J. Mol. Biol. (1995) 249, 493-507

JMB

The Complexity and Accuracy of Discrete State Models of Protein Structure

Britt H. Park and Michael Levitt*

Department of Structural Biology, Stanford University Medical School, Stanford CA 94305, USA

**Corresponding author*

The prediction of protein structure depends on the quality of the models used. In this paper, we examine the relationship between the complexity and accuracy of representation of various models of protein a-carbon backbone structure. First, we develop an efficient algorithm for the near optimal fitting of arbitrary lattice and off-lattice models of polypeptide chains to their true X-ray structures. Using this, we show that the relationship between the complexity of a model, taken as the number of possible conformational states per residue, and the simplest measure of accuracy the root-mean-square deviation from the X-ray structure, is approximately (Accuracy) $\propto$ (Complexity)$^{-1/2}$. This relationship is insensitive to the particularities of individual models, i.e. lattice and off-lattice models of the same complexity tend to have similar average root-mean-square deviations, and this also implies that improvements in model accuracy with increasing complexity are very small. However, other measures of model accuracy such as the preservation of X-ray residue-residue contacts and the cc-helix, do distinguish among models.

In addition, we show that low complexity models, which take into account the uneven distribution of residue conformations in real proteins, can represent X-ray structures as accurately as more complex models, which do not: a selected 6-state model can represent protein structures almost as accurately (1.7 A root-mean-square) as a 17-state lattice model (1.6 Å root-mean-square).

Finally we use a novel optimization procedure to generate eight 4-state models, which fit native proteins to an average of 2.4 A, and preserve 85% of native residue-residue contacts. We discuss the implications of these findings for protein fading and the prediction of protein conformation.

Keywords: protein; discrete model; accuracy; complexity

Introduction

Two major difficulties plague the search for methods to predict the conformation of proteins *ab initio:* the exponentially large number of possible protein conformations, and the lack of effective criteria for differentiating correct from incorrect folds. The usual approach for dealing with the large number of possible protein conformations has been: first, to simplify protein structure by modeling proteins as chains of one or two interacting centers representing individual amino acids; and second, to allow these simplified models to adopt only a small discrete number of conformational states. For the most part, these simplified, discrete models are lattice models. These range from a simple tetrahedral model representing every other residue, with $3^{(n/2)-3}$ possible conformations for an n residue protein (Hinds & Levitt, 1992, 1994), through a tetrahedral lattice presenting all residues (Skolnick & Kolinski, 1989, 1990), a "knight's walk" lattice (Rey & Skolnick, 1991) with 23^{n-3} possible conformations, an extended face-centered cubic lattice with 41^{n-3} possible conformations (Cove11 & Jernigan, 1990; Covell, 1992), to an extended knight's walk model with 55^{n-3} possible conformations (Skolnick *et al.*, 1993). There are also several instances in the literature of the use of a 6-state off-lattice model with 6^{n-3} conformations (Rooman *et al.*, 1991, 1992; Rooman & Wodak, 1992; Dandekar & Argos, 1994; this model also has a 7th state reserved for cis-peptides).

Underlying all of these studies are the geometric characteristics of these various models which make

Abbreviations used: c.r.m.s., coordinate root-mean-square deviation; d.r.m.s., distance root-mean-square deviation.

0022–2836/95/220493–15 $08.00/0

them more or less useful for protein structure prediction, regardless of the actual search and discrimination strategies used. Two competing criteria determine how "good" a particular model is in this sense. First, a model must be as simple as possible. All other factors being equal, a model of complexity 2, i.e. one which has 2^{n-3} possible conformations for a protein of length n, is superior to one of complexity 10, which has $10^{n-3}/2^{n-3} = 5^{n-3}$ times more conformations. The smaller the conformational space to be searched, the more likely one is to find a "correct" conformation. Second, a model must represent the actual geometry of protein conformations accurately Models of low complexity tend to have a lower accuracy than models of high complexity

To date, there has been no complete systematic study of these issues of model complexity and accuracy Various aspects of model properties have been explored in the literature. Gregoret & Cohen (1991) looked at the compactness characteristics of a cubic lattice in comparison to real proteins and a more realistic non-lattice representation of protein conformations. Cove11 & Jernigan (1990) pointed out that superiority of a face-centered cubic lattice to a simple cubic lattice, in that a face-centered cubic lattice has pseudo-bond angles and pseudo-torsion angles more closely matching those found in real proteins. Godzik et al. (1993) studied the quality of fit and preservation of the secondary structure of various lattice models. What is absent from the literature is a complete analysis of the purely geometric properties of both lattice and off-lattice models.

In this paper, we report a study of many different protein models, both on and off lattice, in order to evaluate their relative utility for the prediction of protein structure. In particular, we use a novel method to fit any model to the X-ray coordinates of a large database of proteins. This method, which is over 1000-fold more efficient than previous methods (Rooman et al., 1991), enables us to analyze the relationship between complexity and accuracy, measured in *both a global (close fit to X-rav coordinates) and local (preservation of native secondary structure) sense. The data obtained from this exhaustive analysis provide upper bounds on what degree of accuracy can be expected from attempts to -fold proteins- with particular models.

We note also that models which do take into accou nt the uneven distribution of residue conformations in real proteins have significant advantages over those that do not. We find, for example, that 6-state models can be almost as accurate as 18- or 32-state models, and argue that low complexity off-lattice models are more appropriate for protein structure prediction than higher complexity lattice models. To this end, we present a set of optimized 4-state models which can reproduce protein backbone structure accurately in both the global and local senses.

Finally we examine all of these models in the context of protein structure prediction, by pointing out that the model chosen to represent a polypeptide structure has a strong effect on the prediction of its conformation, independent of the search strategy and discrimination function used. These effects may place fundamental restrictions on our ability to predict protein structure.

Results

The build-up method

The build-up method we describe in Materials and Methods is very efficient, but does not guarantee that the globally optimal fit will be found. In order to test the quality of fits obtained, we used short peptides, the globally best fit of which could be found. We randomly selected one hundred 12-residue segments from our database of X-ray conformations. For each of these, we exhaustively enumerated all their 4^{10} possible conformations using one of the optimized 4-state (ϕ, ψ) models (model C), and found the globally optimal fit. Then, for each of the segments, we performed our usual build-up procedure. With N_{keep} = 200,100 or 50, the build-up method found the global optimum for 99 of the 100 peptide segments. Even with N_{keep} = 10, 88 out of 100 build-up fits were globally optimal.

For structures much larger than the 12 amino acids considered here, there are too many possible conformations for exact enumeration, and we have no way of knowing how close we are to the exact optimum. To investigate the characteristics of the build-up algorithm for large peptides, we fitted a 4-state model (model C) to each of the proteins in our database while varying N_{keep}. Figure 1 illustrates the relationship between the number of conformations kept at each residue addition and (coordinate root-mean-square) $(\langle$ c.r.m.s$\rangle)$, the average c.r.m.s. fit. It is plain that one gains very little extra accuracy by keeping more than N_{keep} = 100 conformations per round of build-up. Thus, we decided that using 200 saved conformations per build-up round is completely adequate for the present study

Using N_{keep} = 200, we can also show that the algorithm is insensitive to the length of the protein fitted. Figure 2 shows the final c.r.m.s. fit of a 4-state optimized model (model C) to each of the proteins in our database as a function of the length of the proteins. Although there is considerable variation from protein to protein, the general trend is plain. Up to a length of about 100 residues, c.r.m.s. deviations rise, but bevond that thev remain constant or even decline slightly This trend is typical of the behavior for other models. In retrospect, the domain structure of proteins made this result predictable. Even large proteins are composed of several 100- to 200-residue domains. Since the build-up algorithm can fit 100- to 200-residue proteins well, it can also fit larger proteins composed of independent domains of similar size.

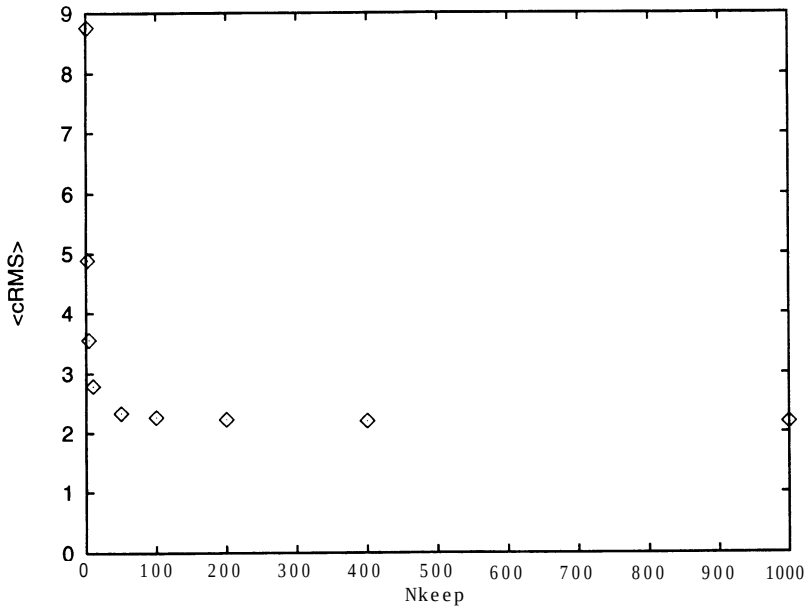

Figure 1. This shows the dependence of (c.r.m.s.), the length-weighted average coordinate r.m.s. deviation of all our database proteins, with N_{keep}, the number of partially built conformations saved at each build-up round. When N_{keep} is unity the effect is similar, but not identical, to choosing the closest state to each local τ angle. Such a strategy is remarkably ineffective, giving (c.r.m.s.) = 8.8 Å. Notice that fitting accuracy reaches a plateau at (c.r.m.s.) = 2.2 A when N_{keep} is greater than 100. For all figures, cRMS denotes c.r.m.s. and Nkeep denotes N_{keep}.

