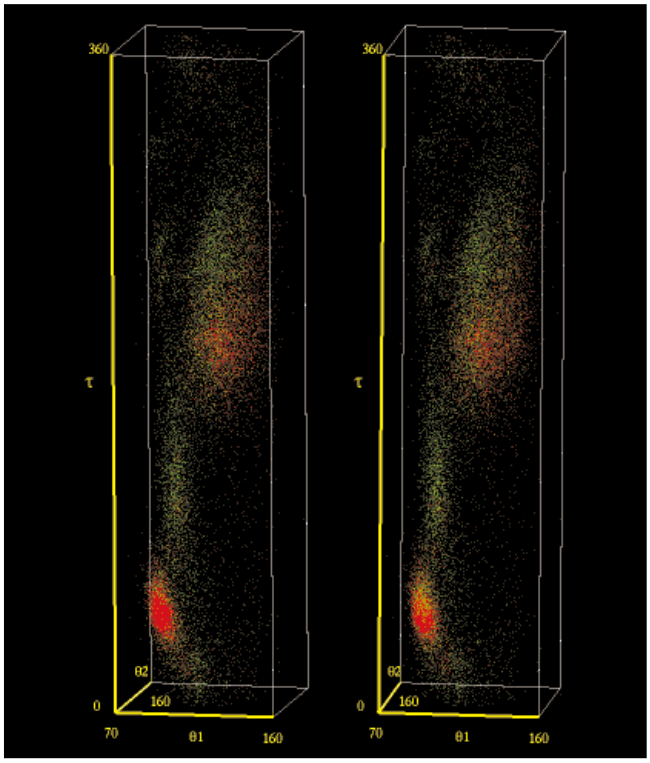


Definition of the three angles θ1, τ, and θ2. The four line segments connect four consecutive C_α atoms, which are numbered *i* to *i*+3.

three angles consist of the virtual torsion angle τ defined by the four atoms (*i*, *i*+1, *i*+2, and *i*+3) and the two flanking virtual bond angles θ1 and θ2 defined by the three atoms (*i*, *i*+1, and *i*+2) and (*i*+1, *i*+2, and *i*+3), respectively (Figure 1). Oldfield and Hubbard [9] have already reported on the frequency distribution of this particular set of angles in the database of protein structures. The description based on the four-residue unit is chosen over other possible C_α-based representations (e.g. see [10]) because four residues are the basic unit of an α-helix or a β-hairpin turn and because any unit longer than four residues will require more than three angles to specify. Flocco and Mowbray [11] recently showed that the torsion angle defined by the four consecutive C_α atoms can be used to analyze the protein conformational changes.

Thus, the state of each quartet of residues (*i* to *i*+3) is described by means of a set of values for the θ1–τ–θ2 angle triplets, which are analogous to the ϕ–ψ angle pairs in the study of Rooman *et al.* [5]. The representation is discretized by allowing only a small number of discrete values, or states, for this triplet of angles. Different choices for the discrete angle values result in different models and the complexity of a model is determined by the number of discrete states allowed (plus the connectivity considerations, see Materials and methods). We study a number of different models of increasing complexity. Structures were built using each of these models and the efficient build-up procedure of Park and Levitt [8], which minimizes the crms from the known native structure. Examination of the quality of these structures yields interesting insights into the relation between the complexity of a model and the fidelity of the structure produced.

Figure 2



Stereoview of the distribution in the θ1–τ–θ2 space of the conformation of all residue quartets in the 139 database proteins. The red dots are for the quartets in the α-helical and the β-strand structures. The figure was made using the in-house program GEMM.

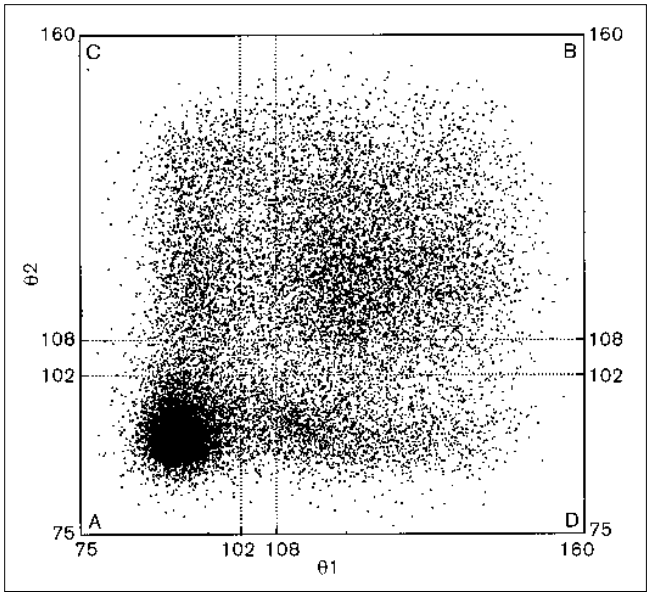
Results

Distribution of the θ1–τ–θ2 angle triplets

Figure 2 gives a stereoview of the distribution of the θ1–τ–θ2 angle triplets found in our database of 139 protein structures (see Materials and methods). The three-dimensional map is consistent with the one described by Oldfield and Hubbard [9]. There are no data points with the θ angle outside the range 75° to 160°. Periodic structures such as α-helices and β-strands must have θ1 ≈ θ2, since θ1 of a residue quartet (*i* to *i*+3) is the same as θ2 of the preceding quartet (*i*–1 to *i*+2). The data points for these structures cluster around the point (θ1 = θ2 = 95° and τ = 50°) for the α-helices and (θ1 = θ2 = 125° and τ = 210°) for the β-strands [12]. The data points outside these two clusters, as well as many within the clusters, represent nonperiodic structures, many of which are linkers and turns that connect the two periodic structural elements.

