J. Mol. Biol. (1994) 243, 668–682

Exploring Conformational Space with a Simple Lattice Model for Protein Structure

David A. Hinds and Michael Levitt

Beckman Laboratories for Structural Biology
Department of Structural Biology, Stanford University School of Medicine, Stanford, CA 94305, U.S.A.

We present a low resolution lattice model for which we can exhaustively generate all possible compact backbone conformations for small proteins. Using simple structural and energetic criteria, for a variety of proteins, we can select for lattice structures that have significant similarities with their known native structures. Our energetic parameters are based on pairwise amino acid contact frequencies in a database of experimentally determined structures. A key step in our method involves the threading of a sequence onto every lattice model, such that a locally optimal pattern of tertiary interactions is formed. We evaluate our results against statistics collected for structures covering all of conformational space, and against statistics collected for permuted sequences. Despite the low resolution of the model, our low energy structures contain many native features. These results indicate that the overall pattern of hydrophobicity of a sequence significantly constrains the range of folds that sequence is likely to adopt.

Keywords: protein structure prediction; protein folding; lattice model; conformational search

1. Introduction

The prediction of protein structures from their amino acid sequences remains an unsolved problem in computational biochemistry. A proteins sequence is generally thought to be sufficient to determine its folded structure (Anfinsen *et al.*, 1961; Anfinsen, 1973). However, the vast diversity of potential structures accessible to even a small protein, and the presence of multiple local energetic minima, combine to make the native state an elusive target. Protein folding pathways represent nature's solution to the search problem (Levinthal, 1968; Wetlaufer, 1973). In principle, with sufficiently accurate potential functions and enough computer power, we could predict the folded conformations of proteins by direct simulation of the folding process. While this is not currently feasible, it would still be infinitely more efficient than systematically searching all possible structures for a global energy minimum. At lower resolution, the Choice between current directed and exhaustive search methods is less clear (Brower et al., **1993;** Abagyan, **1993).**

A common strategy for approaching the protein folding problem is to' use simplified representations and energy functions, in the hope that solutions to these simpler problems n-ill be close to the solution to the real problem (Levitt & Warshel, 1975; Levitt, **1976;** Kuntz *et al.*, **1976;** Wilson & Doniach, **1989).** Lattice representations are a computationally convenient way of limiting the conformational possibilities of a protein. Lattice models can be broadly classified into those that are more abstract, designed for addressing general questions about protein folding (Go, **1978;** Lau & Dill, 1989; Crippen, 1991), and those that can accurately represent native polypeptide backbones, designed for structure prediction (Covell & Jernigan, **1990;** Skolnick and Kolinski, **1990;** Skolnick et *al.*, **1993).** Our approach (Hinds & Levitt, **1992)** is somewhere in the middle: we have picked a simple low resolution model, so that we can perform exhaustive conformational searches and collect statistics about all conformational space, but we also try to make predictions using real sequences. Here we present a more thorough analysis and extension of this earlier work.

The lattice representation we use is designed to only capture the overall path of a polypeptide chain. The model cannot represent, secondary structural differences or side-chain orientations. For each lattice structure, we thread a sequence onto the chain path to form a locally optimal arrangement of tertiary interactions. Our structures are then scored based on their patterns of tertiary contacts, compared to contact; frequencies observed in a database of known structures. We evaluate our predictions by comparison with statistics collected over all conformational space.

For most protein sequences we have tried, simple structural and energetic screens can select for lattice structures that have significant) similarities with their known native conformations. We show that we can extract some tertiary structural information from a sequence, without being able to build detailed models.

0022–2836/94/440668–15 $08.00/0

Table 1

List of 246 representative protein structures

1aai:A	1aai:B	1aak:-	1abe:-	1abm:A	1ace:-	1acx:-	1ak3:A	lake:A	1ala:-
1alc:-	1ald:-	1 atn:B	1bab:A	1 bab:B	1bbh:A	1 bbk:A	1bbk:B	1 bbp:A	1bbt:1
1 bbt:2	1bbt:3	1 bia:-	1bmv:1	1bmv:2	1c2r:A	1ccr:-	1cd8:-	1cmc:A	1col:A
1cox:-	1dri:-	dwc:B	1eca:-	1end:-	1etu:-	1ezm:-	1f3g:-	1fc2:D	1 fcb:A
1fha:-	1fkb:-	1fnr:-	1gal:-	1gd1:O	1gky:-	1gly:-	1gmf:A	1gox:-	1gpl :A
1gpb:-	1gpr:-	1grc:A	1gst:A	1 hge:A	1hge:B	1 hrh:A	1ife:-	1igm:L	1ipd:-
1ith:A	1lap:-	1ldn:A	1lh2:-	1lig:-	1llc:-	1lld:A	1lpe:-	1lts:A	1lts:D
1lz1:-	1mba:-	1mcp:H	1mnr:-	1mpp:-	1nn2:-	1npx:-	1 nsb:A	1ofv:-	lova:A
1paz:-	1pbx:A	1pbx:B	1pgd:-	1phd:-	1phh:-	1pii:-	1pp2:R	1ppa:-	1ppf:E
1ppl:-	1ppn:-	1prc:C	1prc:H	1prc:L	1prc:M	1pyp:-	1r09:1	1r09:2	1r09:3
1rbp:-	1rcb:-	1rhd:-	lrnh:	1rro:-	1rve:A	1sas:-	1sdy:A	1sgt:-	1spa:-
1stp:-	1tfd:-	1tgi:-	1thm:-	1tie:-	1tim:A	1tlk:-	1tmd:-	1tnf:A	1ton:-
1trb:-	1ula:-	1vsg:A	1wsy: A	1wsy:B	1yea:-	256b:A	2act :-	2alp: ▪	2aza:A
2bb2:-	2ca2:-	2ccy:A	2cna: ▪	2cts:-	2cyp:-	2dnj :A	2er7:E	2fcr: ▪	2fx2:-
2gbp: ▪	2hbg:-	2lhb:-	2liv:-	21tn:A	2mcg: 1	2mcm:-	2mev: 1	2mev:2	2mev:3
2msb:A	2nn9:	2pab:A	2pf 1 : ▪	2pfk: A	2pka:B	2plv:1	2plv:2	2plv:3	2pmg:A
2por: ▪	2prk:-	2reb: ▪	2ren:-	2rsp:A	2scp:A	2sdh:A	2sga:-	2sns: ▪	2snv
2sod:O	2stv:-	2tbv:A	2tmv:P	2tpr:A	2trx:A	2tsl:-	2wrp:R	3adk:-	3apr:E
3blm: ▪	3c2c: ▪	3cd4:A	3chy:-	3cla: ▪	3dfr: ▪	3est:-	3fgf: ▪	3gap:A	3hla:A
3icd:	3lad:A	3pfk:-	3pgk: ▪	3pgm:-	3rp2:A	3rub:L	3rub:S	3sdp:A	3sic:E
3sic:1	4bp2:-	4cms:	4cpv:	4dfr: A	4enl: ▪	4fab:L	4fd1:-	4fxn:	4gcr: ▪
4gpd:1	4i1b:-	4lzm:-	4mdh:A	4pep: ▪	4ptp: ▪	4rcr:H	4rcr:L	4rcr:M	4sbv:A
4tgl:A	4tms:-	5cpa:	5cyt:R	5fbp:A	5rub:A	5tim:A	5tmn:E	5tnc: ▪	6ldh: ▪
6q21:A	6taa:-	6xia:-	7aat :A	7acn:-	7api:A	7cat :A	7tim:A	8adh:-	8atc:A
8atc:B	8dfr: ▪	8fab:B	8gch: ▪	9rnt :-	9wga: A				

Protein structures are identified by their 4-letter PDB codes, followed by a chain identifier, or ▪ if the chain identifier is blank.

The distribution of hydrophobic and hydrophilic residues along the sequence is a powerful constraint on a proteins folded conformation, supporting the central role of hydrophobicity in establishing an overall fold (Kauzmann, 1959; Dill. 1990). In its present form. our low resolution approach seems less useful as a practical means for predicting protein structure.

2. Theory and Methods

(a) Da tabase of representative structures

Our structure database is based on a list of representative structures from the Brookhaven Protein Data Bank (PDB†; Bernstein *et al.*, 1977) selected so that no two members have more than 60% sequence identity (Hobohm *et* al.. 1992). Two proteins with this level of sequence identity would be expected to have extremely similar structures. However, they could have at most 36% identical pairwise interactions. For our purposes, this cutoff seemed to be a good balance between maximizing our sample size and minimizing redundancy. The particular list we used contains 380 chains and 78.766 residues.

We filtered this raw list to select the best available structures where there were several alternatives in the PDB. For each entry in the original list, we identified likely relatives by searching for similar PDB headers, confirmed by sequence comparison. Structures were scored based on crystallographic resolution, presence of mutations and heteroatoms, and date of ent ry. We removed entries that had resolutions worse than 3.0 Å. or fewer than 100 residues. This revised list. shown in Table 1. contains 246 chains and 60,721 residues.

† Abbreviations used: PDB, Protein Data Bank: r.m.s., root mean square; R_g, radius of gyration.