Accuracy assessment of lattice and "naive" off-lattice models

Having established a robust algorithm that can fit arbitrary models to X-ray structures, we proceeded to use it to characterize a large number of discrete state protein models. We considered both lattice and off-lattice models of varying complexity For each model, we assessed the quality with which it could represent each member of our database of proteins. Each of the measures of fit quality are reported in Table 1 as sequence length weighted averages over the entire database of proteins.

c.r.m.s. deviations

Figure 3(a) shows the relationship between a model's complexity and (c.r.m.s.), averaged over all of the proteins in our database. Rather surprisingly the curve is quite smooth. Further, the relationship seems to depend only on the complexity of a model and not on its precise nature. For example, the simple cubic lattice (α = 90°, τ = –90°, 0°, 90°, 180°; a = 180°), with its complexity of 5, generates average c.r.m.s. fits quite similar to those of the unrelated off-lattice model, with α = 120° and τ = 108°, –36°, 36°, 108°, 180°, which also has a complexity of 5. The results for the more complex lattice models may seem

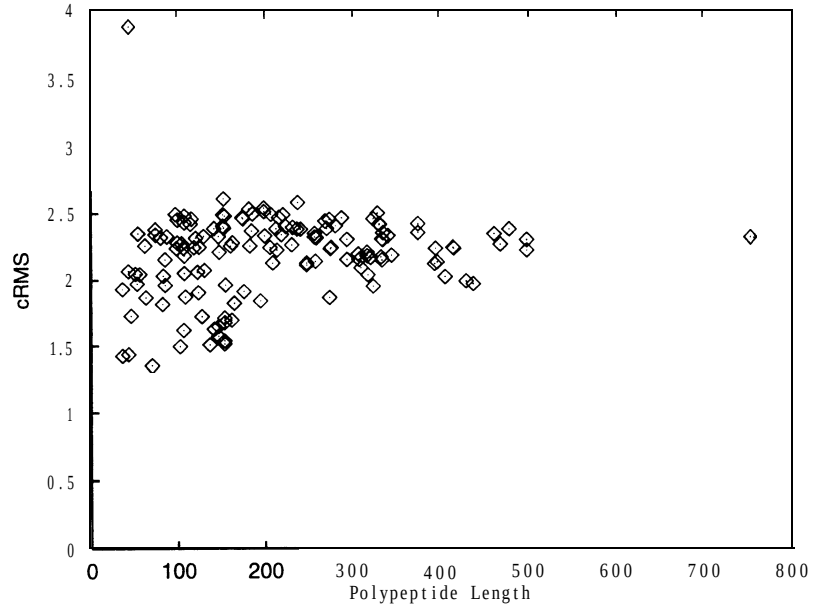

Figure 2. The c.r.m.s. fit of each protein in our database is plotted as a function of polypeptide length for the optimized 4-state model C. The c.r.m.s. deviation rises until the length is above 100 residues, after which it declines slightly The behavior of this particular model is typical of all the others.

Table 1

Deviations of different models fom native protein structures

Model	(τ, a) or (ϕ, ψ) values	Number states	c.r.m.s. (Å)	d.r.m.s. (Å)	% Native Contacts	% Native α	% Native β
Models of increasing complexity							
Tetrahedral/2	α = 109.5, τ = 60,60,180[a]	$\sqrt{3}$	5.39	4.66	60	NA[h]	NA
Tetrahedral	α = 109.5, τ = –60,60,180	3	3.63	2.99	78	68	80
Half cubic	α = 90, τ = –90,0,90,180; α = 180"	$\sqrt{5}$	4.81	4.18	81	NA	NA
Simple-4 A	α = 120, τ = –90,0,90,180	4	3.07	2.51	75	0	79
Simple-4 B	α = 120, τ = –135,–45,45,135	4	3.02	2.46	72	68	73
(ϕ, ψ) 4-state	(ϕ, ψ) = (–90,–90),(90,–90),(–90,90),(90,90)	4	3.22	2.67	77	29	89
Cubic	α = 90, τ = –90,0,90,180; α = 180	5	2.84	2.34	78	0	59
Simple-5 A	α = 120, τ = –144,–72,0,72,144	5	2.37	1.94	82	0	76
Simple-5 B	α = 120, τ = –108,–36,36,108,180	5	3.01	2.47	77	62	68
Simple-6 A	α = 120, τ = –120,–60,0,60,120,180	6	2.69	2.21	76	56	75
BCC	Body-centered cubic lattice[b]	7	2.59	2.14	87	50	74
FCC	Face-centered cubic lattice[b]	11	1.78	1.46	88	66	82
S+FCC	Simple + face-centered cubic[c]	17	1.60	1.31	88	68	80
1 B-state	α = (mean), τ = 160,–140, . . . 0,20, . . . 180[d]	18	1.24	1.02	92	90	88
KW	Knight's walk lattice (2,1,0)[e]	23	1.24	1.02	93	67	89
36-state	a(mean), τ = –170,–160, . . . 0,10, . . . 180[d]	36	0.97	0.80	94	91	91
XFCC	Extended face-centered cubic lattice[f]	41	1.15	0.94	89	79	83
XKW	Extended knight's walk'	55	0.90	0.73	96	79	88
Optimized models							
A	(ϕ, ψ) = (–64,–40),(–123,134),(111,–46),(117,105)	4	2.43	1.99	85	86	80
B	(ϕ, ψ) = (–66,–40),(–119,114),(–36,124),(132,–40)	4	2.31	1.89	86	86	74
C	(ϕ, ψ) = (–63,–63),(–132,115),(–42,–41),(–44,127)	4	2.22	1.82	86	75	75
D	(ϕ, ψ) = (–58,–31),(–127,126),(–97,–24),(109,108)	4	2.28	1.87	85	91	72
E	(ϕ, ψ) = (–71,–57),(–131,122),(–42,–36),(107,–25)	4	2.52	2.08	85	75	72
F	(ϕ, ψ) = (–58,–51),(–133,135),(–33,174),(114,–40)	4	2.42	1.97	84	68	76
G	(ϕ, ψ) = (–56,–48),(–129,128),(–108,35),(–31,–109)	4	2.48	2.03	84	61	76
H	(ϕ, ψ) = (–74,–31),(–131,125),(–101,179),(105,–40)	4	2.37	1.94	85	94	69
6-state	(ϕ, ψ) = (–57,–47),(–139,135),(–119,113), (-49,-26),(-1 06,48),(–1 01,–1 27)	6	1.90	1.55	87	71	80
Rooman et *al.*[g]	(ϕ, ψ) = (–65,–42),(–123,139),(–70,138), (–87,–47),(77,22),(107,–174)	6	1.74	1.42	89	80	81
Vary number of kept chains							
N_{keep} = 1 (for model C)		4	8.76	6.87	56	10	22
N_{keep} = 2		4	4.89	3.96	69	27	37
N_{keep} = 4		4	3.56	2.90	76	44	40
N_{keep} = 10		4	2.79	2.28	82	58	38
N_{keep} = 50		4	2.33	1.90	85	67	41
N_{keep} = 100		4	2.26	1.85	86	66	40
N_{keep} = 400		4	2.19	1.79	86	67	42
N_{keep} = 1000		4	2.17	1.77	86	66	42

[a] On these lattices, every other residue is fitted, with lattice spacing d = 4.95 Å (Hinds & Levitt, 1992).

[b] The BCC lattice allows single-step movements which are the vectors (± 1, ± 1, ± 1). The lattice dimension is 2.19 Å. The FCC lattice allows moves which are all cyclic permutations of the vectors (± 1, ± 1,0). d = 2.69 Å.

[c] The S + FCC lattice is a combination of a simple cubic and a face-centered cubic lattice, allowing moves which are permutations of the vectors (± 2,0,0) and (± 1, ± 1,1); d = 2.69 Å.

[d] These lattices use a values at each τ value which are the mean values for that τ value over our database of proteins.

[e] The underlying lattices are cubic. The knight's walk lattice allows moves which are cyclic permutations of the vectors (± 2, ± 1,0). The extended knight's walk lattice allows additional (± 1, ± 1, ± 1) moves (Skolnick *et al.*, 1993). The underlying lattice dimension for both models is 1.7 Å.

[f] This model allows moves which are permutations of the vectors (± 2,0,0), (± 2, ± 1, ± 1), and (± 1, ± 1,0) (Cove11 & Jernigan, 1990).

[g] Rooman et *al.* (1991).

[h] NA, not applicable.

to belie this. They all have (c.r.m.s.) fits worse than comparably complex off-lattice models. For example, the S + FCC lattice has an average c.r.m.s. fit of 1.60 Å (for 17 states), compared to 1.28 Å for an off-lattice 1 8-state model. It is, however, an exception which proves a rule. In practice, each step of a walk on the S + FCC lattice has approximately 12 sterically acceptable moves. From Figure 3(a) we would expect an average c.r.m.s. deviation of about 1.6 Å for a model with an effective complexity of about 12. Similarly, the other complex lattices have lower effective complexities.

Another surprising feature of Figure 3(a) is the diminishing returns in representational accuracy with progressively more complex models. For instance, a 36-state model produces fits only 0.2 Å better than an M-state model. By plotting the data from Figure 3(a) on a log-log scale (Figure 3(b)),

we find that the relationship is almost (c.r.m.s.) = $k(\text{Complexity})^{-1/2}$. In retrospect, such a simple dependence of c.r.m.s. on complexity can be quite easily explained, as follows.