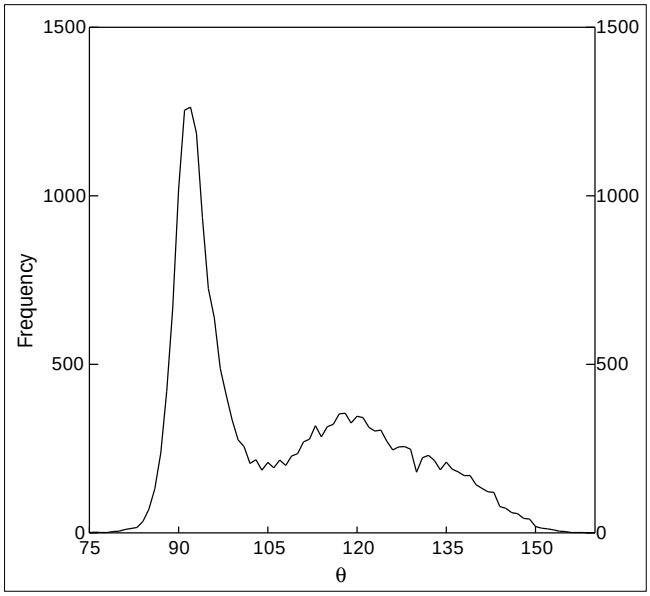
Figure 3 shows the data of Figure 2 projected onto the θ1–θ2 plane. The distribution is nearly symmetric with respect to a flip along the θ1–θ2 diagonal. Four main clusters can be identified in this two-dimensional map. Two of them, in the regions marked A and B in the figure, are

Figure 3



Projection of the θ1–τ–θ2 map onto the θ1–θ2 plane. The plane is divided into four regions, which are labeled A–D. Two sets of dividing lines are shown: one at θ1 = θ2 = 102° and another at θ1 = θ2 = 108°. The latter set was used in this work.

Figure 4



Frequency distribution of the θ angle.

centered on the diagonal of the plot, have a more or less circular shape, and contain all of the α-helix and β-strand states, respectively. The other two, in the regions marked C and D in the figure, are elongated along one of the coordinate axes. The points in region C have the θ1 angles of an α-helix and the θ2 angles of a β-strand structure. These points can serve as linkers that connect an α-helix to a β-strand. Similarly, the cluster of points in region D have the θ1 angles of a β-strand and the θ2 angles of an α-helix and can serve as β-to-α linkers.

In order to give a more precise definition to these clusters, the θ1–θ2 plane was divided into four rectangular regions by means of vertical and horizontal lines (Figure 3). These lines were placed at θ1 = θ2 = 108°, which is near the minimum in the distribution of the data points as a function of the θ angle (Figure 4). The clusters were then defined as all the points that fall within the corresponding rectangular regions.

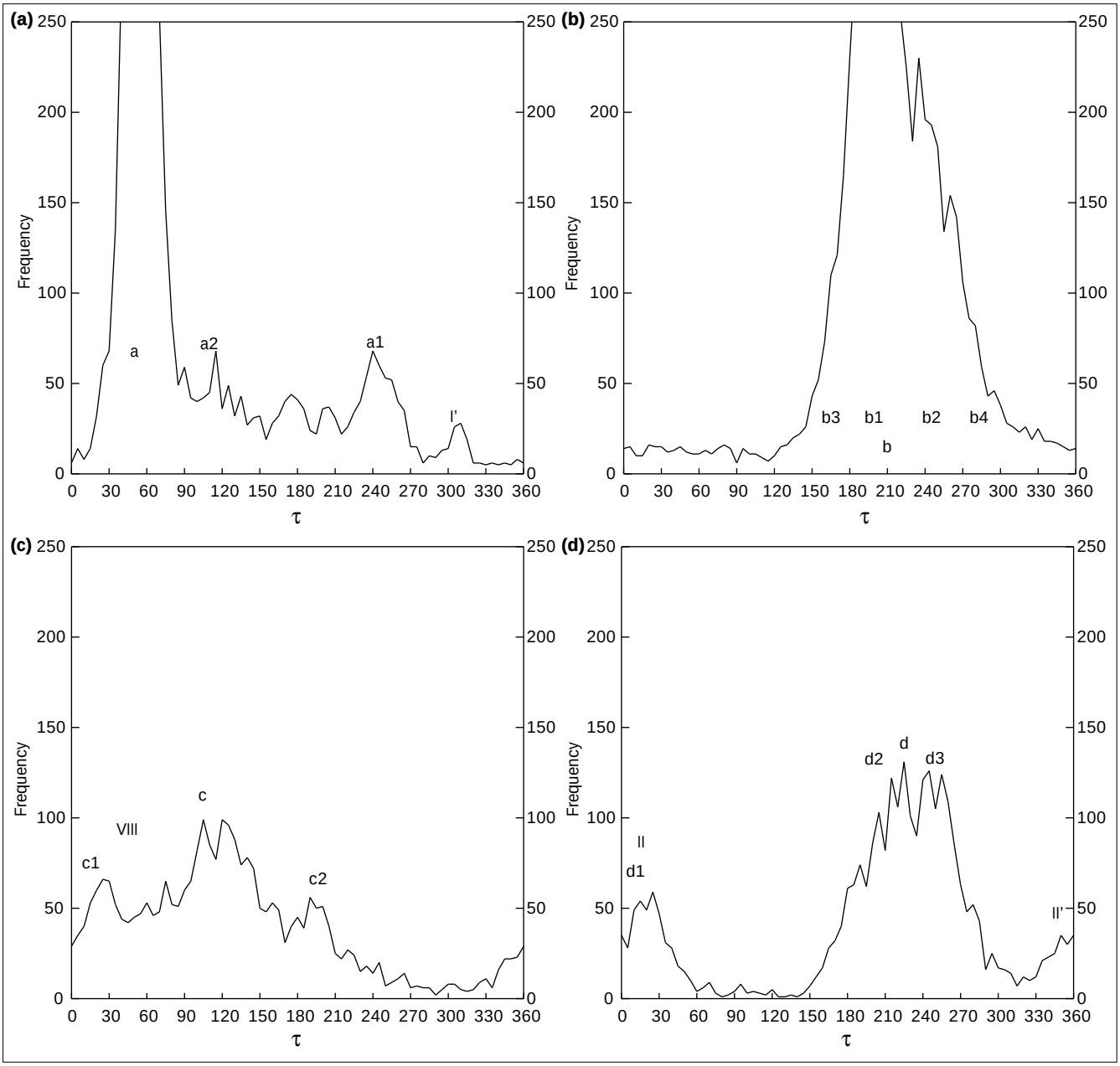
The distribution of the data points as a function of the τ angle is shown for each of the four clusters in Figure 5. Cluster A has a large sharp peak at $\tau \approx 50°$ (Figure 5a), which obviously contains all of the α-helical structures. However, there are also a significant number of points outside this peak which are clearly not α-helical since their τ angles are large. Cluster B contains one large broad peak (Figure 5b). The breadth of the peak presumably reflects the large variation that is possible for β-structures, but the peak also includes many of the points for nonperiodic structures (see Figure 2). Predictably, the linker clusters C and D show broad distributions (Figure 5c,d) that indicate that many discrete states will be required in order to represent these clusters.

Choice of the discrete values for θ1 and θ2

The distribution pattern of Figure 3 indicates that the simplest way to discretize the θ1–θ2 space is to use four discrete points, one to represent each of the A, B, C, and D clusters. Models that use more than one point per cluster, particularly for the B, C, and D clusters, can be interesting, but they are necessarily more complex and have not been investigated in this study. Even when the θ1–θ2 space is restricted to only four discrete points, the representation can become complex because of the need to choose many different values for the τ angle, as will be seen later.