(b) *Calculation of effective pairwise energies*

We base our pairwise residue-residue interaction energies on the frequencies of tertiary contacts in our database. We define a tertiarv contact between two residues wherever a non-hydrogen atom of one residue approaches within 4.5 Å of a non-hydrogen atom of the other residue, and the two residues are at least five sequence positions distant from one another. We do not differentiate between backbone and side-chain atoms. Using these criteria, our database contains 124.3 14 independent' tertiary contacts, corresponding to about four long-range contacts to each residue. Our energy calculations are similar to those used by Miyazawa & Jernigan (1985), or Bryant & Lawrence (1993). The effective energy e_{uv} of a contact between amino acid types u and v is given by:

$$e_{uv} = -kT \ln\left(\frac{\sum_p C_{uvp}}{\sum_p \frac{C_p}{T_p} T_{uvp}} \right), \qquad (1)$$

where p varies over all proteins in the data set, and $kT = 0.59213$ kcal/mole at 298 K. For each protein, C_p is the number of tertiary contacts, C_{uvp} is the number of U-V contacts, T_p is the total number of possible tertiary contacts. and T_{uvp} is the number of possible U-W contacts. T_p and T_{uvp} are given by:

$$T_p = \sum_{i=1}^{N_p} \sum_{j=1}^{N_p} \begin{cases} 1, & (|i-j| \geq 5) = (N_p - 4)(N_p - 5) \\ 0, & (|i-j| < 5) \end{cases} \qquad (2)$$

and

$$T_{uvp} = \sum_{i=1}^{N_p} \sum_{j=1}^{N_p} \begin{cases} 1, & (|i-j| \geq 5) \text{ and } (r_i = u) \text{ and } (r_j = v), \\ 0, & \text{otherwise} \end{cases} \qquad (3)$$

where N_p is the number of residues in protein p. and r_i is the amino acid type of residue i of protein p. T_p can also be thought of as the contact map area that is at least five

sequence positions from the diagonal, and T_{uvp} equals the part of that area consisting of $u - v$ residue pairings. Our e_{uv} values are shown in Table 2.

The cysteine–cysteine interaction energy calculated using equation (1) is biased because the contact totals include both covalent and non-covalent interactions. We would like this term to be the effective energy of just a non-covalent cysteine–cysteine interaction, because we handle covalent linkages separately. The contact counts cannot be easily corrected, because the vast majority of cysteines are covalently linked. We decided to estimate the Cys–Cys energy using the other energies in our table. We found that Cys X energies were most correlated with Met X energies ($r = 0.972$). We fit our Met X energies to the Cys X energies, and used the fit to predict Cys–Cys from Met-Cys. This yielded a value of -0.196 kcal/mole.

The energies calculated using equation (1) are more accurate than the energies we have used previously (Hinds & Levitt, 1992). Our new method correctly normalizes the values so that an energy of zero has a reasonable meaning: an interaction that occurs just as often as would be expected by chance, accounting for sequence composition. We characterized our energy parameters by calculating jackknife estimates (Wonnacott & Wonnacott, 1984) of their standard errors. The average standard error of the e_{uv} values was 0.03 kcal/mole, and 93% of the values had standard errors of 0.05 kcal/mole or less. The standard deviation between values is about 0.3 kcal/mole, so the uncertainties in these values are predicted to be on the order of 10% of the spread between them. These errors are about half as large as errors we calculated for our older table.

(c) *The tetrahedral lattice representation*

Our two criteria in deciding how to represent a protein were that our model should be sufficiently flexible to be able to represent the range of chain threadings seen in real proteins, and that it should be sufficiently restrictive to allow us to make exhaustive conformational searches. We chose to represent a polypeptide as a self-avoiding chain of vertices on a diamond lattice (Figure 1). The diamond lattice has the lowest possible density of connections. Lattices with denser connectivity can fit native structures with higher fidelity, at the expense of exponentially increasing the number of possible shapes that must be searched. We are only interested in distinguishing correctly folded structures from other compact globular shapes, so we also restrict all our lattice structures to fit within an ellipsoidal bounding volume. Finally, allowing one degree of freedom per residue leads to too many possible structures for long polypeptides, so we limit our lattice chain lengths so that there are two residues for every lattice vertex.

An important advantage of lattice models is their computational simplicity. All structures are composed from a small set of predetermined points in space, so self-avoidance is easily enforced, and calculation of many structural properties can be simplified using previously generated look-up tables. Using conventional workstations, it is feasible for us to generate and analyze on the order of 10^8 structures, though our searches more typically cover 10^7 or 10^7 possibilities. We are able to do exhaustive searches of compact, bounded lattices with up to approximately 40 vertices (Figure 2). These searches take on the order of a few hours on a fast workstation, and can easily be executed in parallel over several machines. The bounding volume size is the most important determinant of the search time; larger volumes result in exponentially longer searches, but as larger structures fill more and more of a volume, the boundary constraints become more severe and ultimately limit the diversity of allowed structures.

(d) *Aligning a sequence with a lattice structure*

The discriminatory power of our contact potential is determined by how well our lattice models can represent native tertiary contact patterns. Our lattice structures are defined by fewer vertices than there are residues, so we have some choice about how we map a sequence to a chain fold. Spacing all residues equally along the length of the chain is a poor choice, because the geometry of the lattice makes many arrangements of contacts impossible. In actual protein structures, the spacing of residues along the chain path varies from 1.5 Å per residue in α-helices to 3.4 Å on antiparallel β-sheet. Therefore, we decided that for each lattice structure, we would try to optimize the spacing of residues along the chain path, using the tertiary contact energy as a guide. Each lattice point is associated with a specific residue, and from zero to three residues are positioned between each pair of lattice points along the chain. A particular lattice structure of length N is defined by a list of vertices v_i for $i = 1, \ldots, N$, and an alignment of a sequence with that lattice structure is defined by a corresponding list of sequence positions m_i for $i = 1, \ldots, N$.

To calculate the energy of a lattice structure, we score each lattice contact using our pairwise energy matrix. When counting contacts that contribute to a structure's energy we include nearest and next-nearest neighbors on the lattice. Since there are more residues than lattice points, we average the interaction energies for residues adjacent to the residues actually assigned to the lattice points. The energy of a contact between two vertices mapped to sequence positions m and n is given by

$$E_{mn} = \frac{2e_{r_m r_n} + e_{r_{m-1} r_n} + e_{r_{m+1} r_n} + e_{r_m r_{n-1}} + e_{r_m r_{n+1}}}{6}. \qquad (4)$$

where r_m is the residue type of residue m in the sequence. When counting residue–residue contacts in a lattice structure, for a particular lattice contact, we treat $\langle m, n \rangle$ as a full contact and the other $\langle m \pm 1, n \pm 1 \rangle$ interactions as "half" contacts. Using this rule, the total numbers of long-range contacts in lattice and actual structures are roughly similar. With the averaging in equation (4), if there are no residues in the gap between two lattice vertices, some interactions will be counted twice, and if there are three residues in the gap, the middle residue will not contribute to the total energy at all.

Finding the assignment of residues to lattice points that has the global minimum contact energy is a difficult problem, so in keeping with our strategy of substituting easy problems for hard ones, we use an iterative procedure that quickly converges to a locally optimal arrangement, shown in Figure 3. We start by spacing residues as evenly as possible along the chain path. We then calculate the expected energy changes for shifting the residue assignment of each lattice point, using a window of $+/-2$ residues. Using dynamic programming, we determine the set of mutually compatible moves with lowest predicted final energy. The procedure is repeated until there are no favorable moves. The energies we predict for individual moves assume that all the other residues are fixed, so when several moves are made in one cycle, the energy of the new mapping may not be as predicted, and in fact may get worse. In practice, the method typically converges in three to five cycles, independent of chain length.