Assume that residues 1 to $i-1$ of a protein have been fit perfectly by a model of complexity m. What is the average distance from the fit position of residue i to its actual position? Residue i lies somewhere on a sphere of radius b, the fixed bond length, centered on atom $i-1$. The m possible fit positions for atom i are, we will assume, evenly distributed on the surface of this sphere, which has a surface area $4\pi b^2$. On average, the surface area per state will be $4\pi b^2/m$, and the separation of states along the surface of the sphere will scale as $\sqrt{(4\pi b^2/m)}$. This separation is proportional to the accuracy (c.r.m.s. deviation) with

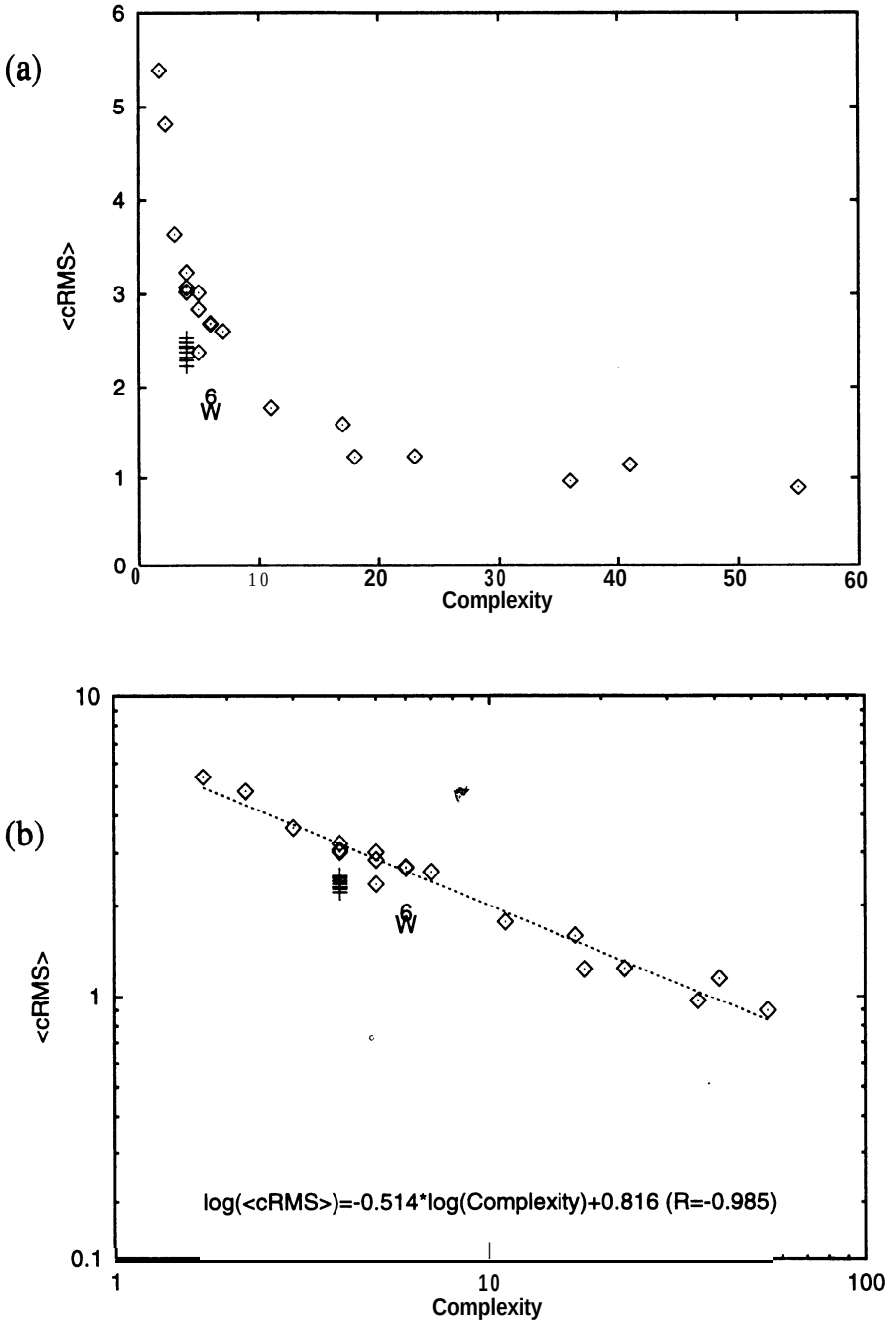

Figure 3. (a) Values of (c.r.m.s.), the sequence length weighted average c.r.m.s. deviations for best fits to all proteins in our database, are plotted as a function of model complexity The diamonds are all the naive models from Table 1. The crosses correspond to our optimized 4-state models, and 6 and W to our and Rooman's selected 6-state models (Rooman et *al.*, 1991). Beyond a certain point, added complexity improves accuracy very little. Optimized models show marked improvement over unoptimized models of the same complexity (b) Values of (c.r.m.s.) are plotted against complexity on a log-log scale. A linear least-squares fit gives (c.r.m.s.) = 6.59 $(\text{complexity})^{-0.514}$, very close to (c.r.m.s.) = $k(\text{complexity})^{-1/2}$ predicted by a simple analysis (see the text).

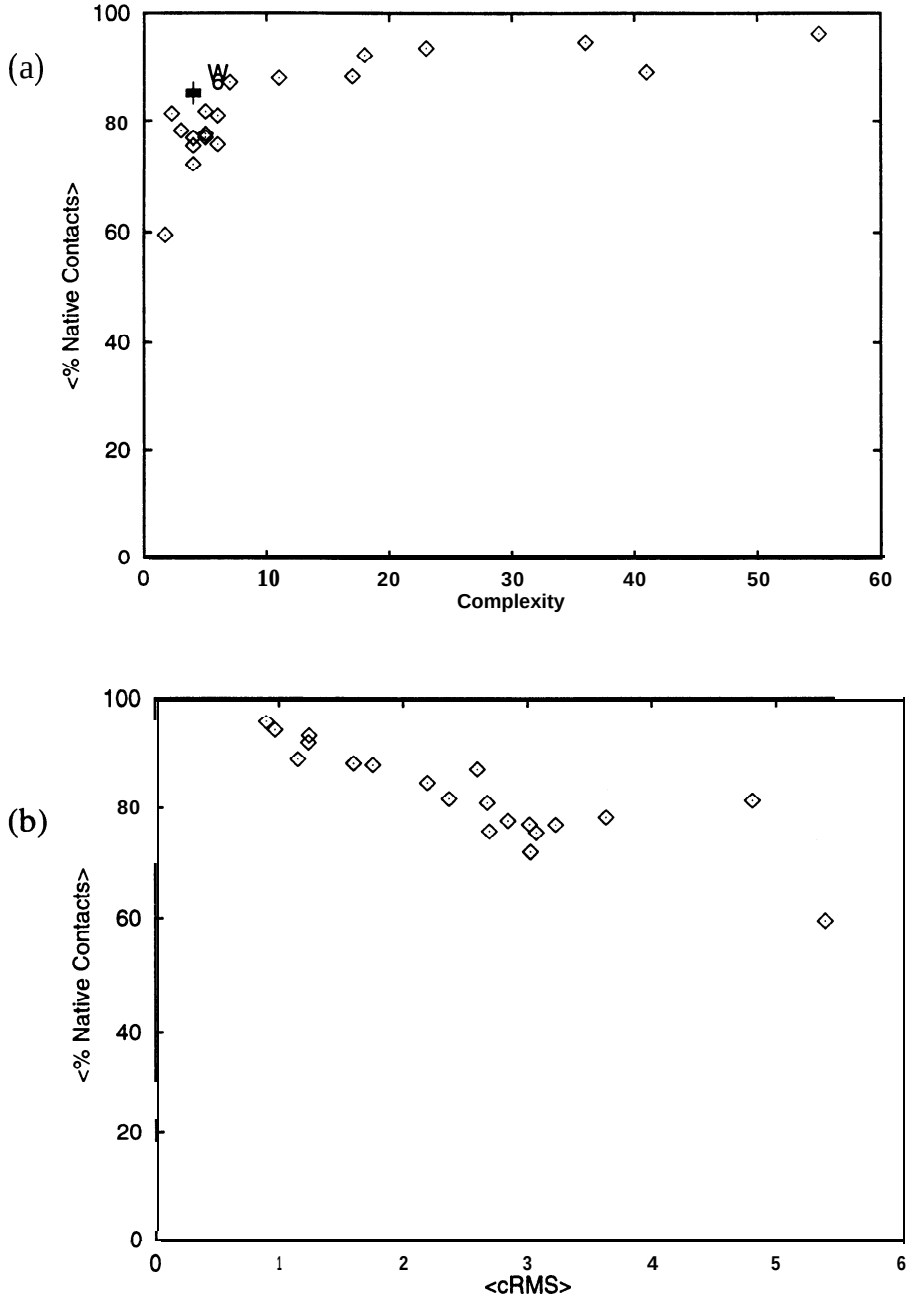

Figure 4. (a) The average proportion of X-ray contacts preserved by model fits, ⟨% Native Contacts), is plotted as a function of model complexity The variation in (% Native Contacts) is not as pronounced as that for (c.r.m.s.), but a similar trend is apparent. Optimized or selected models (shown as crosses and "6" and "W") are once again better than naive models of the same complexity (b) The relationship between preservation of native contacts and (c.r.m.s.). As expected, they are closely related.

which the position of residue i can be fit, so that c.r.m.s. should depend on $\sqrt{(1/m)}$ or $m^{-1/2}$.

Preservation of native contacts

So far, we have only looked at c.r.m.s. fit as a measure of model accuracy We also require the pattern of inter-residue contacts to be well preserved. Most energy functions used to predict protein conformations depend on empirical estimates of the relative likelihood of particular residue-residue contacts (for a recent review, see Wodak & Rooman, 1993). A model whose best representation of a native protein conformation only poorly reproduces the native pattern of inter-residue contacts can only generate relatively poor predictions.