The points representing clusters A and B are placed on the diagonal and their placement requires two angle values, θ_α and θ_β, respectively. The position representing cluster C must then be placed at $\theta1 = \theta_\alpha$ and $\theta2 = \theta_\beta$ because otherwise the state represented cannot function as an α-to-β linker. Similarly, the point that represents cluster D is placed at $\theta1 = \theta_\beta$ and $\theta2 = \theta_\alpha$. Therefore, only two discrete values of θ are needed in order to define all four θ1–θ2 states. The values for θ_α and θ_β were chosen to be 94° and 124°, respectively. These were obtained by averaging θ values of the appropriate sets of data points.

Figure 5



Frequency distribution of the τ angle for the data points in (a) cluster A, (b) cluster B, (c) cluster C, and (d) cluster D of Figure 3. The approximate locations of the τ values used for different discrete states are indicated.

Representation with the minimal set

We started by studying the representation using the minimal set, S1, in which only four discrete states were allowed, one for each of the four clusters in the θ1–θ2 map—these were termed a, b, c, and d (Table 1). The τ angles for these states were chosen as the mean of data points around the main peak in the τ angle distribution for each cluster shown in Figure 5. The values of these angles are listed in Table 2 together with those for all other states used in this work. Although four states are allowed, the complexity index of this model is 2 because of the connectivity constraint.

When S1 was used to rebuild the structure of the proteins in the database, according to the build-up procedure described in the Materials and methods, the average crms

Table 1

The eight sets of discrete states used in this work*.

S1	a, b, c, d
S2	a, b1, b2, c, d
S3	a, a1, b, c, d
S4	a, a1, a2, b, c, d
S5	a, a1, a2, b1, b2, c, d
S6	a, b, c, d, I′, II, II′, VIII
S7	a, a1, a2, b1, b2, c, c1, c2, d1, d2, d3
S8	a, a1, a2, b1, b2, b3, b4, c, c1, c2, d1, d2, d3

*The discrete state codes are defined in Table 2.

Table 2

θ1, θ2, and τ angles, in degrees, of the discrete states used*.

Name	θ1	θ2	τ
a, a1, a2, I′	94	94	51, 243, 119, 309
b, b1, b2, b3, b4	124	124	214, 200, 250, 165, 285
c, c1, c2, VIII	94	124	109, 16, 195, 43
d, d1, d2, d3, II, II′	124	94	228, 9, 203, 259, 18, 342

*All discrete states listed in one row have the same θ1 and θ2 values but different τ values, which are listed in the last column in the same order in which the names of the discrete states are given.

obtained was 5.29 Å (Table 3). Park and Levitt [8] observed that the quality of a discrete representation increases with its complexity and generally follows the relation:

$$\log(\langle\text{crms}\rangle) = -0.514 * \log(\text{complexity}) + 0.816 \qquad (1)$$

Figure 6 shows that S1 gives a result that is poorer than expected from this formula.

Individual secondary structural elements were reproduced with much higher accuracy (Table 3), but still rather poorly than might be expected. For example, in 1eca, the majority of the α-helices were distorted and showed high crms values: 2.9 Å, 2.2 Å, 4.2 Å, 3.1 Å, 3.2 Å, and 2.3 Å for helices B, D, E, F, G, and H, respectively. Helix A was an exception, with a crms of 0.4 Å. This is due to the fact that the build-up procedure used builds the structure from the N to the C terminus of the protein. Helix A is at the N-terminal end of the protein and almost no cumulative error is introduced during the application of the build-up procedure. Helix C was also an exception, with a crms of 0.2 Å. This is a four-residue helix, and the a state was assigned to this quartet. This indicates that independent helices may be well reproduced using our a state.

In order to assess more generally the inherent ability of this model to represent the secondary structural elements, we extracted α-helices and β-strands from four representative proteins and rebuilt them independently of the rest of the structure. The average context-free crms values obtained were 0.52 Å, 0.72 Å, 1.3 Å, and 0.91 Å for the 1eca and 1ilk helices and the 1bcx and 1cbs strands, respectively. (The averages are given in parentheses in Table 3.) Thus, independent pieces of secondary structures can be reproduced with high quality using this set, particularly the α-helices. One helix, the first helix in 1ilk, was an exception and showed a 2.22 Å crms. This helix has a significant distortion at residues 31 and 32, as specifically noted in the header of the PDB file for this protein. The poorer performance for the β-strands was to be expected in view of the greater structural heterogeneity of the β-strands, reflected in the fact that the β-strand state occupies a broader volume in the θ1–τ–θ2 map than the α-helix state.

The linkers (nonperiodic structures) were likewise extracted from the four representative structures and rebuilt independently. The context-free crms values obtained, 1.1 Å, 1.34 Å, 1.36 Å, and 1.26 Å for 1eca, 1ilk, 1cbs, and 1bcx, respectively, were also substantially higher than those for the α-helices.

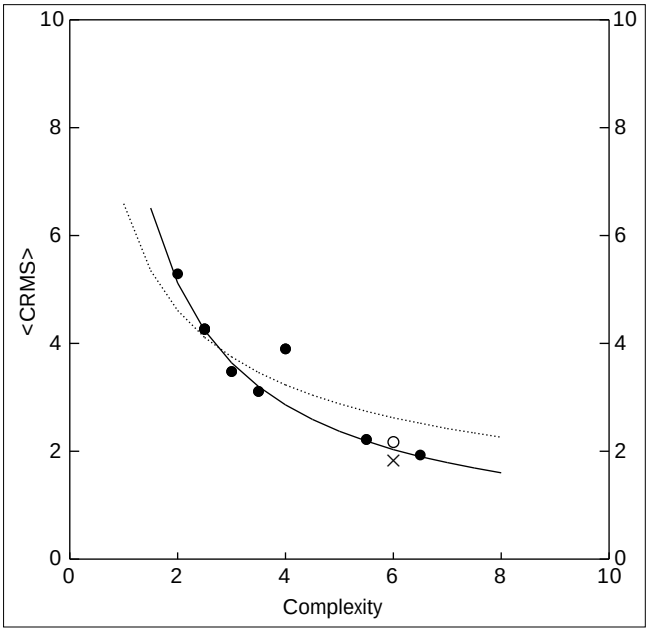
Table 3

Characteristics and performance of the different sets of states used in this work.