Table 2

Pairwise energy parameters, in kcal/mole

	ALA	CYS	ASP	GLU	PHE	GLY	HIS	ILE	LYS	LEU	MET	ASN	PRO	GLN	ARG	SER	THR	VAL	TRP	TYR
ALA	0.272	0.040	0.525	0.496	−0.167	0.422	0.137	−0.134	0.421	−0.118	−0.088	0.336	0.293	0.269	0.215	0.367	0.206	−0.104	−0.242	−0.112
CYS	0.040	−0.980	0.237	0.200	−0.445	0.211	−0.146	−0.287	0.170	−0.233	−0.339	0.141	0.008	−0.059	0.005	0.013	−0.031	−0.246	−0.460	−0.252
ASP	0.525	0.237	0.447	0.485	0.174	0.328	−0.062	0.271	0.022	0.310	0.249	0.205	0.331	0.260	−0.233	0.320	0.233	0.399	−0.091	−0.083
GLU	0.496	0.200	0.485	0.539	0.106	0.489	−0.043	0.187	−0.023	0.258	0.181	0.234	0.324	0.356	−0.188	0.301	0.231	0.250	−0.115	−0.096
PHE	−0.167	−0.445	0.174	0.106	−0.748	0.059	−0.318	−0.585	0.024	−0.613	−0.644	−0.005	−0.101	−0.062	−0.207	−0.002	−0.149	−0.502	−0.690	−0.545
GLY	0.422	0.211	0.328	0.489	0.059	0.387	0.222	0.152	0.421	0.155	0.121	0.311	0.265	0.304	0.135	0.324	0.271	0.171	−0.081	0.023
HIS	0.137	−0.146	−0.062	−0.043	−0.318	0.222	−0.396	−0.090	0.162	−0.088	−0.226	−0.014	−0.034	0.084	−0.160	0.025	−0.041	−0.015	−0.406	−0.365
ILE	−0.134	−0.287	0.271	0.187	−0.585	0.152	−0.090	−0.557	0.150	−0.559	−0.449	0.152	0.084	0.041	−0.104	0.104	−0.100	−0.496	−0.645	−0.488
LYS	0.421	0.170	0.022	−0.023	0.024	0.421	0.162	0.150	0.429	0.154	0.124	0.249	0.287	0.145	0.196	0.311	0.256	0.163	−0.144	−0.183
LEU	−0.118	−0.233	0.310	0.258	−0.613	0.155	−0.088	−0.559	0.154	−0.551	−0.432	0.190	0.067	0.033	−0.062	0.134	−0.024	−0.452	−0.598	−0.442
MET	−0.088	−0.339	0.249	0.181	−0.644	0.121	−0.226	−0.449	0.124	−0.432	−0.531	0.080	−0.069	0.019	−0.085	0.083	−0.132	−0.334	−0.718	−0.494
ASN	0.336	0.141	0.205	0.234	−0.005	0.311	−0.014	0.152	0.249	0.190	0.080	0.031	0.187	0.159	−0.002	0.244	0.170	0.220	−0.106	−0.105
PRO	0.293	0.008	0.331	0.324	−0.101	0.265	−0.034	0.084	0.287	0.067	−0.069	0.187	0.189	0.107	0.007	0.296	0.164	0.095	−0.358	−0.221
GLN	0.269	−0.059	0.260	0.356	−0.062	0.304	0.084	0.041	0.145	0.033	0.019	0.159	0.107	0.100	−0.036	0.230	0.091	0.019	−0.235	−0.176
ARG	0.215	0.005	−0.233	−0.188	−0.207	0.135	−0.160	−0.104	0.196	−0.062	−0.085	−0.002	0.007	−0.036	−0.116	0.032	−0.034	−0.017	−0.369	−0.372
SER	0.367	0.013	0.320	0.301	−0.002	0.324	0.025	0.104	0.311	0.134	0.083	0.244	0.296	0.230	0.032	0.244	0.185	0.154	−0.070	−0.061
THR	0.206	−0.031	0.233	0.231	−0.049	0.271	−0.041	−0.100	0.256	−0.024	−0.132	0.170	0.164	0.091	−0.034	0.185	0.054	−0.070	−0.195	−0.099
VAL	−0.104	−0.246	0.399	0.250	−0.502	0.171	−0.015	−0.496	0.163	−0.452	−0.334	0.220	0.095	0.019	−0.017	0.154	−0.070	−0.450	−0.508	−0.352
TRP	−0.242	−0.460	−0.091	−0.115	−0.090	−0.081	−0.406	−0.645	−0.144	−0.598	−0.718	−0.106	−0.358	−0.235	−0.369	−0.070	−0.195	−0.508	−0.679	−0.605
TYR	−0.112	−0.252	−0.083	−0.096	−0.545	0.023	−0.365	−0.448	−0.183	−0.442	−0.494	−0.105	−0.221	−0.176	−0.372	−0.061	−0.099	−0.352	−0.605	−0.475

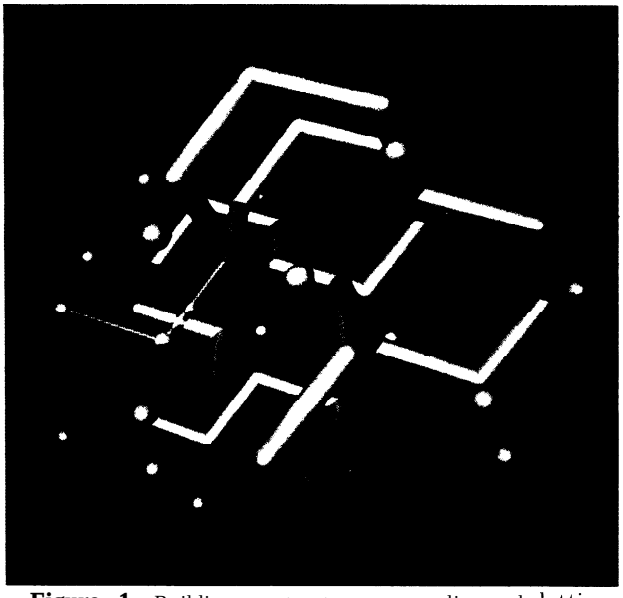


Figure 1. Building a structure on a diamond lattice. Unoccupied lattice points are shown as spheres. When adding a new link to a growing chain. there are at most 3 neighboring lattice points to choose from. Possible next steps are shown in red; in this case. 1 of the 3 options would violate the self-avoidance rule because the vertex is already occupied.

(e) *Conformational* search: selection criteria

Our strategy for evaluating our model is to exhaustively enumerate all lattice folds for a protein of known structure. filter these folds using a series of structural and energetic criteria. and then see if the resulting high-scoring folds resemble the known native structure. We first find the locally optimal alignment of a sequence to each fold: this is effectively an energetic screen. in which we keep the one lowest energy alignment out of many possible alignments. We then screen for structures with low radii of gyration and correct disulfide linkages. Finally. we pick out the 1000 structures with lowest total conformational energy.

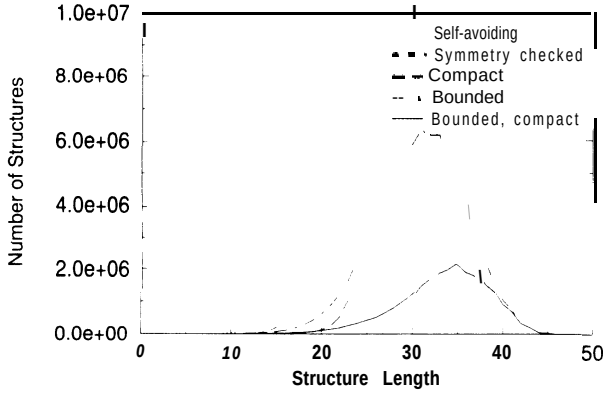


Figure 2. Number of lattice structures *versus* chain length. Symmetry checking removes structures that differ only in their orientation. Compact structures have radii of gyration no larger than 1.10 times that of a sphere with equal volume. Bounded structures are constrained to fit within an ellipsoidal volume containing 50 lattice points.

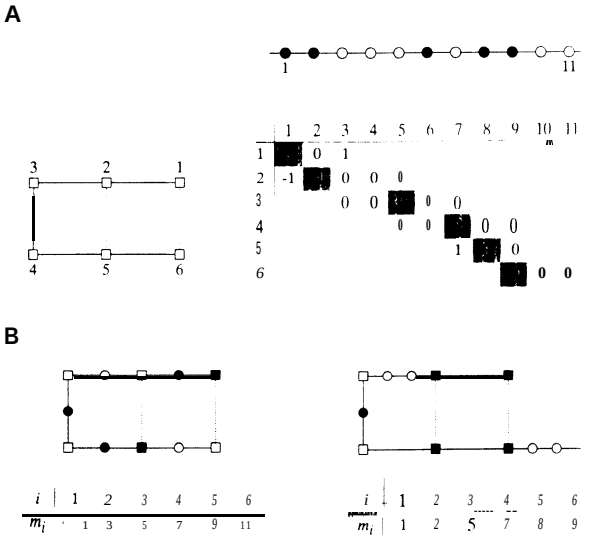


Figure 3. Illustration of optimization procedure in 2 dimensions. We use a model peptide composed of just hydrophilic (○) and hydrophobic (●) residues. We arbitrarily assign hydrophobic-hydrophobic contacts an energy of − 1. and hydrophilic-hydrophilic contacts an energy of + 1. An 11-residue peptide (A, top) is to be threaded onto a 6-vertex lattice structure with 2 long-range contacts (A, left; contacts are dotted). To start, the residues are spaced evenly along the structure (B, left). For each vertex: we calculate the interaction energies of that position with the rest of the structure, for a window of allowed moves of + / − 2 steps (A. lower right). We use dynamic programming to determine the lowest energy mapping of residues to lattice points (shaded boxes). The final mapping (B, right) places hydrophobic residues on all the lattice points that make long-range contacts.

We have experimented with a variety of other structural criteria. including the number of tertiary contacts. tests for the presence of cavities or deep crevices. and measures of chain flexibility. These measures were predictive in some cases. but were either unreliable or sufficiently correlated with our other selection steps that they added little predictive power.

It should be possible to integrate all of our separate "pass/fail" criteria into a single discriminant function, so that a single number could be used to rank all the lattice structures. We have stayed with our multi-step approach because it separates out the influence of each variable. Many of our criteria are strongly correlated, so, for instance, structures that are more compact (i.e. low radius of gyration) also tend to have low energies. The order in which we apply different criteria can thus affect their apparent predictive abilities: in general. if a test is used early. it will be more effective than if it is used late, when the remaining structures may be already enriched for that property We save our energy selection step for last. so that we have already selected for non-sequence-specific factors that might have energetic contributions.