Figure 4(a) shows the relationship between the proportion of preserved native contacts and model complexity First of all, we note that, as with the (c.r.m.s.) *versus* Complexity (Figure 3), little improvement in model performance is gained by models of complexity greater than 18. Second, the

spread of preservation values is rather narrow, from 60% to 95%, with most being greater than 70%. Another way of looking at this is shown in Figure 4(b), where the percentage of native contacts is plotted as a function of (c.r.m.s.) deviation. Structures as far as 3.0 Å (c.r.m.s.) from native folds still preserve more than 75% of native contacts. Third, the variation in contact preservations for models of similar complexity is quite high. Notably, the 3-state tetrahedral lattice preserves a greater than expected 78% of native contacts, the same as the considerably more complex 5-state simple cubic lattice. The cubic lattice which fits every other residue also preserves a larger proportion (81.3%) of native contacts than expected. This, however, is almost certainly an artifact of lattice scaling. For this model, we used a lattice dimensions of 4.95 A, which gives reasonably good c.r.m.s. fits, but also gives structures with an uncharacteristically high number of contacts per residue.

Preservation of secondary structure

Our third criterion for a "good" model of protein structure is that it preserve a large proportion of native secondary structure. Protein structure is hierarchical and, in general, dominated by well-defined secondary structure. Native conformations of proteins can be thought of essentially as arrangements of a-helix and P-sheet. Therefore, a good model should be capable of reproducing the secondary structure of native protein conformations.

Figure 5 shows that these characteristics, particularly the preservation of α-helix, vary much less regularly with model complexity than c.r.m.s. deviation or the percentage of preserved native contacts. In general, more complex models preserve more secondary structure, but there are large variations, particularly among the simpler models. Because of their simpler structural nature, (they are straight lines to a first approximation), p-strands are more easily reproduced. Godzik *et al. (1993)* have also noted the relative ease with which lattice models preserve p-structure.

It is in the preservation of α-helix that the deficiencies of complex lattice models are most apparent (Figure 5(a)). The body-centered cubic, face-centered cubic, knight's walk, simple-plus face-centered, extended face-centered cubic and extended knight's walk lattices all reproduce an abnormally low proportion of α-helix. In comparison, 4-state, 5-state, and 6-state models can each preserve over 60% of a-helix. Whether a particular 4 or 5-state model preserves a-helix depends on the model having a τ value in the helical region (τ = 22.9" to 71.6"). Thus the model Simple-4B, which has $\tau = 45°$, performs better than the model Simple-4A, which has τ = 0" and 90°.

Optimization of models

Since many low-complexity models are able to reproduce the important characteristics of protein structure, native contacts, secondary structure and c.r.m.s. deviation, we attempted to find the "best" 4 and 6-state models. Most of the models we have discussed in the previous section were naive in that, even when they were off-lattice models, they made no use of the added flexibility of not being on a lattice (for example, Simple-4A). In order to reap the benefits of stepping off-lattice, a model must take advantage of the non-uniform distribution of angles and torsion angles in real proteins. We first followed Rooman *et al*. (1991), by using a 6-state model in which each of the (ϕ, ψ) states were representative of the conformational states found in real proteins. The actual (ϕ, ψ) values for our sets and Rooman's are shown in Table 1. It is clear that both 6-state models are considerable improvements, with c.r.m.s. deviations of 1.74 Å and 190 Å, respectively as opposed to the 2.7 Å expected from naive 6-state models.

Models with six states are still rather complex, in that a chain of ten residues will have 6^{10} (60 million) conformations. A 4-state model, on the other hand, would only have a 4^{10} (1 million) conformations. Encouraged by the ease with which we could generate-a superior 6-state model of protein structure by simply selecting a set of six "reasonable" (ϕ, ψ) states, we undertook a more rigorous optimization of 4-state models, using the procedure described in Materials and Methods. From the initial enumeration of 325 4-state models in stage 1 of the optimization, we chose the eight best to be brought through stages 2 and 3. (We designate these as models A through H.) Each of the eight best models at this point could fit the small test proteins to within an average of 2.1 A. Random optimization improved fits to an average of 2.0 A. Minimization gave a final average of 1.9 A. Table 2 shows the actual (ϕ, ψ) values and the course of optimization. The final average c.r.m.s. fit for the eight optimized models over our 149 protein database was 2.38 A, a large improvement over the 3.0 to 3.2 A for naive 4-state models (Table 1 and Figure 3).

The other characteristics of the eight optimized 4-state models are also impressive. They uniformly manage to preserve 85% of native contacts, as opposed to 75% for naive 4-state models (Figure 4(a)). Both the optimized 4-state models and the selected 6-state models preserve native contacts much better than naive models. This is encouraging, as the 4-state models were optimized to minimize (c.r.m.s.) deviation, not maximize contact preservation. The proportion of native x-helix preserved is significantly improved for six of the models; model H, in particular, preserves 94% of a-helix. Native P-strand preservation, however, is not improved (Figure 5).

This concomitant improvement in average c.r.m.s. fit and preservation of the a-helix is not accidental. Improvement in a-helix fitting is, in fact, a good way to improve c.r.m.s. fitting. The explanation for this hinges on the interaction of local fitting and global fitting. The build-up algorithm makes no distinction between these two processes; it simply saves the N_{keep} best conformations each time it adds a new residue to a growing polypeptide. Depending on the model

and the residue position in the protein being fitted, the build-up algorithm may save conformations which are globally similar but locally diverse, or locally similar and globally diverse. One can therefore imagine that a cleverly designed model might be able to take advantage of the build-up algorithm's flexibility to mimic the hierarchical structure of proteins. For instance, if a model were able to accurately fit secondary structure with only a small number of states, it could force the build-up algorithm to save more globally diverse, but locally uniform, conformations when fitting particular stretches of secondary structure. This smart model would then be able to generate better global fits, because it could explore more of the conformational space.

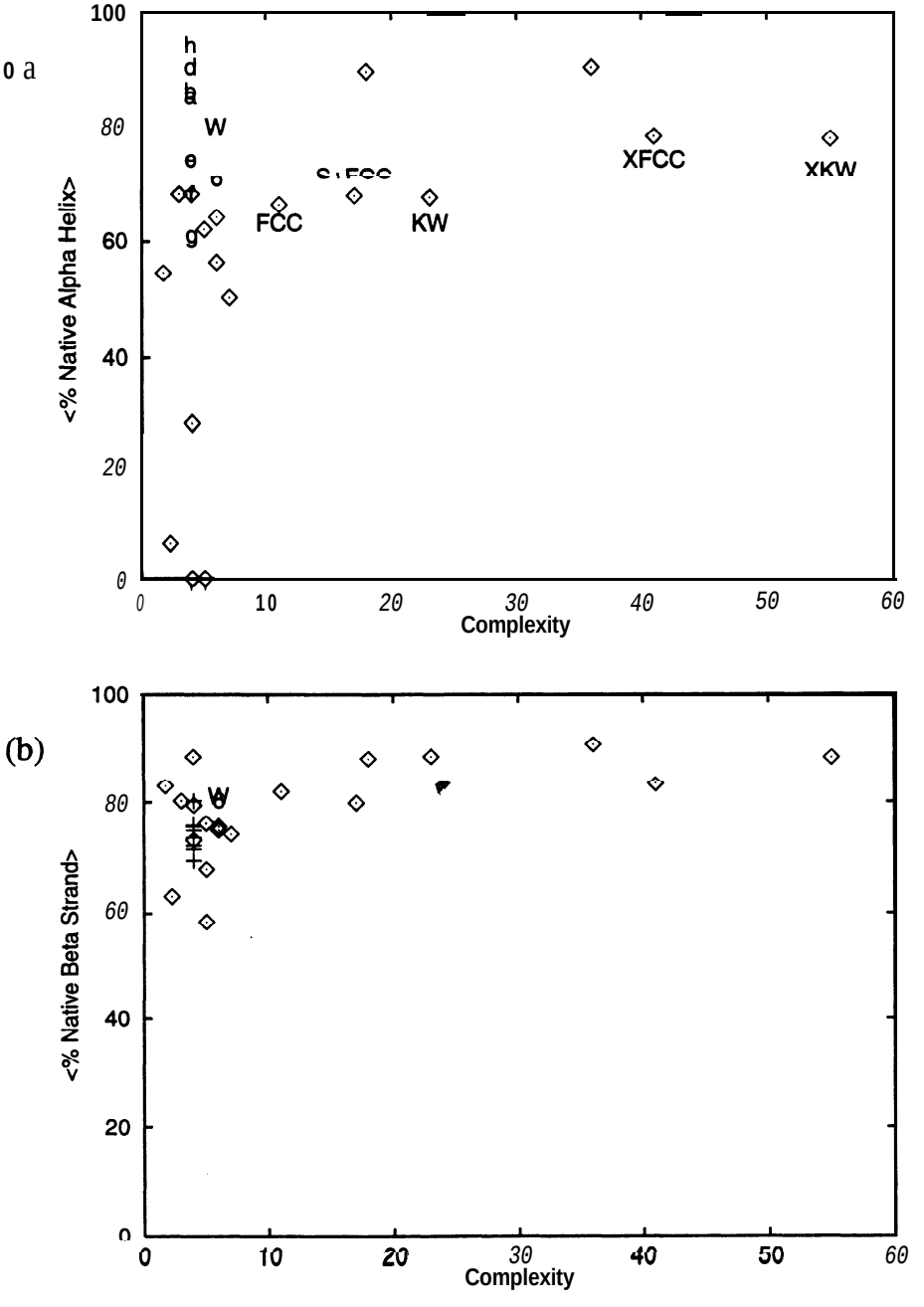

Figure 5. The average percentages of (a) X-ray a-helical and (b) X-ray P-strand structure preserved are plotted as a function of model complexity The relationships here are much less regular than for (c.r.m.s.) or (% Native Contacts). This is not entirely surprising, since the naive models know nothing of secondary structure. It is purely chance if a model happens to form τ angles within the range characteristic of a-helices. Most interesting, however, is the fact that complex lattice models are bad at preserving a-helices, and that optimization usually improves the preservation of the X-ray a-helix, relative to naive models, while it does little for the preservation of the P-strand. Optimized models are shown as crosses (b) or as the letters a through h (a). Our 6-state models and that of Rooman et *al.* (1991) are marked by 6 and W.