	Number of states	Complexity index	⟨crms⟩*	⟨a⟩†	⟨b⟩†	⟨c⟩†
S1	4	2	5.29	2.25 (0.62)	1.56 (1.12)	1.91 (1.25)
S2	5	2.5	4.27	2.15 (0.62)	1.20 (0.88)	1.71 (1.21)
S3	5	2.5	4.26	2.00 (0.60)	1.49 (1.11)	1.68 (1.07)
S4	6	3	3.48	1.41 (0.56)	1.50 (1.08)	1.52 (0.98)
S5	7	3.5	3.11	1.36 (0.56)	1.20 (0.87)	1.43 (0.96)
S6	8	4.0	3.90	1.51 (0.56)	1.42 (1.10)	1.62 (0.87)
S7	11	5.5	2.22	1.09 (0.56)	0.98 (0.72)	1.09 (0.67)
S8	13	6.5	1.93	0.97 (0.56)	0.84 (0.58)	0.96 (0.65)

*Average crms over all the proteins in the database used (see Materials and methods). †Average 'local' crms for α-helix (⟨a⟩), β-strand (⟨b⟩), and loop/coil (⟨c⟩). These were computed after superimposing the secondary structure element of the rebuilt protein on the same secondary structure element of the X-ray structure. The numbers in parentheses are the 'context-free' crms values obtained when the secondary structure was excised out and rebuilt independently of the rest of the protein.

Figure 6



Average crms obtained as a function of the complexity index for different models. The solid circles represent our models, with increasing complexity from S1 to S8. (The points for models S2 and S3, both with complexity 2.5, are superimposed.) The cross and the open circle are the data for the set of Rooman *et al.* [5] with (circle) and without (cross) the bump check. These calculations were made using our data set of proteins. The solid line is the plot of equation 2. The dotted line is the plot of equation 1, which represents the best fit line for the data of Park and Levitt [8].

Interestingly, when built within the context of the native structure, the average in-context crms was best for the β-strands and poorest for the α-helices (Table 3). This contrasting behavior implies that the build-up procedure was forced to use non-α-helical discrete states for some α-helical residues in order to minimize the overall crms. A similar observation was made earlier by Rooman *et al.* [5] in their study with the discrete ϕ–¬ψ angle sets. Thus, the requirement for an overall crms minimization introduces an interdependence among the reproduction qualities of different secondary structural elements; a poor ability to reproduce the β-strand and loop/coil structures also leads to low quality α-helices, which otherwise could have been correctly built.

More complex models

By selecting different combinations of additional τ values, we generated and tested 30 different models with varying complexity. In this paper, we describe only a small subset (listed in Table 1). The τ angle values that define different states used in these models are given in Table 2. These were selected according to the scheme described in the Materials and methods. The complexity indexes and different performance measures of these models are summarized in Table 3. Other models use somewhat different τ angles and/or different combinations of them. Their characteristics are broadly similar to those of the models given in the table.

The minimal set S1 consists of four states, one state to represent each of the four θ1–θ2 clusters A, B, C, and D. Set S2 uses two states for cluster B, while S3 uses two states for cluster A. States b1 and b2 of S2 both represent the broad β-strand peak in the τ distribution of cluster B (Figure 5b). On the other hand, state a1 of S3 represents a non-α-helical turn conformation in cluster A (Figure 5a). Set S4 is an extension of S3 and represents the wide τ distribution of the non-α-helical conformations of cluster A (Figure 5a) by two states, a1 and a2. Both of these represent rather low peaks in the distribution, but their presence seems to give rise to a surprisingly large improvement in the reproduction quality (see below). Set S5 is a combination of sets S2 and S4.

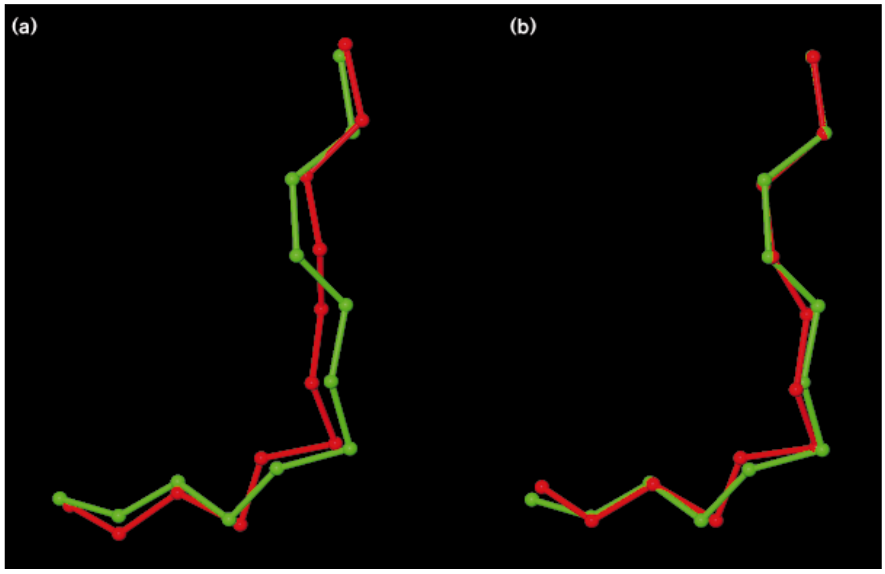
The four states used in S6, in addition to the minimal set, are based on the β-turn types identified by Hutchinson and Thornton [13]. These authors described eight β-turn types. We discarded the three type VI turns, which have a *cis*-proline at position *i+2* of the turn, because of their very low occurrence in known protein structures [13]. The θ1–τ–θ2 angles for the remaining five β-turn types, I, II, I′, II′, and VIII, were determined from the reported average ϕ–ψ angle values and discretized according to the procedure described in the Materials and methods. We then discarded type I because this type became essentially the same as state a in terms of the discretized θ1–τ–θ2 angles.

In set S7, clusters C and D are each represented by three states. Finally, set S8 is obtained by adding two more β-strand states to set S7. The relation between the overall reproduction quality of these different models and their complexity is shown in Figure 6.

β-strand reproduction qualities

As noted earlier, the main peak in the τ angle distribution of cluster A is rather sharp, but that of cluster B is broad (Figure 5a,b). Thus, α-helix is expected to be well represented by one τ angle, but the β-strand is not. The quality of the β-strand reproduction can be followed by the context-free crms values (numbers in parentheses in Table 3). The effect of using more than one state to represent the β-strand can be seen by comparing these crms values for sets S1, S3, S4, and S6 on the one hand and those for S2 and S5 on the other. Thus, when β-strands are rebuilt independently of the rest of the protein, S1 gives the crms value of 1.12 Å and this accuracy is not improved by the use of a more complex set, S3, S4, or S6, which all use only one state to represent the β-strand. In contrast, sets S2 and S5, which have complexities that are

Figure 7



The superposition of the rebuilt (red) and the native (green) structures, using **(a)** model S7 and **(b)** model S8, for the β-strand of residues 163–174 of xylanase (1bcx), which has a β-bulge.

comparable to these sets but which use two states to represent the β-strand, give the noticeably lower crms values of 0.88 and 0.87 Å.