3. Results

We will describe results for a representative subset of the small proteins we have tested, shown in Table 3. These proteins were chosen to span a variety of

Table 3
Properties of known structures used to evaluate the method

PDB code	Residues	Bonds†	Class	Description	Reference
1 cbn	46	3	$\alpha + \beta$	Cram bin	Teeter et *al.* (1993)
8rxn	52	1	β	Rubredoxin	Dauter *et al.* (1992)
2gb1	56	0	$\alpha + \beta$	St reptococcal protein G (B 1 domain)	Gronenborn *et al.* (1991)
5pt i	58	3	$\alpha + \beta$	Bovine trypsin inhibitor	Wlodawer *et al.* (1984)
1r69	63	0	α	434 repressor (N-terminal domain)	Mondragon *et al.* (1989*a*)
2cro	65	0	α	434 Cro protein	Mondragon *et al.* (1989h)
2sn3	65	4	$\alpha + \beta$	Scorpion neurotoxin variant-3	Almassy *et al.* (1983)
1ctf	68	0	α/β	L7/L12 ribosomal protein (C-terminal domain)	Leijonmarck & Liljas (1987)
1ubq	76	0	$\alpha + \beta$	Ubiquitin	Vijay-Kumar *et al.* (1987)
4i cb	76	0	α	Bovine calbindin D_{9k} (minor A form)	Svensson *et al.* (1992)
2fxb	81		$\alpha + \beta$	Ferredoxin	Fukuvama *et al.* (1989)

† Disulfide bonds. except in the case of 8rxn. which has a metal binding site.

tertiary structural classes. None of these proteins are homologous to any protein in the database used to generate our energy parameters.

We constructed a diamond lattice with an edge length of 5.08 Å. corresponding to a volume per vertex of 200 Å^3. This edge length gave the best fits between the most accurate lattice models and their corresponding native structures. judged by comparing distributions of inter-C" distances. We enumerated the lattice points contained within an eliptical bounding volume with major axes of 23.5. 23.5, and 3 1.5 Å. We removed anv vertices with fewer than two nearest neighbors. The resulting bounded lattice contains 50 vertices. For the three largest proteins (PDB codes luby. 4icb and 2fxb). we used a different bounding elipse, enclosing 51 vertices, that allows for more self-avoiding structures at the longer chain lengths. This second elipse has major axes of 23.5. 23.5. and 33.5 Å. but is oriented differently so that these 51 vertices are not a simple superset of the 50 vertex lattice. For the chain lengths we are using. these bounding volumes cont ain 20% t o 50 % more vertices than will be used by any particular structure. These vacant vertices are an important source of structural diversity in our model, and prevent the bounding volume from dictating the shapes of allowed structures.

We compare structures using the r.m.s. deviation of corresponding C^α–C^α distances, in addition to superimposing and comparing C^α coordinates directly (Cohen & Sternberg. 1980). because the distance r.m.s. correlates better with the preservation of local structural features. We are less interested in our lattice models as predictions of actual coordinates, than in seeing the extent to which our models capture native structural features. The distance r.m.s. deviation (d r.m.s.) is given by:

$$d \text{ r.m.s.} = \sqrt{\frac{\sum_{i=1}^{N}\sum_{j=1}^{N}(D(v_i, v_j) - R(m_i, m_j))^2}{N(N-1)}}, \quad (5)$$

where N is the length of the lattice structure, v_i is the lattice vertex at position i along the chain, m_i is the sequence position mapped to position i, $D(v_i, v_j)$ is the

Table 4
Energetic8 of contacts in native structures

PDB code	Average contact energy (kcal/mole)†			Normalized energies (SD units)‡		
	Native	Smoothed	Best half	Native	Smoothed	Best half
lcbn	− 0.039	− 0.002	− 0.136	− 0.41	− 0.34	− 1.30
8rxn	− 0.071	0.039	− 0.131	− 0.76	− 0.52	− 1.76
2gb1	− 0.034	0.025	− 0.105	− 0.59	− 0.52	− 1.21
5pti	− 0.122	− 0.068	− 0.243	− 0.69	− 0.66	− 2.15
1r69	− 0.126	0.023	− 0.136	− 0.94	− 0.52	1.78
2cro	− 0.107	0.016	0.165	− 0.54	− 0.31	− 1.43
2sn3	0.004	0.045	− 0.120	− 0.45	− 0.42	− 1.44
1ctf	− 0.047	0.072	− 0.076	− 0.85	− 0.66	− 1.02
1ubq	− 0.144	0.022	− 0.185	− 0.88	− 0.60	− 1.74
4icb	− 0.152	0.03 1	− 0.145	− 1.08	− 0.63	− 1.54
2fxb	− 0.036	0.037	− 0.114	− 0.59	− 0.50	− 1.18

† The native energy is the average over all true contacts in the native structure. The smoothed energy has had the energy of each native cont act averaged using equation (4). The best half energy only considers smoothed contacts between the most strongly interacting half of the residues in each sequence.

‡ Normalized energies have been converted to units of standard deviations from a mean. We determined the mean and standard deviation oft he contact energies of all possible contacts. based on the composition of the specified sequence. The smoothed statistics were calculated using energies averaged using equation (4). The best half statistics were calculated using the composition of the best half subset of each sequence.

Table 5
Numbers of conformations searched

PDB code	Path length	Number of structures		
		R_g cut†	Bounded‡	Compact§
1cbn	23	1.14	1.7×10^6	3.7×10^5
8rxn	26	1.12	3.6×10^5	5.1×10^5
2gb1	28	1.12	4.9×10^5	8.1×10^5
5pti	29	1.12	5.7×10^6	1.0×10^5
1r69	32	1.10	6.2×10^6	8.0×10^5
2cro	33	1.10	6.2×10^6	9.4×10^5
2sn3	33	1.10	6.2×10^6	9.4×10^5
1ctf	34	1.10	0.4×10^7	9.9×10^5
1ubq	38	1.08	4.7×10^7	1.5×10^6
4icb	38	1.08	4.7×10^7	1.5×10^6
2fxb	41	1.08	2.6×10^7	1.2×10^6

† Upper bound for radius of gyration, relative to that of a sphere with the same volume.

‡ Self-avoiding, symmetry-unrelated, before using radius of gyration cut.

§ Bounded structures that also pass radius of gyration cut.

distance between lattice vertices v_i and v_j, and $R(m_i, m_j)$ is the distance between the C^α atoms of residues m_i and m_j. This only considers distances between the subset of residues that are mapped to lattice points; if we interpolate positions for residues between lattice points, the resulting all-residue dr.m.s. is nearly the same, but is much **more expensive** to compute over several million structures.

We also score lattice structures by counting the contacts they have in common with the native structure, and by the ratio of native contacts to total contacts. When counting contacts, we use the same weights we use when calculating the conformational energy of a structure, in equation (4). These measures are useful because they can identify locally correct features in misfolded structures, and because of their connection with the derivation of our energy parameters from contact counts. In our experience, the ratio of native contacts to total contacts is a somewhat less useful measure than the simple sum of native contacts, because it favors extended structures with few total contacts.

(a) *Energetics*

Our simplified energy function will only be a useful predictor when applied to lattice structures if a protein's native structure has a low energy. We verified this by calculating the energies of actual contacts in the native structures of our test proteins (Table 4). The actual contact energies are very sequence-dependent so we also calculated normalized energies based on the expected values for random contacts for each sequence. Native contacts do in fact have low energies, and continue to do so when we average the interactions around each contact using equation (4).

We tried picking out the best half of the residues in each sequence, defined by summing the interaction energies (Table 2) of each residue with the rest of that sequence and picking the half of the residues with the lowest sums. The selection of this subset does not depend on the native structure, only on the sequence.

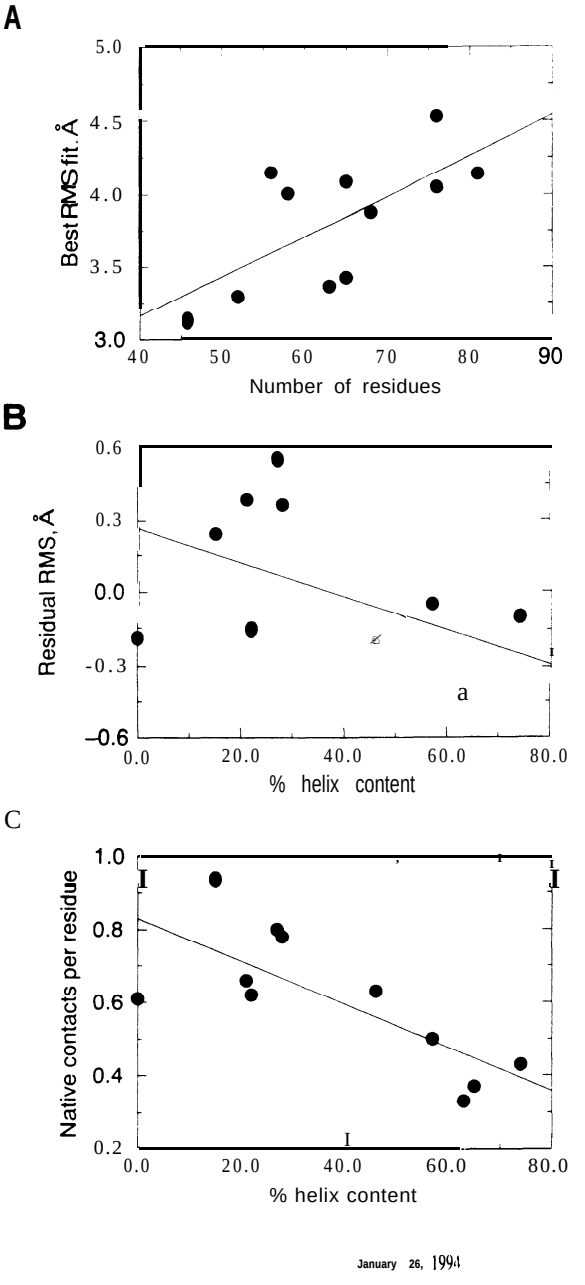


Figure 4. Dependence of best lattice fits on native secondary structure composition and size. The best lattice fits for larger structures tend to have higher distance r.m.s. deviations(X). Structures with higher helix content can be fit somewhat more accurately (B, residual r.m.s. is the best fit r.m.s. after subtracting the contribution of the sequence length from A), but the best fits tend to have fewer contacts in common with the native structure (C).