Table 2

Optimization of (ϕ, ψ) states in different 4-state models

Model	Set[a]	Enumerated[b] (ϕ, ψ)	Enumerated[b] Score[e]	Random[c] (ϕ, ψ)	Random[c] Score	Minimized[d] (ϕ, ψ)	Minimized[d] Score
A	α	**(-57,-47)**	**2.068**	(–62,–42)	**2.043**	(–64,–40)	**1.963**
	β	**(-129,124)**		**(-129,129)**		**(-123,134)**	
	γ	(108,–36)		(108,–36)		(111,–46)	
	δ	**(108,108)**		**(113,103)**		**(117,105)**	
B	α	(–57,–47)	**2.077**	(–67,–42)	**1.924**	(–66,–40)	**1.783**
	β	**(-129,124)**		**(-124,119)**		**(-119,114)**	
	γ	(–36,1OS)		**(-31,118)**		**(-36,124)**	
	δ	(108,–36)		(133,–41)		(132,–40)	
C	α	(–57,–47)	**2.086**	(–67,–67)	**1.932**	(–63,–63)	**1.847**
	β	**(-129,124)**		**(-139,114)**		**(-132,115)**	
	γ	(–108,–36)		(–46,–46)		(–42,–41)	
	δ	**(108,108)**		**(-16,118)**		(–44,127)	
D	a	(–57,–47)	**2.093**	(–57,–32)	**1.949**	(–58,–31)	1.881
	β	**(-129,124)**		**(-129,124)**		**(-127,126)**	
	γ	(–108,–36)		(–98,–31)		(–97,–24)	
	δ	**(108,108)**		**(108,108)**		**(109,108)**	
E	a	(–57,–47)	**2.094**	**(-72,-57)**	**1.995**	(–71,–57)	**1.922**
	β	**(-129,124)**		(–129,10)		**(-131,122)**	
	γ	(–36,–36)		(–41,–36)		(–42,–36)	
	δ	(108,–36)		(108,–26)		(107,–25)	
F	a	(–57,–47)	**2.096**	(–57,–52)	**2.047**	(–58,–51)	**1.996**
	β	**(-129,124)**		**(-134,124)**		**(-133,135)**	
		(–36,–1 80)		**(-31,175)**		**(-33,174)**	
	δ	(108,–36)		(113,–41)		(114,–40)	
G	a	(–57,–47)	**2.098**	(–57,–47)	**2.098**	(–56,–48)	**2.037**
	β	**(-129,124)**		**(-129,124)**		**(-129,128)**	
	γ	(–108,36)		(–108,36,)		(–108,35)	
	δ	(–36,–1 08)		(–36,–1 08)		(–31,–109)	
H	α	(–57,–47)	**2.102**	**(-72,-32)**	**2.028**	(–74,–31)	1.943
	β	**(-129,124)**		**(-134,124)**		**(-131,125)**	
	γ	(–108,–1 80)		**(-103,175)**		**(-101,179)**	
	δ	(108,–36)		(103,–36)		(105,–40)	

[a] **The (ϕ, ψ) states are labeled *a*, β, γ, 6, wherep and β are a-helix and P-sheet, respectively**
[b] **The (ϕ, ψ) values and scores of the eight best enumerated 4-state models.**
[c] **The (ϕ, ψ) values and scores after random optimization.**
[d] **The (ϕ, ψ) values and scores after Nelder-Mead simplex minimization (Press *et al.*, 1988).**
[e] **The scores are the length weighted averages of the best-fit c.r.m.s. deviations of the models to each of the eight small test Proteins.**

Our optimization procedure seems to have found versions of this smart model. The optimized models preserve an unusually large proportion of a-helix, indicating that they might indeed be forcing the build-up algorithm to save globally diverse but locally uniform populations of conformations, and thus be improving global fits. This hypothesis becomes certain when we examine the diversity of residue conformations during the build-up process. Indeed, for the "helix improved" models, the build-up algorithm saves fewer locally different conformations and more globally different conformations when building helices than when building non-helical regions. The question remains, however, of why optimization chooses a-helix fitting to improve, and not P-strand fitting. The answer comes from the different characteristics of the two types of secondary structure. Any a-helix is more or less identical to any other a-helix, i.e. there is little structural diversity among them. P-Strands, on the other hand, are often dissimilar among themselves; as a class, they have much more flexibility than a-helices. It is possible, therefore, for a low-complexity model to fit the a-helix well locally without sacrificing too many of its possible conformational states. To fit P-strands well locally would require a model to sacrifice too many states.

Figure 6 makes the improved characteristics of these selected and optimized models graphically clear. It illustrates, first, that a selected 6-state model is almost as good as an M-state naive model, especially for representing helical proteins, and second, that an optimized 4-state model drastically improves the accuracy of representation of helical proteins, compared to a naive 4-state model.

Discussion

The relationship between complexity and accuracy

We have found the accuracy of fit of lattice and naive off-lattice models to X-ray structures to follow a simple law:

$$(\text{c.r.m.s.}) \propto (\text{Complexity})^{-1/2}$$

This indicates that increasing model complexity above a certain point, yields little improvement in accuracy Other measures of accuracy however, can distinguish between models with similar average c.r.m.s. fits. For instance, when using a simple lattice representation of protein structure, a tetrahedral lattice is preferable to a cubic lattice. In addition to having a complexity 3/5 of that of the cubic lattice, the tetrahedral lattice preserves X-ray contacts slightly better with a 78.2% preserved, against 77.6% for the cubic. The average c.r.m.s. fit for the cubic lattice is better, at 2.84 A against 3.63 A. Nevertheless, for most prediction strategies, which rely on residue-residue contacts, the value for X-ray contacts is likelv to be more important. Thus, even though the cubic lattice is capable of representing protein structures more accurately than the tetrahedral, there are likely to be many conformations on the cubic lattice which have a high percentage of X-ray contacts correct, but are inaccurate in a c.r.m.s. sense. The reason for the better than expected preservation of X-ray contacts by the tetrahedral lattice is not clear. One explanation is that the 109.5" pseudo-bond angle of the tetrahedral lattice allows a more natural representation of certain protein structural features, in particular, P-strands. Indeed, the tetrahedral lattice preserves an average of 80.2% of X-ray β structure. It seems likely that the geometry of a tetrahedral lattice allows not only actual X-ray P-strands to be preserved, but also strand-strand contacts.

Another result of these studies is that optimized off-lattice models can, for the same complexity represent X-ray protein conformations much more accurately than lattice models. For example, any one of our optimized 4-state models is considerably more accurate than the (5-state) cubic lattice. A rationally selected set of &states, either ours or those described by Rooman *et* al. (1991), is as good as a naive 18-state model. Clearly any attempt to predict protein

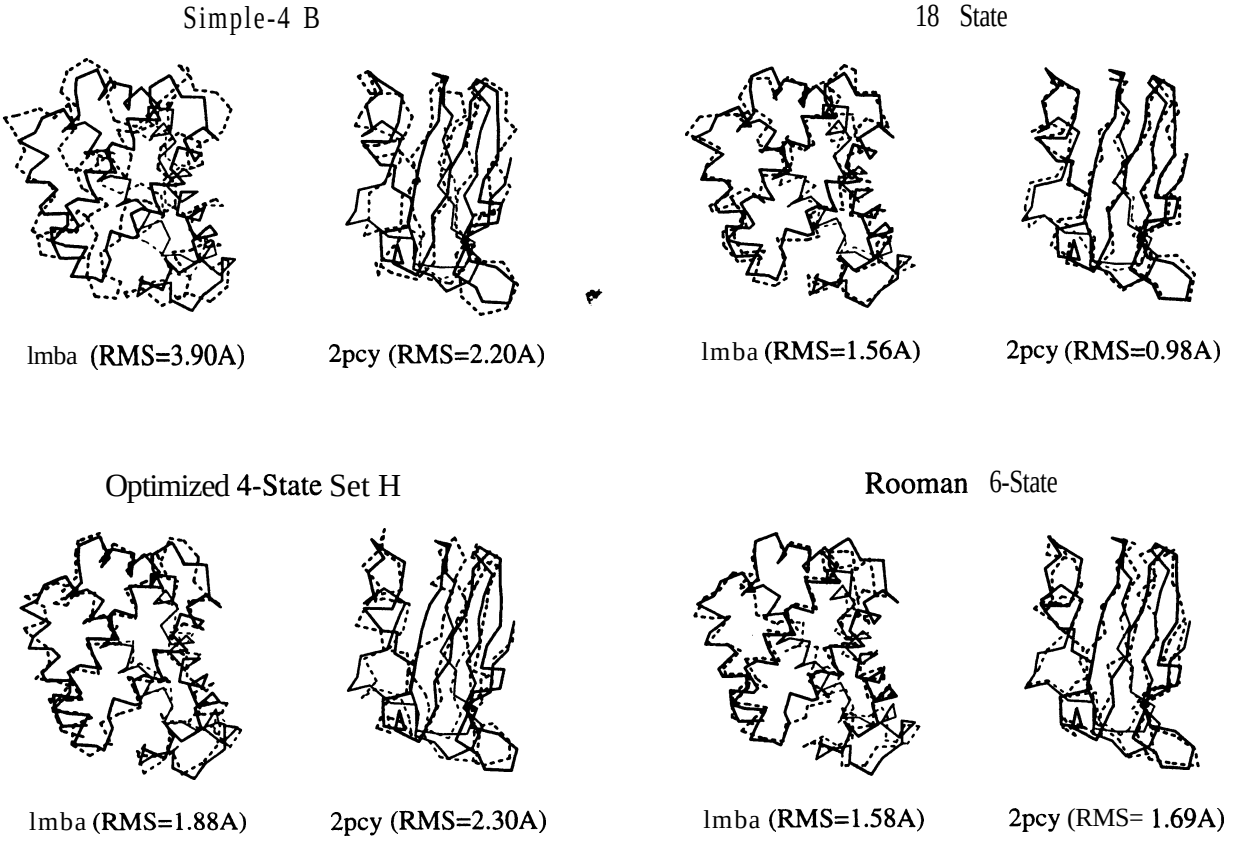

Figure 6. This shows fitted models (broken lines) superimposed over the X-ray conformations (continuous lines) of myoglobin (1 mba) and plastocyanin (1 pcy). Note the difference between a naive 4-state model (upper left) and an optimized 4-state model (lower left). The improvement in fit for the all a-protein, 1 mba, is remarkable. There is no improvement in fit for the all-p protein, 1 pcy. On the right-hand side, we compare the fits of an M-state model and a "hand" optimized 6-state model (Rooman et *al.*, 1991). **1** mba is fitted almost as well by the simple model as by the complex one, whereas plastocyanin is fit significantly better by the complex model. However, for many purposes, the poorer fit of the 6-state model is still adequate.