It is interesting that there is a slight further improvement between sets S5 and S7, even though they both use the same two states, b1 and b2, to represent the B cluster. The reason that addition of linker states improves the reproduction quality of β-strands is probably related to the fact that both b1 and b2 states use only one θ angle, θ_β; use of the linker states introduces another θ angle, which probably helps the reproduction quality by increasing the degree of freedom. Use of four states, b1, b2, b3, and b4, in S8 further improves the β-strand reproduction ability to a level similar to that for the α-helix. (Compare the numbers in parentheses in columns 5 and 6 in Table 3.) This last improvement was partly due to a better handling of β-bulges. For example, the β-strand from residues 163 to 174 of xylanase (1bcx) has a β-bulge. Its context-free crms is 0.94 Å when S7 is used to represent it, but drops to 0.59 Å when S8 is used (see Figure 7).

The accuracy with which the β-strands can be reproduced in the context of the whole protein (in-context reproduction) generally follows the same trend as that for the context-free reproduction (Table 3, column 6). However, the crms values for the former are 30–50% higher than those for the latter. The quality of the in-context reproduction is similarly worse than that of the context-free reproduction for the loop/coil residues and considerably worse for the helical residues. This again implies that the local structure had to be deliberately distorted from the best achievable for a given model set in order to optimize the reproduction quality of the overall structure.

A notable feature, when comparing S1 and S2, is that the addition of a B cluster state significantly improves the in-context reproduction quality of α-helices and coils also. This is partly due to the interdependence among reproduction qualities mentioned earlier; a better representation of β-strands makes it possible to build non-β-strand structures better also. Another reason, though, is simply that more states are available in S2 for use in building the coil structures. Indeed, we find that 24% of the coil/loop residues are modeled using the b state when S1 is used, whereas 33% of them are modeled using the b1 and b2 states when S2 is used. Similar improvements in the in-context reproduction qualities of α-helices and coil/loop residues are observed upon addition of two more B cluster states in going from S7 to S8, in which case the number of such states used in coils increases from 19% for S7 to 29% for S8.

Helix and turn reproduction qualities

Helices can be built with very high accuracy even with the minimal set when rebuilt independently of the rest of the protein. This is to be expected in view of the rather narrow width of the main peak in Figure 5a. However, the in-context reproduction quality is particularly poor for helices. For example, the average crms for helices in the S1 model is 2.25 Å (Table 3), which is more than 3.5 times the average context-free crms of 0.62 Å. Thus, substantial improvements in the quality of reproduction of helices can come only from improving the β-strand and the loop/coil conformations.

Splitting the b state into two (S2 compared to S1 and S5 compared to S4) does produce small improvements in the reproduction quality of helices and turns (Table 3).

Addition of the a1 turn state to the minimal set (S3 compared to S1) produces larger improvements, particularly in the quality of the context-free reproduction of turn residues (from 1.25 to 1.07 Å in crms, a 20% reduction). Addition of one more A cluster state, a2 (S4 compared to S3), produces an unexpectedly large improvement (from 2.00 to 1.41 Å in average crms) in the quality of the in-context reproduction of helices. S4 performs substantially better than S1, both in terms of the overall crms, 3.48 Å, and in terms of the local in-context secondary structure crms values, 1.41 Å, 1.50 Å, and 1.52 Å, for the α-helix, β-strand, and loop/coil, respectively.

A glimpse of the mechanism by which these improvements (S4 over S1) are brought about can be obtained from the following observations. We found that the percentage of α-helical residue quartets modeled using the proper helical state, a, was 68% in S4, compared to 53% in S1. This implies that the two new states in S4, a1 and a2, were used to build nonhelical structures better so that less distortion was necessary when building the helices. Indeed, the usage of the two new states in different secondary structures were 24%, 23%, and 53%, respectively, for the α-helices, β-strands, and loop/coils. Since the total number of residue quartets in the database is 34% α-helical, 23% in β-strands, and 43% in loop/coil states, the two new states are preferentially used in the loop/coils and less in α-helices. Thus, at least part of the improvement in the reproduction quality of the α-helices was brought about because the a1 and a2 states helped to allow better position and orientation for them. At the same time, and partly because of this, over 80% of the α-helical residue quartets use the A cluster states in S4 compared to 53% for S1. Since A cluster states use the proper θ values for the α helix, this will also help improve the quality of helix reproduction. It is notable that this improvement in the reproduction quality of α-helices was achieved with a substantially smaller improvement in that of the turn conformations; the improvement in the in-context reconstruction of α-helices is 37% (1.41 Å for S4 versus 2.25 Å for S1) whereas the improvement for the turn residues is only ~ 20% for both the context-free (0.98 Å versus 1.25 Å) and in-context (1.52 Å versus 1.91 Å) reproductions.

The a1 and a2 states are obviously not the only way to represent the turn conformations. Instead of these states, we tried in S6 the four states derived from the turn conformation list of Hutchinson and Thornton [13]. S6 reproduces the turn conformations better when they are isolated (0.87 Å versus 0.98 Å crms) and the β-strands better in the context of the whole protein (1.42 Å versus 1.50 Å crms). Surprisingly, however, it produces worse results than S4 in all other categories despite the fact that its complexity index is higher than that of S4. It is probable that this result is, at least in part, due to fact that the θ angles were changed from the original turn conformations in order to discretize them in accordance with the connectivity constraint.

The majority of turn residues are, of course, in clusters C and D. These clusters have wide τ distributions (Figure 5c,d) which require a large number of discrete states for their fair representation. Set S7 includes three discrete states for each of these clusters. This expansion in the number of states over S5 results in a substantial improvement in the quality of the context-free reconstruction of the turns (0.67 Å versus 0.96 Å average crms, Table 3) and β-strands (0.72 Å versus 0.87 Å average crms), but no improvement for the α-helices. On the other hand, the in-context average crms values improve more for the turns and helices than for the β-strands so that they are now quite similar to each other (1.09 Å, 0.98 Å, and 1.09 Å for the α-helix, β-strand and loop/coil residues, respectively).