When we only consider contacts between these strongly interacting residues, these tend to have much more favorable energies than would be expected by chance, taking into account the new sequence composition. This is significant because our alignment procedure increases the importance of these residues in our conformational energy calculations.

Table 6
Effects of alignment optimization

Population property	Average and range: Native sequences	Average and range: Permuted sequences
Distance r.m.s. (Å)		
Change in mean	− 0.41 (− 0.85 to − 0.11)	0.03 (− 0.15 to + 0.14)
Change in lowest value	− 0.48 (0.98 to − 0.09)	− 0.28 (− 0.73 to − 0.05)
Coordinate r.m.s. (Å)		
Change in mean	− 0.58 (− 0.90 to − 0.29)	− 0.14 (− 0.22 to + 0.04)
Change in lowest value	− 0.15 (− 1.15 to − 0.18)	− 0.27 (− 0.79 to + 0.05)
Number of native contacts		
Change in mean	***+ 8.8 (+ 3.6*** to + 12.6)	+ 1.8 (− 0.3 to + 2.7)
Change in highest value	+ 18.1 (+ 7.5 to + 25.5)	+ 13.3 (+ 6.5 to + 21.5)
Fraction of native contacts (%)		
Change in mean	+ 4.8 (+ 2.0 to + 6.3)	+ 1.0 (− 0.2 to + 1.8)
Change in highest value	+ 9.5 (+ 5.1 to + 15.3)	***+ 7.3*** (+ 4.1 to + 13.8)

(b) *Generation of lattice structures and alignment*

For each of' our test proteins. we generated all self-avoiding structures on the appropriate bounded lattice, using all svmmetrv-unrelated vertices within the elliptical bounding vdlume as starting points. A svmmet ry checking step guaranteed that for each starting point, no structures would be generated that could be related by simple rotations or inversions. The total numbers of bounded structures ranged from 1.7 x 10" to 4.7 x 10^7, shown in Table 5. We chose relative radius-of'-gyration limits to decrease gradually with increasing protein size. since smaller proteins are often more irregular in shape. For the longer chain lengths. a tighter cutoff also still gave reasonably large diversities of structures. After this selection step. we were left n-it h on the order of 10^6 structures for each protein.

The best fits for larger proteins tend to have larger r.m.s. deviations than the fits for smaller proteins (Figure 4). After correcting for this size effect by subtracting the predicted r.m.s. values. the best fits for the more helical proteins tend to have lower r.m.s. deviations. However: they also tend to have fewer native contacts. because the usual $\langle i, i + 3)$ and $\langle i, i + 4)$ helical interactions are too short range to be represented in our model.

We find that optimizing the spacing of residues along our low-resolution lattice models significantly improves the extent to which these models can represent native structural features. After shifting residues to optimize their tertiary interactions. we

Table 7
Summary of screening results based on distance r.m.s.

PDB code	dr.m.s. cut (A)	Hits‡	Selection steps§: R_g	Opt	s-s	E	Total
1cbn	3.09	0		2 x	5.8 x	0 x	0 x
8rxn	***3.53***	4	3.1 x	18 x	***3.0 x***	4 x	706 x
2gb1	***4.39***	2	4.6 x	15 x		7 x	495 x
5pti	4.21	16	2.3 x	38 x	15.6 x	3 x	4570 x
1r69	3.46	0	2.7 x	31 x		0 x	0 x
2cro	***3.50***	0	2.3 x	2 x		0 x	0 x
2sn3	4.24	18	2.0 x	33 x	7.2 x	1% x	5571 x
1ctf	4.02		2.7 x	40 x		17 x	1886 x
1ubq	4.59	3	0.3 x	7000 x		3 x	7039 x
4icb	3.92	12		996 x		28 x	28156 x
2fxb	4.30	0	1.1 x	62 x	2.3 x	0 x	0 x

† Cutoff for which there are 20 structures as close as this to the native. out of all bounded structures. prior to optimizing their sequence mapping.

‡ Among the 1000 lowest energy structures. the number that meet the dr.m.s. cut off.

§ The selectivity of each step is shown as a purification factor: the ratio of concentrations of structures meeting the dr.m.s. cutoff before and after that step. R_g is the radius of gyration screen. Opt is the alignment optimization step. S-S is the disulfide screen. and E is the energy screen. Total is the proportion of structures that meet the dr.m.s. cutoff in the final 1000 low-energy structures. compared to the proportion in the population of all bounded structures. 0. Indicates that no low-energy structures met the cutoff. For lcbn and 4icb, none of the best unaligned structures passed the R_g cutoff. so the Opt value represents the combination of the R_g and Opt steps.

Table 8
Summary of screening results based on nat ive contacts

PDB code	Contacts†	Hits‡	Selection steps§ R_g	Opt	S-S	E	Tot al
1 cbn	27.0	38	4.1 x	64 x	8.0 x	1.5 x	3230 x
8rxn	30.0	114	5.9 x	273 x	1.2 x	10.0 x	20520 x
2gb1	41 .0	87	5.2 x	634 x		6.5 x	21528 x
5pti	41.5	273	4.7 x	719 x	5.8 x	3.3 x	77805 x
1 r69	21.5	589	5.5 x	10000 x		2.4 x	182590 x
2cro	20.5	29	5.2 x	1000 x		1.4 x	8990 x
2sn3	57.0	99	5.9 x	200 x	5.9 x	4.2 x	30690 x
1 ctf	29.5	244	4.2 x	3200 x		5.0 x	65880 x
lubq	47.0	14	3.1 x	4000 x		2.8 x	32900 x
4icb	28.5	449	3.1 x	72500 x		4.i x	1055150 x
2fxb	47.0	12	2.2 x	1850 x	1.2 x	3.3 x	16160 x

† Out of all bounded structures, there are 20 with this many native contacts. prior to optimizing their sequence mapping. Contacts are weighted as they are in equation (4). so fractional contacts are possible.

‡ Among the 1000 lowest energy structures. the number that meet the native contact cutoff.

§ R_g. is the radius of gyration screen. Opt is the alignment optimization step. S-S is the disulfide screen. and E is the energy screen. Total is the proportion of structures that meet the native contact cutoff' in the final 1000 lowest energy structures. compared to the proportion in the population of all bounded structures.

find that the resulting structures have more native contacts, and lower r.m.s. deviations from their corresponding native structures, as shown in Table 6. This improvement is largelv sequence-dependent: if we permute the sequence of a protein. keeping its amino acid composition constant. the optimization step has much less effect on the distributions of these native properties. After optimizing the permuted sequences, the distributions are broader than those for an equal spaced mapping, but their means show only small shifts towards greater agreement with the correct structure.

We had hoped that periodic spacings of favorable interactions would enable our method to pick out secondary structural features like amphipathic helices. to assign residues to lattice points with the correct residue spacing. In practice. these patterns do not seem to be distinct enough for our method to converge to the right spacings. The energy-optimized spacings of residues along segments of lattice structures that correspond to stretches of a-helix or P-sheet are not significantly different. We think this is because our lattice structures need to be only topologically accurate to have favorable energies. Stretching and squeezing a structure, preserving its pattern of tertiarv contacts, will not change its total energy using our method.

Residues positioned at lattice vertices contribute more to the energy function than those between vertices. One effect of the alignment step is to shift residues that tend to make favorable contacts onto lattice vertices. The arrangement of these residues, which are mostly hydrophobic. then becomes the most important determinant of the structure's tot al energy The success of this step. which in effect throws away most of the information contained in the positions of the hydrophilic residues, indicates the importance of forming a hydrophobic core in establishing a protein's overall chain fold.

(c) Structure prediction

We evaluated the success of our selection procedure by comparing distributions of native properties in our low energy subset with these distributions for the entire population of bounded structures. We looked at the overall shifts in these distributions: but also focused on their extremes, because the most native-like structures are most important from the standpoint of making useful predictions. As a test for sequence specificity of our prediction procedure, we randomized the sequences of all our test proteins. preserving sequence composition. For each protein. we generated a new permuted sequence for every bounded, compact lattice structure. We could then compare these results with permuted sequences with our results for the native sequence ordering.