structure which uses some form of enumeration should consider the many advantages afforded by these simple off-lattice models.

For other kinds of non-enumerative searching procedures, the loss in continuity of phase space for models of very low complexity may make these models less suitable. Simulated annealing, for example, is only useful for finding optima or fairly smooth functions; when a model makes large jumps in phase space, a smooth potential function no longer behaves smoothly and one can expect Monte Carlo methods, like simulated annealing, to suffer. If this expectation is born out in actual practice, the problem of continuity can be effectively handled by selecting an off-lattice model which gives the requisite smoothness while being minimally complex.* Figure 3 shows plainly that there is little need to resort to exceedingly complex models. An 18-state model, for instance, will almost certainly have the necessary continuity in phase space, and be almost as accurate as a 36-state model.

In fairness to lattice models, we must note certain inherent advantages which they have relative to off-lattice models. First, the computational load for a lattice model of complexity m is much lower than that for an off-lattice model of the same complexity Almost all the geometric book-keeping for lattice models can be performed with integer arithmetic. Second, excluded volume effects are handled automatically for many lattice models, and for others with little computational cost. These advantages are significant,- and have been used successfully Because of the extremely efficient computational characteristics of the tetrahedral lattice, Hinds & Levitt (1992, 1994) were able to exhaustively enumerate all compact conformations of several small proteins, and find native-like structures among the energetically best 100 or so. Skolnick et al. (1993), using their extremely accurate 55-state lattice model and a Monte Carlo algorithm, have consistently generated native-like conformations for 4-helix bundles. Covell & Jernigan (1990) used a face-centered cubic lattice to enumerate all possible conformations for several small proteins on a lattice bounded by the known protein shapes, and found that native-like conformations were always among the energetically best 1 or 2%. All of these studies were feasible because of the computationally tractable nature of their respective lattice models.

However, our results make it clear that off-lattice models can be computationally efficient by virtue of their low complexity and high accuracy Indeed, the literature already supports this contention. Rooman *et* al. (1991) and Dandekar & Argos (1994) have already used a selected 6-state model to predict early folding segments in proteins and conformations of 4-helical bundles, respectively We are currently using our optimized 4-state models to evaluate the ability of various empirical energy functions to discriminate native from non-native folds.

Implications for structure prediction

This study has also shown that for protein structure prediction, the inherent characteristics of the models used cannot be ignored. The folding problem consists of three parts. (1) A model of protein structure; (2) a conformational search method; (3) a method for discriminating native from non-native conformations. In many studies of protein structure prediction, it is difficult to tell which part of the investigation is at fault when predictions are less than perfect. The model, the search method or the discriminating function may be at fault. Studies in which some conformational space has been exhaustively enumerated (Hinds & Levitt, 1992, 1994; Covell & Jernigan, 1990) have been very useful because they have eliminated one of the structure prediction components, namely the search strategy so that all shortcomings are attributable either to the model or the energy function used for discrimination.

This current study looks at the first part of the protein folding problem, the models used, in isolation, and shows that certain of their characteristics may fundamentally limit their use for the prediction of protein structure. Each model is limited in its possible representational accuracy, and thus limited in its prediction accuracy Beyond this obvious observation there is also a more subtle limitation of structure prediction. Every discrete-state model is limited by its ability to preserve native residue-residue contacts. The simple cubic lattice, for instance, has an average best fit c.r.m.s. of 2.84 A; that best fit, on average, only preserves 77.6% of native contacts. We expect, therefore, that most energy functions (which are usually based on relative residue-residue contact frequencies) will select many false positives which have poorer c.r.m.s. deviations from the correct structure, but which preserve a larger proportion of native contacts. For geometric reasons, an energy function will not be able to discriminate native from non-native conformations. Since accuracy in a c.r.m.s. sense, generally varies closely with a model's ability to preserve native contacts, there is an uncertainty principle for protein structure prediction: the lower the accuracy of the model used to make the prediction, the more ambiguous predictions will be, regardless of the discriminating function or search method used.

Implications for protein folding

Complexity and resolution are intimately associated with protein folding, in that the folding time depends on the number of conformations that have to be searched. A protein chain can change its conformation at no more than the thermal speed of its constituent atoms (approximately 3 Å/ps, or the speed of sound in air). If different conformational states are further apart than 1 A, a chain could examine no more than 10^{13} conformations per second. Small proteins with fewer than 100 residues take about 10^{-3} seconds to fold, and in this time some 10^{10}

conformations could be examined. On the other hand, if there are m states per residue, a chain of length n has m^{n-3} possible conformations. For modest values of m (5 is often used), a chain of 100 residues will have an astronomical number of possible conformation (5^{97} or 10^{68}). This argument was used almost 30 years ago by Levinthal (1968) to pose the famous "Levinthal Paradox": a protein has too many conformations to fold by a random search of conformational space, and must fold by a directed pathway With the results of the present study we have a quantitative relationship between m, the complexity or number of states per residue, and the resolution of the resulting conformations (how close the best one would be to an actual X-ray structure).

For five states, the resolution or c.r.m.s. value is 2.7 A, which is much too low to precisely define a native conformation. To define main-chain atoms to a resolution of 0.5 A, the accuracy of a high-resolution X-ray structure, requires some 150 states per residue (the 100 residue chain would have 10^{218} possible conformations). Even polypeptide chains as short as six residues would have too many states to fold by exhaustive search ($150^6 = 10^{13}$).

One can also ask what resolution model could be exhaustively searched in the fastest folding times of 10^{-3} seconds. This model could only have 1.32 states per residue (or 4 residues for each 3-state site), as $1.32^{100} = 10^{12}$. The expected resolution of such a model is very low, with c.r.m.s. = 10 A. Molten globules with resolutions of about 5 Å would have at least 10^{23} possible conformations (two residues for three states, or $\sqrt{3}$ states per residue). Clearly folding to a compact low-resolution intermediate structure like the molten globule intermediate, also has to follow a directed pathway rather than a random search of all conformations. This suggests that folding of long polypeptide chains, like those found in proteins, must proceed via a hierarchy of pathways in which there are a succession of directed pathways at different levels of resolution. The organization of this hierarchy is, of course, unknown. The suggestion has been made that certain segments of nascent polypeptides have a high propensity for local structure (p-turns, a-helix), and that these fold rapidly and help "guide" the polypeptide to its final three-dimensional conformation (Wetlaufer, 1973). Another hypothesis is that certain key hydrophobic residues rapidly form clusters, which then guide the rest of the folding process (Lesk & Rose, 1981; Bashford et al., 1988). Our results show the necessity of this hierarchical ordering of protein folding, whatever its precise nature.

Future applications

Our optimized 4-state models provide low-complexity models that fit proteins well, yet have a very small number of different conformations. These models, and other models like them, will be useful for protein structure prediction, loop fitting, exhaustive enumeration of peptide conformations, and low-resolution structure determination by NMR or X-ray crystallography (with fewer possible conformations, less experimental data are required).

Table *3*

Database of polypeptides

Protein Data Bank four-letter name followed by chain name[a]						
labp	1fc2:c	1prc:h	2cab	2pab: a	3b5c	4sgb:i
1acx	1fdx	lprcl	2ccy: a	2pcy	3blm	4tln
lbds	1fx1	1prcm	2cdv	2pka:a	3ca2	4tmn:e
1bmv:l	lgrc	1pyp	2cna	2pka:b	3dfr	4tsl :a
1bmv:2	lhip	1rbb:a	2cpp	2r06:3	3fxc	5cpa
lbp2	1hoe	lrei	2cyp	2rsp:a	3gap	5cpv
lcc5	1112	lrhd	2dhf:a	2sbt	3gap: a	5ebx
lcd4	llhl	1rmu:1	2fd2	2sga	3gpd	51dh
1 cho:i	1lh4	lrnt	2gbp	2sns	3gs	5mbn
1cla	1lz1	lsgt	2gd1:o	2sod	3hmg:a	5tnc
1cms	lmba	1 teci	2gls:a	2sod:b	3hmg:b	5xia:a
1 coh:b	lmbd	1tim	2gn5	2ssi	3icd	6acn
1crn	lovo	1tnf:a	2hhb:a	2stv	3pgk	7cat:a
1csc	1p09:a	lwrp	2hhb:b	2taa	3pgm	8adh
1cse	lpaz	1 wrp:r	2hla:a	2taa:a	4ait	8api:a
1cts	1pcy	1 wsy:b	2hla:b	2tbv	4ape	8api:b
lcy3	1pfk	256b:a	2ilb	2tbv:a	4dfr:a	8cat
1eca	1pfk:a	2aat	2kai:b	2tmv:p	4er4:e	9pap
lest	lphh	2act	21iv	2utg:a	4hvp:a	1fc1:a
1f19:h	1pp2	2alp	21zm	2ypi:a	4mdh:a	1prc:c
1f19:l	1ppt	2atl:b	2mev:1	351c	4sbv	2aza
2mev:3	3adk					

[a] Bernstein et al. (1977).