Discussion

When a small number of discrete states are used to represent a protein structure, the structure is necessarily distorted. The results of the studies reported herein show that there are at least three levels to this distortion. At the most elementary level, the use of the discrete states will introduce distortions to the individual secondary structural elements even when they are isolated from the protein structure and rebuilt independently of the rest of the structure. This level of distortion is reflected in the reported context-free crms values (Table 3). Additional distortions are introduced when these structures are built as a part of the whole structure, as can be seen from the fact that the in-context crms values are considerably poorer than the context-free crms values. As pointed out by Rooman *et al.* [5] before us, this implies that structures are deliberately distorted away from the best possible local structure in order to achieve the globally best fit. Finally, even when the different secondary structural elements are built relatively well locally, their relative global arrangement can be poor. This results in a further deterioration in the overall crms of the whole structure (Table 3) over and above the 'local' crms values of the individual secondary structures. Distortions are significant at each of these levels. With set S8, for example, the overall crms value is nearly twice the in-context crms values for individual secondary structural elements, which are in turn approximately twice their context-free crms values.

The wider conformational possibilities of the β-strands and turns mean that more discrete states are required to represent these structures than the α-helices. Therefore, S1, which uses one state for each of the A, B, C, and D clusters, produces poor context-free crms values for β-strands and turns. Strikingly, the poor ability of this model

to reproduce the β-strands and turns results in the poorest in-context fidelity of reproduction for the helices. An even distribution of the context-free reproduction qualities across different local structural types occurs only after inclusion of a sufficient number of non-α-helical states in the discrete set. These sets (S7 and S8) also give a similarly even distribution of the in-context reproduction qualities for different local structural types.

The overall accuracy of the structure reproduction will generally increase with the number of states included in the discrete set. Park and Levitt [8] found that the overall crms values obtained using many different models they generated decreased smoothly with the complexity of the model and closely follow equation 1. The overall crms values of our models are plotted against the complexity index in Figure 6. It shows that they also fall smoothly with the complexity index, but that the rate of fall is significantly faster than expected from Park and Levitt's formula. The data can be fitted to equation 2:

$$\log(\langle\text{crms}\rangle) = -0.839 * \log(\text{complexity}) + 2.214 \qquad (2)$$

with a correlation coefficient of 0.997. Set S6 is clearly an exception and was left out of this correlation.

In order to see why our models show more rapid improvement with the increase in complexity index, it is instructive to review Park and Levitt's argument for expecting approximately the $m^{-1/2}$ dependence that they observe, where m is the complexity index. Suppose that the protein chain has been perfectly built up to residue $i+1$. If the protein chain is built one residue at a time, the next residue, $i+2$, can now be placed at m different places on the surface of the sphere of radius b, the fixed Cα–Cα distance, centered at residue $i+1$. The average distance between different choices will then be given by the square-root of the search area (the surface area of the sphere) divided by m. This gives the $m^{-1/2}$ dependence to this average distance and, presumably, to the overall crms of the model.

Suppose, however, that after building residue $i+1$, residue $i+2$ can now be placed at m different places on the perimeter of a circle instead of the surface of a sphere. Following the line of reasoning of Park and Levitt, the crms will now depend on m^{-1}. Our build-up procedure does not exactly place residue $i+2$ on a circle, since both θ and τ are varied. However, the main source of variation among our different models comes from τ. Therefore, the expected dependence will be between m^{-1} and $m^{-1/2}$, as observed. In any case, the best fit curve of our data and that of Park and Levitt meet between the complexity indices of 2.5 and 3.0 and all the sets that include the a1 and a2 states perform better than expected by Park and Levitt's formula.

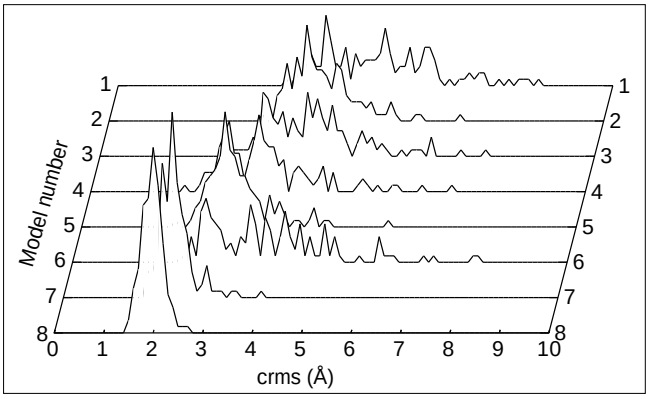
We note also that set S6 is an exception and performs worse than expected from either formula. There are other sets among the 30 that we tried that perform similarly poorly. We have not investigated the reason why some sets give poor results, but one possibility is that one or more of the states included in such sets is rarely used in the final structure. In the case of S6, the I′ state is in fact used very little; only 3% of all residue quartets in the reconstructed structure are in the I′ state. It is probable that states like I′ are critical for the attainment of the finely tuned final structure of a protein molecule. During initial folding attempts at low resolution, however, it may be more important to reduce complexity by substituting other more versatile states for these rarely occurring specialized states.

The complexity of the ϕ–ψ dihedral angle set of Rooman *et al.* [5] is 6 and produces an average crms of 2.17 Å when applied to our set of proteins with bump check and 1.84 Å when used without the bump check. This latter value is comparable to the 1.74 Å value reported for this set by Park and Levitt using their protein structure database. The value of 2.17 Å is only slightly worse than expected from equation 2, which is a correlation using the crms values obtained after the bump check.

Sets S5, S7, and S8 are probably suitable for initial folding trials. These sets produce α-helices, β-strands, and turns with roughly equal accuracy. The overall crms is 3.1 Å or better on average. In a recent article, Orengo *et al.* [14] reported on comparison studies of a series of protein structures whose similarities vary from 1.9 Å to 6.4 Å. However, for the vast majority of the comparisons, the crms was below 4.0 Å. In the same study, the authors use a threshold of 4.5 Å to identify recurring structural motifs in proteins. This suggests that, if an *ab initio* folding program produces a structure that is 4.0–4.5 Å crms away from the true structure, the latter may be recognizable from the calculated structure. This level of accuracy may be achievable with discrete sets S5, S7, and S8 since their average inherent fidelity of reproduction is at least 1 Å better than this level of accuracy.

The accuracy of reproduction that we cite above is average crms values over the 139 proteins. Since we deal with only one protein at a time in any real folding experiment, it is of interest to know the spread of accuracy over different proteins. The distribution of crms values over different proteins is shown in Figure 8 for the different models. It can be seen that the distribution tightens as the average crms value decreases. All models, except S8, build some proteins with a crms greater than 3 Å. This suggests that, in order to build the structure of a given protein that can later be refined to a useful degree of accuracy, one probably has to use a model that is at least as complex as S8.

Figure 8



The distribution of crms values over the database of 139 proteins for models S1 to S8.