For a particular native property we consider a "good" structure to be one as good as the 20 best models for that protein out of the entire population of bounded structures, before alignment, optimization. This cutoff value varies from protein to protein, depending on how well the lattice is able to represent the features of a particular structure. We rate our method by seeing how many of our low-energy structures meet this cutoff. The proportion of these 'good' structures in the low energv set. compared to the proportion in the population of all structures, is a measure of the selectivity of the method. We quantifv the predictive power of individual screening steps the same way. using the ratio of the concentrations of 'good structures' before and after the screen. While this protein-specific measure is less useful for judging the utility of our method for

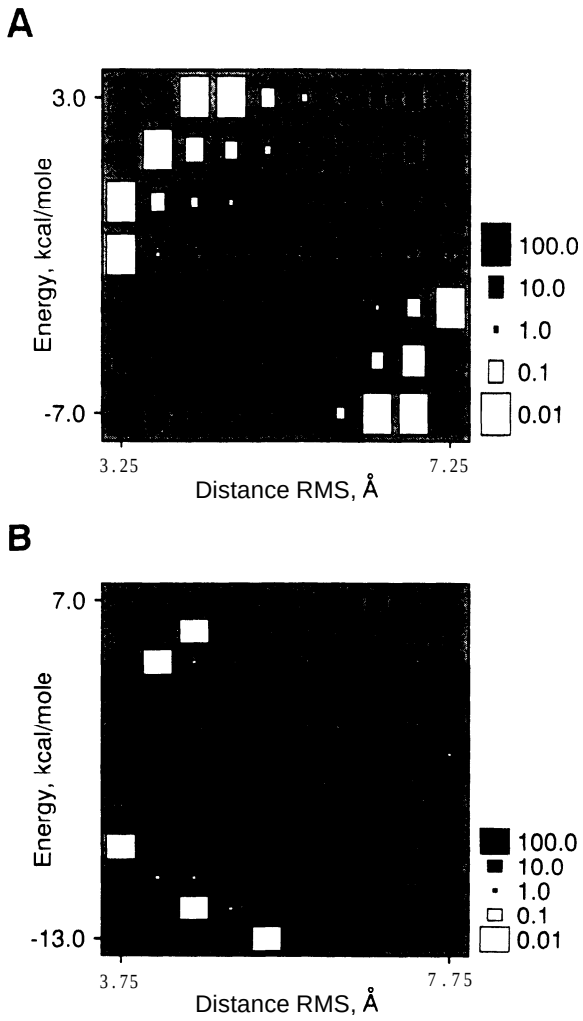


Figure 5. Box plots showing correlation between distance r.m.s. deviation and conformational energies for the C-terminal fragment of ribosomal protein L7/L12 (1ctf). A box plot is a scaled scatterplot. depicting the interaction between 2 variables. The box size indicates the ratio of the number of points observed in that region of the plot, to the number that would be expected by chance, given the overall distributions of the 2 variables. Filled boxes are used to indicate ratios greater than one. and open boxes indicate ratios less than one: the box edge length is proportional to the log oft he ratio. So. a positive correlation is indicated by filled boxes along the diagonal from lower left to upper right, and open boxes in the upper left and lower right. In this plot, energy is clearly correlated with distance r.m.s. for the set of all bounded. compact lattice structures (A) (r = 0.275). If the sequence is permuted (B), the correlation disappears (r = 0.024).

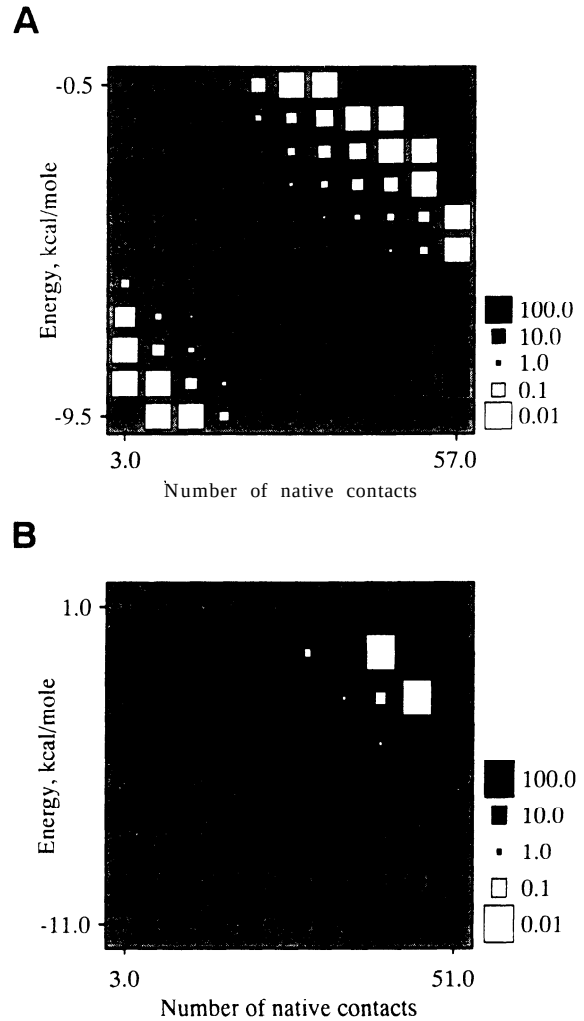


Figure 6. Box plots (as in Figure 5) showing correlation between numbers of native contacts and energies for streptococcal protein G (2gb1). Energy is inversely correlated with native contact counts (A) (r = - 0.354). If the sequence is permuted (B), the correlation drops considerably (r = - 0.096).

predicting unknown structures. we think it is the best way of evaluating our selection scheme in the context of the lattice model we have chosen.

The selectivities of each screening step. with respect to distance r.m.s. and native contact counts. are shown in Table 7 and Table 8. We are more successful at finding structures with many native contacts, than at finding structures with the lowest r.m.s. deviations. Most of our predictive power comes in the two energy-dependent) steps. For these small proteins, at such low resolution, enforcing native disulfide linkages is less helpful than might be expected: in the case of scorpion neurotoxin, with four disulfide bonds, these constraints add only a factor of seven in selectivity

A structure's conformational energv is correlated with its r.m.s. deviation from the native structure (Figure 5). and with its number of native contacts (Figure 6). During the course of our selection procedure. the distributions of native properties shift towards greater agreement) with native structures (Figures 7. 8 and 9). While the positions of these distributions shift, their widths remain roughly constant; low energy lattice structures tend to be closer to the native structure, but there is still a lot of variation within the low energy set. We summarize

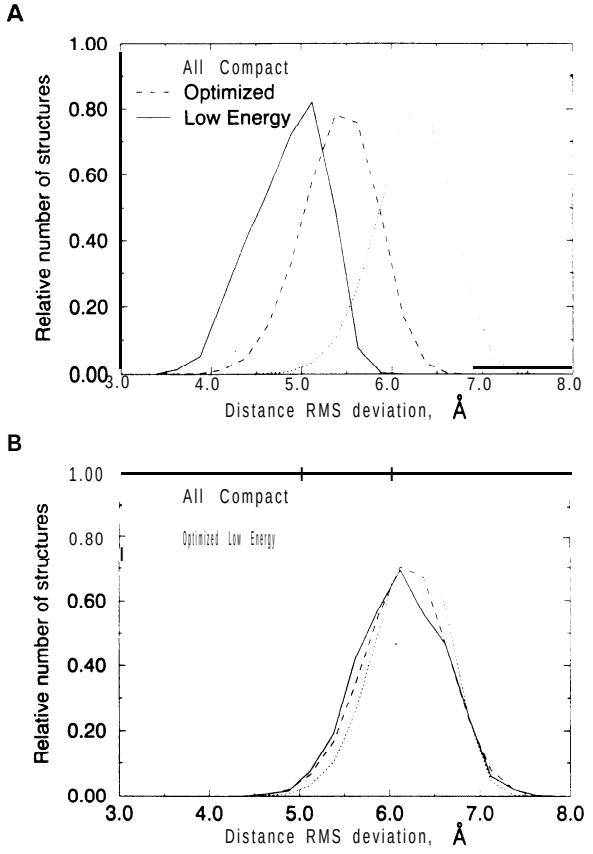


Figure 7. Population distributions of distance r.m.s. deviations for calbindin (4icb). We show distributions for all compact bounded structures before and after optimizing the alignment of the sequence to each structure. and for the 1000 lowest energy structures (A). Both energy-driven steps shift the distribution towards the native structure. but the width of the distribution remains roughly constant. If the sequence is permuted for each lattice structure (B). the selection steps have almost no effect on the distribution.

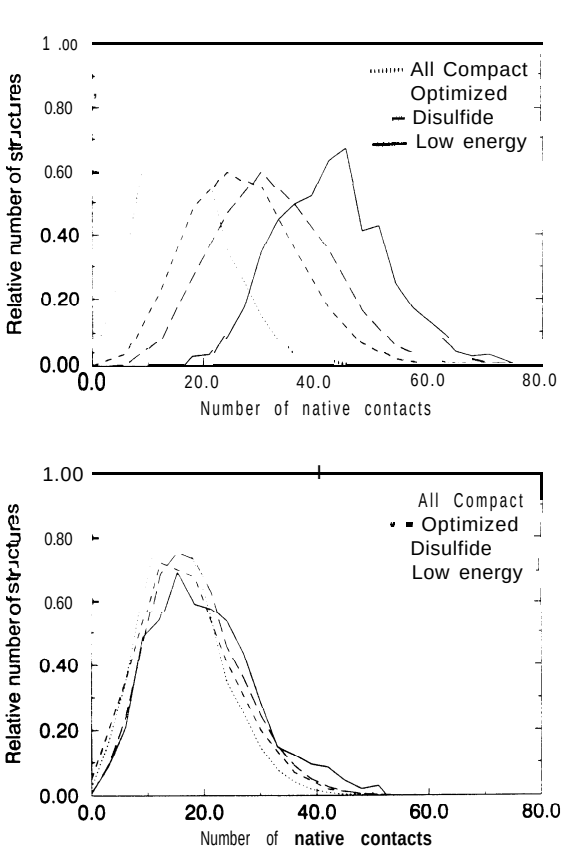


Figure 8. Population distributions of numbers of native contacts for scorpion neurotoxin (2sn3). We shou distributions for all compact bounded structures before and after alignment. after selecting structures for approximately native disulfide connectivities. and for the 1000 lowest energy structures (A). If the sequence is permuted (B). none of the steps significantly improve the distribution.

the magnitudes of the shifts in Table 9. In every case. permuting a protein's sequence removes the correlations between the conformational energies of our models and their native properties, and removes the shifts in population distributions in the screening steps. This verifies that the selective power of our model is completely dependent on information contained in the native ordering of residues along the sequence.