Materials and Methods

Here, we first list the sets of proteins we used in this study Then we present methodological details of our particular implementation of discrete state models, emphasizing their generation and fitting to test set proteins, and our criteria for assessing structural accuracy Finally we present a heuristic method for optimizing the selection of the (ϕ, ψ) values which characterize 4-state models.

Database of well-refined protein structures

Throughout this study we used a database of 149 peptide chains. Their Protein Data Bank designations are shown in Table 3. In order to simplify handling, the list, which is based on Hobohm et *al.* (1992), was modified to exclude proteins which had missing residues. The database contains polypeptides from 36 to 753 residues in length. Essentially, all structural motifs are represented.

Test set of proteins

For part of this study, in particular the optimization of (ϕ, ψ) parameters of 4-state models, we used eight small proteins: lctf, lr69, 1sn3, 1ubq, 2cro, 3icb, 4pti, and 4rxn. We used this set of small proteins (each less than 80 residues in length) to make our computationally intensive optimization procedure tractable. In other ways, they are more or less typical of all proteins.

Discrete state models

In this study we worked with discrete state models of protein chains. By discrete state, we mean that the internal coordinates (bond lengths, bond angles, and torsion angles)

can only assume certain particular values. For simplicity we only consider the cc-carbons. Conventional representations use (ϕ, ψ) coordinates, defined, as usual, by the torsion angles formed by atoms $[C_{i-1}, N_i, C_i^\alpha, C_i]$ and $[N_i, C_i^\alpha, C_i, N_i]$. In models using only C^α atoms, as in this study it is more convenient to use a different pair of angles (α, τ). α_i is defined as the pseudo-bond angle formed by $[C_{i-1}^\alpha, C_i^\alpha, C_{i+1}^\alpha]$ and τ_i is defined as the pseudo-torsion angle formed by $[C_{i-1}^\alpha, C_i^\alpha, C_{i+2}^\alpha, C_{i+2}^\alpha]$ (pseudo because these atoms are not connected by chemical bonds). Our nomenclature for α and τ is opposite to that used in earlier studies (Levitt, 1976), but is more mnemonic, as α is an "angle" and τ is a "torsion".

Using (α, τ) angles or (ϕ, ψ) angles converted to (cc, τ) angles (see below), we consider different discrete state models of varying "complexity", by which we mean the number of possible conformations per residue. More precisely, a protein of n amino acid residues, represented by a model of complexity m, will have m^{n-3} possible conformations for an (α, τ) based model and m^{n-2} conformations for a (ϕ, ψ) based model.

Building Cartesian coordinates from (ϕ, ψ) and (a, τ) coordinates

Internal coordinates a and τ are the easiest from which to generate Cartesian coordinates, but local states of amino acid residues (e.g. α-helix, P-strand) are more conveniently defined in terms of (ϕ, ψ) angles. For an amino acid with standard bond lengths and angles, there is a strict dependence of (a, τ) on (ϕ, ψ): angle α_i is determined by the ϕ and ψ angles of residue i; torsion angle τ_i is determined from the ϕ and ψ angles of both residues i and $i + 1$. While the a angles depends only on the (ϕ, ψ) angles of a single residue, the τ angle depends on the (ϕ, ψ) angles of two adjacent residues. Thus, for a $(\phi, \$)$-based discrete state model of complexity m, there will be m possible a values and m^2 possible τ values. While approximate analytical relationships between the two coordinate systems have been described by Levitt (1976), we use exact relationships by first generating a tetrapeptide from specific (ϕ, ψ) torsion angles and standard bond lengths and angles (Schulz & Schirmer, 1979), and then calculating a and τ angles from the C^α coordinates of the peptide.

From such a set of angles and torsions, α_i and τ_i, we generate C^α Cartesian coordinates for residues 1 to n in the following way The Cartesian coordinates for the first three C^α atoms are calculated using:

$$x_1 = 0, \quad y_1 = 0, \quad z_1 = 0$$

$$x_2 = 3.8, \quad y_2 = 0, \quad z_2 = 0$$

$$x_3 = x2 + 3.8\cos(\pi - \alpha_1),$$

$$y_3 = y_2 + 3.8\sin(\pi - \alpha_i), \quad z_3 = 0 \qquad (1)$$

Each additional residue's coordinates, $\mathbf{r}_i = (x_i, y_i, z_i)$, are calculated from the coordinates of the preceding three residues. We first calculate the three orthogonal unit vectors u, v, and w:

$$\mathbf{u} = \frac{\mathbf{r}_{i-1} - \mathbf{r}_{i-2}}{|\mathbf{r}_{i-1} - \mathbf{r}_{i-2}|}$$

$$\mathbf{v} = \frac{(\mathbf{r}_{i-3} - \mathbf{r}_{i-2}) - [(\mathbf{r}_{i-3} - \mathbf{r}_{i-2})\cdot\mathbf{u}]\mathbf{u}}{|(\mathbf{r}_{i-3} - \mathbf{r}_{i-2}) - [(\mathbf{r}_{i-3} - \mathbf{r}_{i-2})\cdot\mathbf{u}]\mathbf{u}|}$$

$$\mathbf{w} = \mathbf{u} \times \mathbf{v} \qquad (2)$$

The vector u is the unit vector along the pseudo-bond between C_{i-2}^α and C_{i-1}^α. Vectors u and v together define the plane containing atoms C_{i-3}^α, C_{i-2}^α, and C_{i-1}^α. Vector w simply completes the right-handed coordinate system. The Cartesian coordinates for the next residue are then given by the simple relation:

$$\mathbf{r}_i = \mathbf{r}_{i-1} + 3.8\cos(\pi - \alpha_{i-1})\mathbf{u} + 3.8\sin(\pi - \alpha_{i-1})\cos(\tau_{i-2})\mathbf{v} + 3.8\sin(\pi - \alpha_{i-1})\sin(\tau_{i-2})\mathbf{w} \qquad (3)$$

where 3.8 Å is the standard α-carbon to a-carbon distance.

c.r.m.s. and d.r.m.s. deviations

There are two commonly used measures for the similarity of two sets of protein coordinates. The first, the coordinate root mean squared deviation (c.r.m.s.), is calculated by:

$$\text{c.r.m.s.} = \left(\frac{\sum_{i=1}^{n} |\mathbf{r}_{ai} - \mathbf{r}_{bi}|^2}{n} \right)^{1/2} \qquad (4)$$

where $\mathbf{r}_{ai}$ and $\mathbf{r}_{bi}$ are the positions of atom i of structure a and structure b, respectively, and where structures a and b have been optimally superimposed (Kabsch, 1978). The second measure, the distance root mean squared deviation (d.r.m.s.), is calculated:

$$\text{d.r.m.s.} = \left(\frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (|\mathbf{r}_{ai} - \mathbf{r}_{aj}| - |\mathbf{r}_{bi} - \mathbf{r}_{bj}|)^2}{(n(n-1))/2} \right)^{1/2} \qquad (5)$$

where $\mathbf{r}_{ai}$ and $\mathbf{r}_{bi}$ are defined as before. This calculation does not require the superposition of coordinates. For a particular pair of structures, the c.r.m.s. deviation, which measures the similarity of atomic positions, is usually larger than the d.r.m.s. deviation, which measures the similarity of interatomic distances.

Discrete state model fitting to X-ray structures

In order to generate accurate discrete state models of X-ray structures, we used a brute force build-up algorithm, reminiscent of a method used to find low-energy conformations of peptides (Vasquez & Scheraga, 1985, 1988). Starting at the N terminus of the target protein, we enumerated all the possible conformations for the first four residues, saving the N_{keep} conformations which were closest in conformation to the first four C^α atoms of the X-ray structure (lowest c.r.m.s. deviation). Then we added single residues to the C terminus of each of the N_{keep} saved conformations in all m possible states, and again saved the N_{keep} conformations with the lowest c.r.m.s. deviation from the appropriate portion of the X-ray structure. We repeated this iterative procedure until the entire protein had been fitted. The characteristics of the process are illustrated in two dimensions in Figure 7.

For the more complex lattice models, like the extended face centered cubic lattice, knight's walk, and extended knight's walk, the use of internal angles is needlessly complex, because the possible internal angles for each step depend on the previous steps taken. For these models we use Cartesian coordinates to build a chain directly

Finding the best fit of a model to an X-ray structure is computationally equivalent to enumerating all possible folded structures, i.e. the problem scales as m^n, where m is the number of possible states per residue and n is the

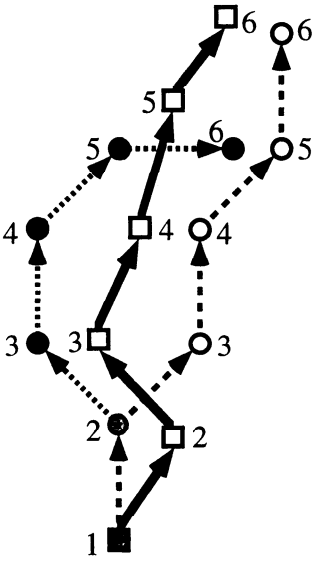

Figure 7. This illustrates the build-up algorithm in two dimensions. Consider the bold path with residues marked by squares to be the X-ray conformation. With the first three residues fitted, the left-hand path (marked with filled circles) is plainly better than the right (marked with open circles). (The stippled circles mark the overlap of the left- and right-hand paths.) However, by the time all six residues have been built, the right-hand path is a better fit. By saving an ensemble of "virtual" paths when adding each new residue, the build-up algorithm can find globally good fits which are composed of locally sub-optimal fits. Note that the model chain paths illustrated here are: (1) Discrete in that the chain can be continued in two ways (τ = 0 or 180° with α = 135°); (2) off-lattice, as evidenced by the uneven distribution of allowed positions.

number of residues. The best fit among these m^n possible folds is easily identified as having the minimum of c.r.m.s. deviation. Despite the well-defined nature of the problem, enumeration of all m^n possibilities is not computationally feasible. Our build-up procedure takes advantage of the reasonable expectation that a globally optimal fit of a particular model to an X-ray structure is likely to be composed of shorter segments which are locally optimal. In other words, if, for a segment of 20 residues, a near optimal set of states is ($S^1, S^2, \ldots S^{20}$), then any continuous subset of these states, e.g. (s", S^4, S^5, S^6, S^7), is likely to be near optimal for the corresponding section of X-ray structure.