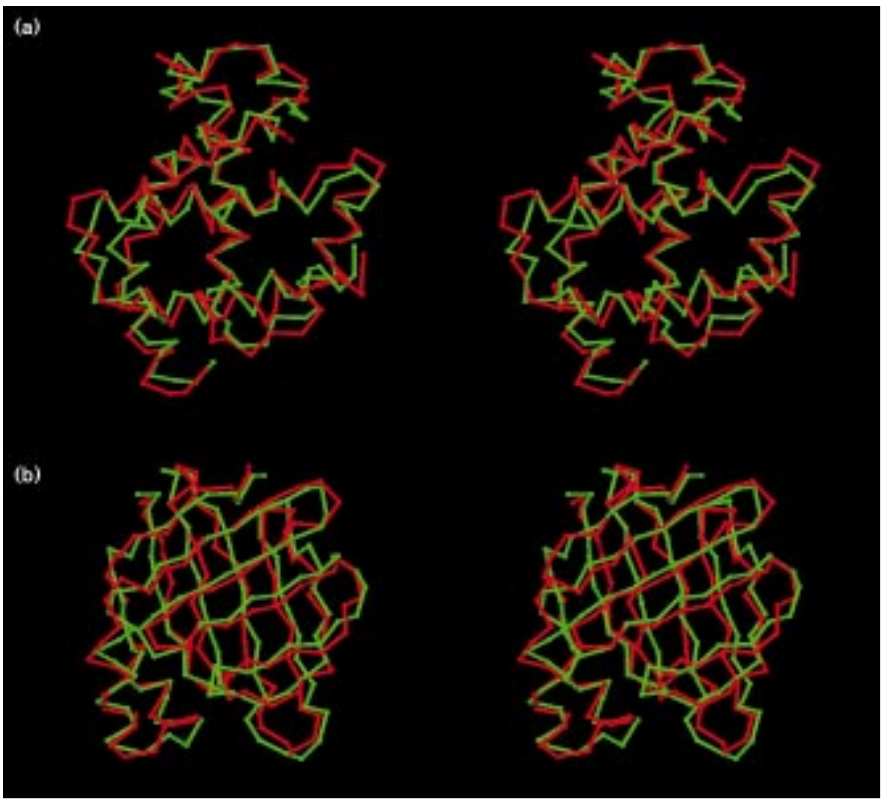
The results of Figure 6 also indicate that accuracies better than these can be achieved only by increasing the complexity substantially. As pointed out by Park and Levitt [8], the reproduction accuracy increases progressively more slowly as the number of states in the set increases, particularly after the crms reaches the 2–3 Å range. Fortunately, higher accuracy may not be needed. The quality of the reproduction with set S8 is generally excellent when inspected visually. Two examples are shown in Figure 9. Dandekar and Argos [6,7] used the set of Rooman *et al.* [5], which has the comparable complexity index of 6, in a series of structure prediction studies and showed that it can yield the correct topology for a collection of proteins. Our representation, which uses four residue units at a time, should perform better since it includes built-in correlation between adjacent residues. Such correlation is of course absent in the ϕ–ψ angle model, which handles each residue independently. The $\theta1$–τ–$\theta2$ representation incorporates the correlation between adjacent residue quartets as well, through the requirement of the connectivity constraint. The price that must be paid for these advantages is that the peptide group must be built after the C_α chain has been built. This can be done by an efficient algorithm (unpublished data) that builds peptide groups in optimum orientation for a given C_α chain.

Materials and methods

The protein database

The set of proteins we used were obtained using the OBSTRUCT program [15]. This program allows the selection of a subset of proteins from the PDB [16] according to a series of criteria provided by the user. The selection criteria we used were: no NMR structures, number of residues between 50 and 300, sequence identity <25%, and resolution >1.8 Å. From the set of proteins given by OBSTRUCT, we

Figure 9



Superposition of the rebuilt (red) and the native (green) structures of (a) erythrocruorin (1eca) and (b) retinoic acid binding protein (1cbs). The reconstruction used model S8.

eliminated the ones having missing residues. The final number of protein chains was 139.

Division of the θ1–θ2 space and the determination of θ_α and θ_β

In order to determine the two discrete values of θ used in this study, the data of Figure 3 were projected onto the θ1 or θ2 axis. Since θ1 of a quartet of consecutive residues is θ2 of the previous quartet, the two projections give identical results. The common θ distribution for the whole data set (Figure 4) shows a minimum at θ angles in the range between 102° and 108°. We chose the value of 108° for both θ1 and θ2 to divide the θ1–θ2 plane into four rectangular regions, A, B, C, and D, as shown in Figure 3. The figure indicates that a lower value might have been a better choice, but choosing 104°, for example, changes the θ_α, θ_β, and τ values (see below) by at most 3°, which would make no difference in the quality of the structure reproductions. The A, B, C, and D regions include the points representing the α-helix, β-strand, α-to-β linkers, and β-to-α linkers, respectively. The values of θ_α and θ_β were obtained as the average of all θ angles with θ < 108° and with θ > 108°, respectively.

Discretization of the τ angle

The discrete state for a quartet of residues was obtained by assigning discrete values to the θ1–τ–θ2 angles for the quartet. All of the discrete states described in this paper are listed in Table 2. The θ1 and θ2 values were determined as described above. The τ angles were determined as follows.

For states other than I′, II, II′, and VIII, the discrete τ angle values were chosen from the frequency distribution along the τ coordinate (Figure 5) for each of the four clusters of data points in regions A, B, C, and D of the θ1–θ2 plane. For most of the states, the discrete value chosen was the weighted average of the data points that define chosen peaks in these distributions. Peaks were defined by the minimum and maximum τ values (in degrees) as follows: a (10, 90); a1 (215, 280); a2 (90, 155); b (145, 305); c (45, 170); c1 (–25, 45); c2 (170, 230); d (145, 315); d1 (–45, 60); d2 (145, 235); and d3 (235, 315). Since there is only one broad peak in the τ angle distribution for region B, the above strategy could not be used to discretize the data points in this region. Instead, the b state was initially split into two that are 50° apart to generate the b1 and b2 states. Later, two more states, b3 and b4, were added which were 35° further out in either direction along the τ axis.

For states I′, II, II′, and VIII, the four-residue C_α chain structure was constructed using the φ–ψ angles of these turn types as given by Hutchinson and Thornton [13]. The τ angles were then measured from these structures. The θ1 and θ2 values were also measured and then re-set to either θ_α or θ_β, whichever was closer to the measured value.