The best low-energv lattice structures are consistent with the overall folds of their corresponding native structures (Figure 10). even for the proteins that did not score well in Table 7. The lattice structures are distorted. and many native features are poorlv represented. The cores of these best models are held together bv dense networks of native contacts (Figure 11). The most native-like models cannot be identified by their energies. but in some cases. the absolute minimum energy lattice model is at least recognizably similar to the native structure (Figure 12).

4. Discussion

A shortcoming of our lattice model is that it is so low resolution that it is impossible to represent differences between secondary structural units: a β strand looks the same as an a helix. As a result. our structures do not contain any information about side-chain orientations. To the extent that these structural details are important in determining a protein-s overall topology this limits the accuracy of our predictions. Our exhaustive conformational searches are also ultimately limited by their exponential dependence on chain length. We can work around this to some extent by increasing the ratio of residues to points on the lattice. and we have had some success using this approach with larger proteins. However. the number of lattice points in our chains places an upper limit on the number of secondary structural segments and turns that can be even roughly approximated.

In even the worst cases, our energy function is at least weaklv predictive, but not always to the extent that any of the most native-like structures (by r.m.s.) show up in the 1000 lowest energy structures. We

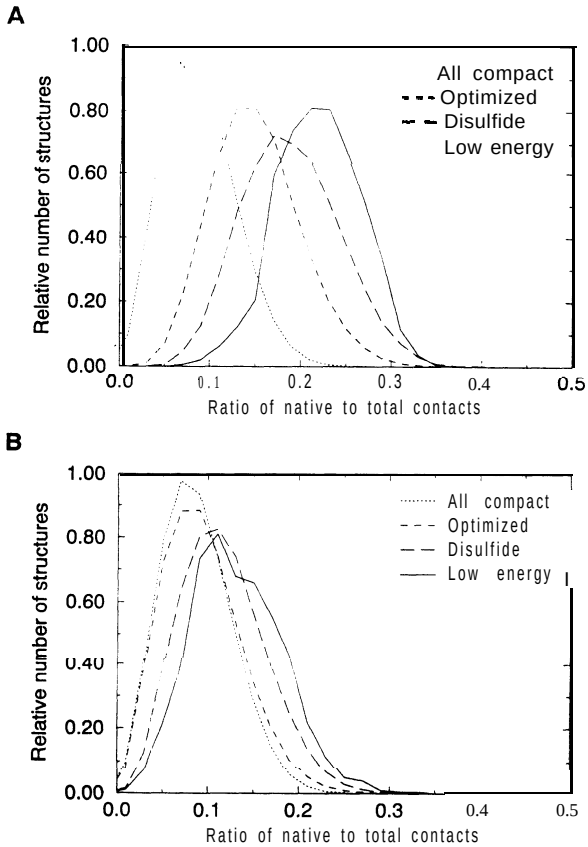


Figure 9. Population distributions of ratios of native contacts to total contacts for bovine pancreatic trypsin inhibitor (5pti). We show distributions for all compact bounded structures before and after alignment. after selecting structures for approximately native disulfide connectivities. and for the 1000 lowest energy structures with the native (A) and permuted (B) sequences.

know that for these proteins. the native contacts do have low energies. Therefore. our best explanation for this failure is that due to the geometry of the lattice. the low r.m.s. models for these proteins contain either relatively few native contacts. or an unrepresentative subset of native contacts. In other words. the lowest r.m.s. models are sacrificing local correctness to get a bet ter overall r. m .s. fit. These less successful examples may reflect an inevitable problem with discrete simplified models: given the structural diversity of proteins. for any model. there will he pathological cases for which that model is unable to represent key structural features.

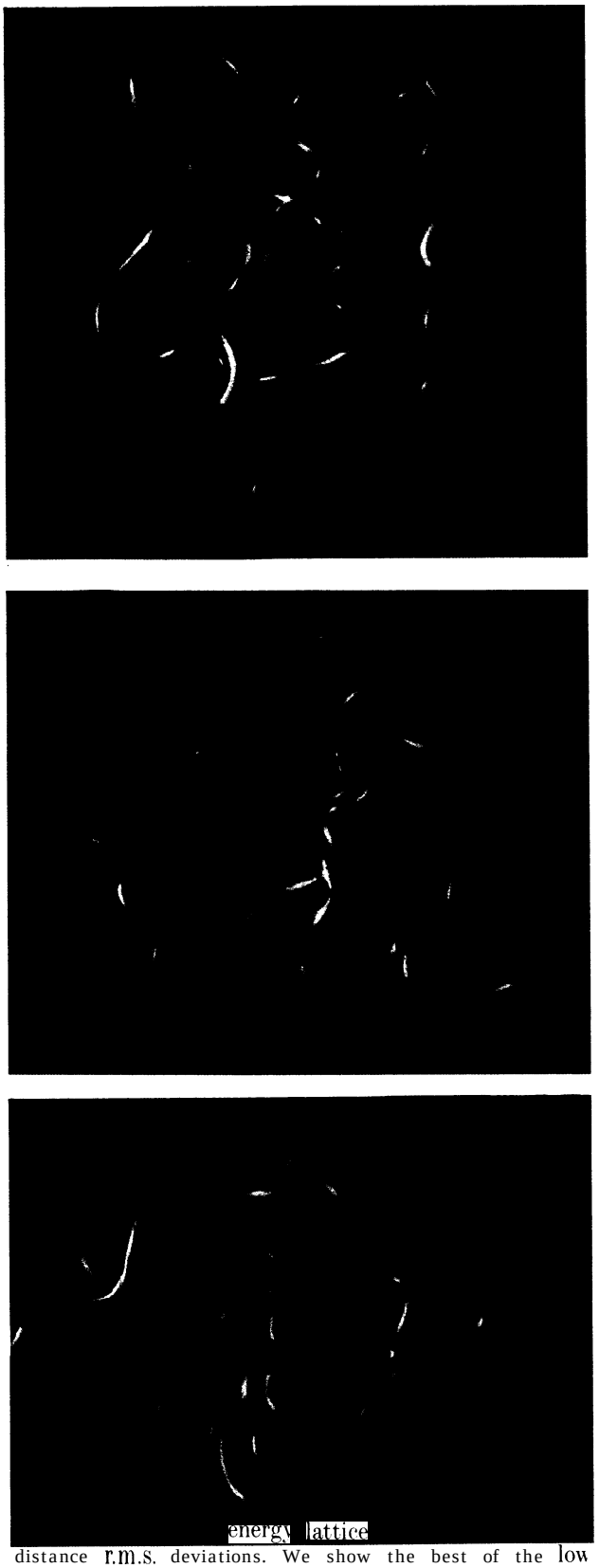
Our results are roughly comparable to those of another recent low resolution folding study (Covell, 1992, 1994) using a different lattice representation and a Monte Carlo sampling procedure. The fractions of native contacts we see in our lattice structures are much lower, however. Our most native-like lattice models typically have no more than **20%** or 30% native contacts, and our random compact structures average closer to **5%** native contacts. These compare to about 50% and 25% native contacts. respectively in Covell's work. The likely reason for this difference is that all of our lattice contacts are truly long range, while many of the native contacts seen by Cove11 may have been relatively short range? such as contacts bet ween residues within the same secondary structural element.

This particular low resolution model does not seem to be well suited for extension to higher resolution. Using conventional energy minimization techniques. we can build all-atom models, but they are of no higher quality than the original lattice coordinates. The errors in the model coordinates are fairly large, so we still need to search a relatively large window of conformations around any particular lattice model. Our low energy structures do not cluster well, so we would need to explore windows around each one of these models. A practical method for going to higher resolution would need to make these search windows large enough to catch good structures. but small enough that the windows around different lattice structures do not overlap.