By keeping a repertoire of good local fits at each stepwise residue addition, the build-up procedure is able to find good approximations of optimal fits at little computational cost. For example, with a loo-residue protein keeping 200 partially built conformations at each step, the build-up algorithm takes about one minute on a Silicon Graphics Indigo (MIPS R3000) for a 4-state model. The overall algorithmic complexity of the method can be derived as follows. During each round of build-up, c.r.m.s. deviations from the X-ray conformation (the dominant computational task) for mN_{keep} conformations are calculated. Each of these c.r.m.s. calculations takes time proportional to k, the number of residues already built. Therefore, the total running time for a protein of n residues is approximately proportional to $\Sigma_{k=1}^{n} mN_{keep}k$, or $(n^2 + n)mN_{keep}$, so that, asymptotically the running time scales as n^2mN_{keep}.

Assignment of secondary structure

Since the models we deal with in this study are composed of a-carbon backbones only standard methods for the assignment of secondary structure to protein conformations (Kabsch & Sander, 1983) are not applicable We devised simple and fairly robust criteria for assigning α-helix and P-strand secondary structure states to a-carbon coordinates. The a states are assigned for those residues for which the value of the pseudo-torsion angle, τ, is between 22.9° and 71.6'. More specifically if the τ angle formed by residues i, $i + 1$, $i + 2$, and $i + 3$ is between 22.9' and 71.6°, then residues $i + 1$ and $i + 2$ are considered to be α-helices. In contrast, β states are assigned in a somewhat less obvious way, not directly dependent on τ values. We define residue i to be in a β state if the angle between the vector from residues $i - 2$ to $i + 1$ and the vector from residues $i - 1$ to $i + 2$ is less than 30°.

This definition of β state has the desirable property of being insensitive to short kinks, which occur reasonably frequently in P-strands (especially in discrete models), while still identifying longer range changes in chain direction, such as at the termini of P-strands. For example, if one considers conformations in which α = 120° and τ = 0° or 180° (*cis* or *trans*), our criterion will identify a chain with conformation *trans-trans-cis-trans-bans* as a P-strand, but one with conformation *trans-trans-cis-cis-trans-trans* as two separate short P-strands. This fits very well the requirement that P-strands should be regions of extended conformation.

Residue-residue contacts

Two residues in a protein conformation are considered to be in contact if the distance between the C^{α} coordinates is less than 8.0 A. The proportion of native contacts preserved by a model is calculated over all distinct pairs of residues (i,j), where $i < j$.

Optimization of (ϕ, ψ) states

To generate an optimal set of (ϕ, ψ) states for 4-state models, we used the following procedure. (1) As a starting point, we chose two states (ϕ = 57°, ψ = -47") and (ϕ = –129°, ψ = 124°) which represent standard α-helical and P-strand conformations, respectively With these two states fixed, we then enumerated the possible values for the other two states at intervals of 72° in both ϕ and ψ, giving 325 different 4-state models (there are 25 possible individual (ϕ, ψ) states (5 ϕ values and 5 ψ values) and (25*25 + 25)/2 = 325 unique pairs of states). (2) The build-up procedure, defined above, is then used to generate coordinates of each of the small test proteins for each 4-state model. We calculated the weighted average of their c.r.m.s. deviations from their X-ray coordinates as a measure of how well a particular 4-state model fits real protein structures. (3) We optimized the eight models (A through H) with the best c.r.m.s. values further by making small random changes in each of the (ϕ, ψ) values. If the perturbed set produced a better overall fit of the test set of proteins to their X-ray coordinates, we accepted this new set as the next starting point. We repeated this until no further improvement was found. (4) Finally we optimized these states further using a Nelder-Mead simplex minimizer to reduce the average c.r.m.s. value (Press *et al.*, 1988).

Acknowledgements

This work was supported by National Institutes of Health Award GM30387 and National Science Foundation Award DMB8720208. One of us (B.H.P) holds a National Science Foundation fellowship from the Program in Mathematics and Biology We thank S. Subbiah and J. Tsai for their constructive criticism.

References

Bashford, D., Cohen, F. E., Karplus, M., Kuntz, I. D. & Weaver, D. L. (1988). Diffusion-collision model for the folding kinetics of myoglobin. *Proteins: Struct. Funct. Genet 4, 211-227.*

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). The protein data bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol.* 112, 535-542.

Covell, D. G. (1992). Folding protein a-carbon chains into compact forms by Monte Carlo methods. *Proteins: Struct. Genet. 14, 409-420.*

Covell, D. G. & Jernigan, R. L. (1990). Conformations of folded proteins in restricted spaces. *Biochemistry*, 29, 3287-3294.

Dandekar, T. & Argos, I? (1994). Folding the main chain of small proteins with the genetic algorithm. *J. Mol. Biol.* 236, 844-861.

Godzik, A., Kolinski, A. & Skolnick, J. (1993). Lattice representations of globular proteins: How good are they? *J. Comp. Chem.* 14, 1194-1202.

Gregoret, L. M. &Cohen, F. E. (1991). Protein folding: Effect of packing density on chain conformation. *J. Mol. Biol.* 219, 109-l 22.

Hinds, D. A. & Levitt, M. (1992). A lattice model for protein structure prediction at low resolution. *Proc. Natl Acad. Sci. USA*, 89, 2536-2540.

Hinds, D. A. & Levitt, M. (1994). Exploring conformational space with a simple lattice model for protein structure. *J. Mol. Biol.* 243, 668-682.

Hobohm, U., Scharf, M., Schneider, R. & Sander, C. (1992). Selection of representative protein data sets. *Protein Sci.* 1, 409-417.

Kabsch, W. (1978). A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallog. sect. A, 34, 827-828.*

Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers, 22, 2577-2637.*

Lesk, A. M. & Rose, G. D. (1981). Folding units in globular proteins. *Proc. Natl Acad. Sci. USA*, 78, 4304-4308.

Levinthal, C. (1968). Are there pathways for protein folding? *J. Chim. Phys.* 65, 44-45.

Levitt, M. (1976). A simplified representation of protein conformation for rapid simulation of protein folding. *J. Mol. Biol.* 104, 59–107.

Press, W. H., Flannery B. I?, Teukolsky, S. A. & Vetterling, W. T. (1988). *Numerical Recipes in C*, Cambridge University Press, Cambridge, U.K.

Rey, A. & Skolnick, J. (1991). Comparison of lattice Monte Carlo dynamics and Brownian dynamics folding pathways of α-helical hairpins. *Chem. Phys.* 158, 199-219.

Rooman, M. J. & Wodak, S. J. (1992). Extracting information on folding from the amino acid sequence: Consensus regions with preferred conformation in homologous proteins. *Biochemistry*, 31, 10239-10249.

Rooman, M. J., Kocher, J. A. & Wodak, S. J. (1991). Prediction of protein backbone conformation based on seven structure assignments. *J. Mol. Biol.* 221, 961-979.

Rooman, M. J., Kocher, J. P. & Wodak, S. J. (1992). Extracting information on folding from the amino acid sequence: Accurate predictions for protein regions with preferred conformation in the absence of tertiary interactions. *Biochemistry*, 31, 10226-10238.

Schulz, G. E. & Schirmer, R. H. (1979). *Principles of Protein Structure*, Springer-Verlag, N.Y.

Skolnick J. & Kolinski, A. (1989). Dynamic Monte Carlo simulations of globular protein folding/unfolding pathways. I. Six-member, Greek key P-barrel proteins. *J. Mol. Biol.* 212, 787-817.

Skolnick, J. & Kolinski, A. (1990). Dynamic Monte Carlo simulations of globular protein folding/unfolding pathways. II. Z-Helical motifs. *J. Mol. Biol.* 212, 819-836.

Skolnick, J., Kolinski, A., Brooks, C. L., Godzik, A. & Rey, A. (1993). A method for predicting protein structure from sequence. *Curr. Biol. 3*, 414423.

Vasquez, M. & Scheraga, H. A. (1985). Use of buildup and energy-minimization procedures to compute low-energy structures of the backbone of Enkaphilin. *Biopolymers*, 24, 1437-1447.

Vazquez, M. & Scheraga, H. A. (1988). Calculation of protein conformation by the build-up procedure. Application to bovine pancreatic trypsin inhibitor using limited simulated nuclear magnetic resonance data. *J. Biomol. Struct. Dynam. 5*, 705–755.

Wetlaufer, D. B. (1973). Nucleation, rapid folding, and globular intrachain regions in proteins. *Proc. Natl Acad. Sci. USA*, 70, 697-701.

Wodak, S. J. & Rooman, M. J. (1993). Generating and testing protein folds. *Curr. Opin. Struct. Biol. 3, 247-259.*

Edited by F. E. Cohen

(Received 6 January 1995; accepted 10 March 1995)