Complexity index

When the state of a residue quartet (*i* to *i*+3) is fixed, the state of the next quartet (*i*+1 to *i*+4) is restricted since the θ1 angle for the latter must be the same as the θ2 angle for the former. For example, the S1 model uses four states, a, b, c, and d (Table 1). If a quartet of residues (*i* to *i*+3) is in the a state in this model, its θ2 value is θ_α and the state of the next residue quartet (*i*+1 to *i*+4) can only be either another a or the α-to-β linker state, c. We define the complexity index of a set of discrete states as the number of connectable states available to a residue quartet once the state of the preceding quartet has been fixed. This number is significantly smaller than the total number of discrete states used in the model (see Table 3). For some models, the number of connectable states available to a quartet varies depending on the state of the last one built. For such models, the complexity index was computed as the straight average of all possible values. Therefore, this number can be calculated from the composition of the discrete states of the model, before any actual build-up trials. The complexity index is similar to the 'effective complexity' described by Park and Levitt [8]. The average number of available states is slightly higher than this number due to the end effect.

Build-up procedure

The protein structures were reconstructed using the build-up procedure described by Park and Levitt [8] after some slight changes in the run parameters. In the first step, all (discrete) conformations of the first six N-terminal residues were generated and the 200 conformations with the best crms relative to the crystal structure were saved. Then, one residue was added to each one of the saved conformations and the allowed states for the last residue quartet were exhaustively searched. Again, the best new 200 structures were kept. For each generated structure, a bump check was made in order to prevent close non-native contacts. The minimum distance between a pair of non-adjacent C_α atoms was assumed to be 3.8 Å. The C_α–C_α virtual bond distance was set to the same value.

Assessment of the quality of the reproduction

We use the coordinate root-mean-square deviation (crms) as the only quantitative measure of the quality of the reproduced structure. These were computed after best superposition of the rebuilt structure to the native structure using the procedure of Kabsch [17,18]. This measure was used for individual proteins during the build-up procedure and as an average over the data set to assess the quality of each model tested. The percentage of native contacts is another quality measure [8], but initial testings indicated that this measure entirely paralleled the crms measure for our series of models. The root-mean-square deviation of the θ and τ angles can also be used. However, since good reproduction of the overall structure requires choosing some of these angles very differently from those of the native structure, we believe that this is not a good quality measure for discrete state models of low complexity.

The degree with which the secondary structures are reproduced was monitored using crms values calculated in two different ways. In method 1, the α-helices, β-strands, and loops in the structures that were built in the usual manner were superimposed individually onto their respective counterparts from the native structure. This produces the 'local' 'in-context' crms values, which measure the degree with which the different structural elements were reproduced within the context of the whole protein. In method 2, the secondary structural elements were excised out of the protein and rebuilt independently from the rest of the structure. The crms between the structure built this way and its native counterpart is the 'context-free' crms, which measures the degree with which a given model is capable of reproducing the secondary structural element. The context-free crms calculations were made using only four representative proteins: erythrocruorin (PDB ID, 1eca) and interleukin-10 (1ilk), which are predominantly α-helical, and xylanase (1bcx) and retinoic acid binding protein (1cbs), which are predominantly β-sheet structures. In the results shown in the last three columns of Table 3, the crms values computed this way are in parentheses. The secondary structure assignments used for these calculations were made using the program DSSP [19]. As the number of states assigned by this program is more than three, we transformed the DSSP output to a three-state assignment using the criteria given by [20].

Acknowledgements

We greatly appreciate and thank Joe Cammisa for his invaluable help with the maintenance of the computer systems and particularly for making many of the figures presented in this article. We also thank members of the laboratory, Jean-Pierre Kocher, George Vasmatzis, and Yanli Wang, for helpful discussions.

References

1. Jernigan, R.L. (1992). Protein folds. *Curr. Opin. Struct. Biol.* 2, 248–256.
2. Chan, H.C. & Dill, K.A. (1993). The protein folding problem. *Phys. Today* Feb, 24–32.
3. Levitt, M. (1991). Protein folding. *Curr. Opin. Struct. Biol.* 2, 224–229.
4. Kolinski, A. & Skolnick, J. (1994). Monte Carlo simulations of protein folding. I. Lattice model and interaction scheme. *Proteins* 18, 338–352.

5. Rooman, M.J., Kocher, J.-P. & Wodak, S.J. (1991). Prediction of protein backbone conformation based on seven structure assignments. *J. Mol. Biol.* 221, 961–979.
6. Dandekar, T. & Argos, P. (1994). Folding the mainchain of small proteins with the genetic algorithm. *J. Mol. Biol.* 236, 844–861.
7. Dandekar, T. & Argos, P. (1996). Identifying the tertiary fold of small proteins with different topologies from sequence and secondary structure using the genetic algorithm and extended criteria specific for strand regions. *J. Mol. Biol.* 256, 645–660.
8. Park, B.H. & Levitt, M. (1995). The complexity and accuracy of discrete state models of protein structure. *J. Mol. Biol.* 249, 493–507.
9. Oldfield, T.J. & Hubbard, R.E. (1994). Analysis of C_α geometry in protein structures. *Proteins* 18, 324–337.
10. DeWitte, R.S. & Shakhnovich, E.I. (1994). Pseudodihedrals: simplified protein backbone representation with knowledge-based energy. *Protein Sci.* 3, 1570–1581.
11. Flocco, M.M. & Mowbray, S.L. (1995). C_α-based torsion angles: a simple tool to analyze protein conformational changes. *Protein Sci.* 4, 2118–2122.
12. Levitt, M. (1976). A simplified representation of protein conformations for rapid simulation of protein folding. *J. Mol. Biol.* 104, 59–107.
13. Hutchinson, G.E. & Thornton, J.M. (1994). A revised set of potentials for β-turn formation in proteins. *Protein Sci.* 3, 2207–2216.
14. Orengo, C.A., Flores, T.P., Taylor, W.R. & Thornton, J.M. (1993). Identifying and classifying protein fold families. *Protein Eng.* 6, 485–500.
15. Heringa, J., Sommerfeldt, H., Higgins, D. & Argos, P. (1992). OBSTRUCT: a program to obtain largest cliques from a protein sequence set according to structural resolution and sequence similarity. *Computer App. Biosci.* 6, 599–600.
16. Bernstein, F.C., *et al.*, & Tasumi, M. (1977). The protein data bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* 112, 535–542.
17. Kabsch, W. (1976). A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. A* 32, 922–923.
18. Kabsch, W. (1978). A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. A* 34, 827–828.
19. Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22, 2577–2637.
20. Rost, B. & Sander, C. (1993). Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.* 232, 584–599.

Because *Folding & Design* operates a 'Continuous Publication System' for Research Papers, this paper has been published via the internet before being printed. The paper can be accessed from http://biomednet.com/cbiology/fad.htm – for further information, see the explanation on the contents pages.