A better strategy may be to use a low resolution model tailored to the structural features commonly found in proteins. A discrete model composed of idealized building blocks (helices, turns, strands, and loops) should be able to more accurately represent real structures than our lattice model, without increasing the total number of allowed conformations. This new model would have the advantage of determining the positions and approximate orientations of side-chains. Our contact energy function

Table 9
Effects of all selection steps on distributions of native properties

Population propert y	Change in mean: average and range Native sequences	Permuted sequences
Distance r.m.s. (A)	- 0.85 (- 1.43 to - 0.19)	- 0.05 (- 0.23 to + 0.12)
in SD units	- 2.0 (- 3.4 to - 0.5)	- 0.1 (- 0.5 to + 0.2)
Coordinate r.m.s. (A)	- 1.20 (2.22 to - 0.50)	- 0.19 (- 0.40 to - 0.03)
in SD units	- 1.2 (- 2.2 to - 0.5)	- 0.2 (- 0.4 to - 0.0)
Number of native contacts	+ 16.3 (+ 5.7 to + 26.4)	+ 3.6 (+ 0.9 to + 5.4)
in SD units	+ 3.5 (+ 2.1 to + 5.8)	+ 0.8 (+ 0.2 to + 1.5)
Fraction of nat ive contacts (%)	+ 8.3 (+ 2.5 to + 13.9)	+ 1.6 (+ 0.0 to + 2.7)
in SD units	+ 3.2 (+ 1.8 to + 5.8)	+ 0.7 (+ 0.0 to + 1.4)



energy lattice

distance r.m.s. deviations. We show the best of the low energy lattice models for 4icb (A). 2cro (B). and 2fxb (C). The native backbones are in blue. and the lattice structures are in red. The overall folds appear to be correct. even for the proteins that did not fare well based on Table 7. These models have distance deviations of 3.65, 4.22. and 4.81 A. and coordinate deviations of 5.52. 6.64, and 6.67 A, respectively.

seems to have potentially much more predictive power than we have been able to exploit in our simple model. A model specifying side-chain orientations should be able to fit native contact patterns more accurately and make better use of these contact energies.

We think it is remarkable that such a limited energy function correlates with native properties of' our equally limited models for polypeptides. Though folded proteins are onlv marginally stable. the folding "signal" in a sequence'must be very robust to survive this amount of abuse. The overall pattern of hydrophobic and hydrophilic residues in a protein seems to be a very strong determinant of its overall fold. A similar conclusion was reached by Hecht and co-workers (Kamtekar et *al.*, 1993). who found that a fen- simple rules were sufficient for generating sequences able to fold into stable four-helix bundles. Together these validate the use of "low resolution" and hierarchical approaches to the folding problem. in that low resolution properties of a sequence are sufficient to determine a low resolution structure. This suggests that some aspects of the protein folding problem will prove to be not so difficult after all.

This work was supported by the National Institutes of Health, grant GM30387-12. D. H. is a Howard Hughes Medical Institute Predoctoral Fellow. Color graphics images in Figures 10 and 12 were generated using the MidasPlus software system from the Computer Graphics Laboratory,

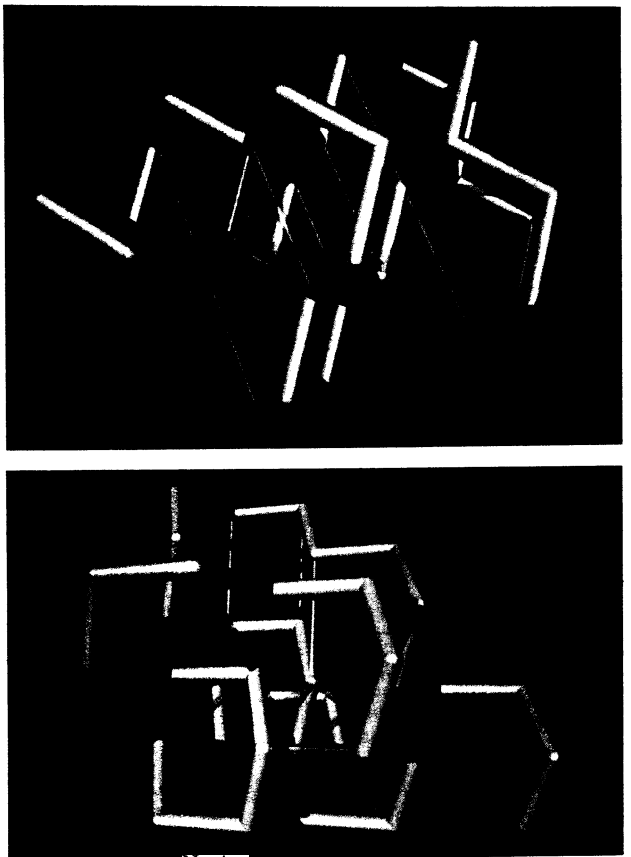


Figure 11. Native contacts are densely distributed throughout the cores of the best models for lctf (A) and 4icb (B). The lattice structure is shown in purple, and contacts are shown in red.

Figure 12. Selected low energy lattice models of rubredoxin (8rxn). We show the structures with lowest r.m.s. deviation (A). most native contacts (B). and minimum energy (C). These structures have distance r.m.s. deviations of 3.1 1. 3.89. and 4.19 Å from the native structure: coordinate r.m.s. deviations of 4.85. 6.52. and 6.17 A: and 40.5. 42.5. and 33.5 native contacts. respectively.

University of California, San Francisco. We thank M. Gerstein and E. Huang for help with this manuscript.

References

Abagyan. R. A. (1993). Towards protein folding by global energy optimization. *FEBS Letters, 325, 17-h.*

Almassy. R. J.. Fontecilla-Camps, J. C., Suddath. F. L. & Bugg, C. E. (1983). Structure of variant-3 scorpion neurotoxin from *Centruroides sculpturatus* ewing. refined at 1.8 Å resolution. *J. Mol. Biol.* 170. 497–527.

Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science.* **181,** 223-230.

Anfinsen, C. B., Haber, E., Sela, M. & White, F. H. (1961). The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proc. Nat. Acad. Sci., U.S.A.* **47**, 1309–1314.

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.**

Brower. R. C.. Vasmatzis. G., Silverman M. & Delisi. C. (1993). Exhaustive conformational search and simulated annealing models of lattice peptides. *Biopolymers.* **33, 329-334.**

Bryant, S. H. & Lawrence, C. E. (1993). An empirical energy function for threading protein sequence through the folding motif. *Proteins, Struct. Funct. Genet.* **16.** 92-l 12.

Cohen, F. E. & Sternberg, M. J. E. (1980). On the prediction of protein structure: the significance of the root-mean-square deviation. *J. Mol. Biol.* **138,** 321-333.

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

Covell. D. G. (1994). Lattice model simulations of polypeptide chain folding. *J. Mol. Biol. 235.* 10321043.

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

Crippen. G. M. (1991). Prediction of protein folding from amino acid sequence over discrete conformation spaces. *Biochem is t ry. 30. 4232-4237.*

Dauter. Z.. Sieker. L. & Wilson, K. (1992). Refinement of rubredoxin from *Desulfovibrio vulgaris* at 1.0 Å with and without restraints. *Acta Crystallogr. Sect. B. 48.* 42–59.

Dill. K. A. (1990). Dominant forces in protein folding. *Biochem istry,* 29. 7 133-7 155.

Fukuyama. II;.. Matsubara. H.. Tsukihara. T. & Katsube. Y. (1989). Structure of [4Fe-4S] ferredoxin from *Bacillus thermoproteolyticus* refined at 2.3 A resolution. Structural comparisons of bacterial ferredoxins. *J. Mol. Biol.* **210. 383-398.**

Go, N. (1978). Respective roles of short- and long-range interactions in protein folding. *Proc. Nat. Acad. Sci.. U.S.A.* 75. 559-563.

Gronenborn. A. M.. Filpula. D. R., Essig. N. Z.. Achari. A.. Whitlow. M.. Wingfield. I? T. & Clore. G. M. (1991). A novel. highly stable fold of the immunoglobulin binding domain of streptococcal protein G. *Science,* 253. 657-661.

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

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

Kamtekar, S., Schiffer, J. M., Xiong, H., Babik, J. M. & Hecht, M. H. (1993). Protein design by binary patterning of polar and nonpolar amino acids. *Science*, **262**, 1680–1685.

Kauzmann, W. (1959). Some factors in the interpretation of protein denaturation. *Adv. Protein Chem.* **14**, 1–63.

Kuntz, I. D., Crippen, G. M., Kollman, P. A. & Kimelman, D. (1976). Calculation of protein tertiary structure. *J. Mol. Biol.* **106**, 983–994.

Lau, K. F. & Dill, K. A. (1989). A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules*, **22**, 3986–3997.

Leijonmarck, M. & Liljas, A. (1987). Structure of the C-terminal domain of the ribosomal protein L7/L12 from *Escherichia coli* at 1.7 Å. *J. Mol. Biol.* **195**, 555–580.

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

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

Levitt, M. & Warshel, A. (1975). A computer simulation of protein folding. *Nature (London)*, *253*, 694–698.

Miyazawa, S. & Jernigan, R. L. (1985). Estimation of effective interresidue contact energies from protein crystal structures: Quasi-chemical approximation. *Macromolecules*, **18**, 534–552.

Mondragon, A., Subbiah, S., Almo, S. C., Drottar, M. & Harrison, S. C. (1989*a*). Structure of the amino-terminal domain of phage 434 repressor at 2.0 Å resolution. *J. Mol. Biol.* *205*, 189–200.

Mondragon, A., Wolberger, C. & Harrison, S. C. (1989*b*). Structure of phage 434 Cro protein at 2.35 A resolution. *J. Mol. Biol.* *205*, 179–188.

Shackhnovich, E. I. & Gutin, A. M. (1990). Enumeration of all compact conformations of copolymers with random sequences of links. *J. Chem. Phys.* 93, 5967–5971.

Skolnick, J. & Kolinski, A. (1990). Simulations of the folding of a globular protein. *Science*, 250, 1121–1125.

Skolnick, J., Kolinski, A., Brooks, C. L. III, Godzik, A. & Rey, A. (1993). A method for predicting protein structure from sequence. *Curr. Biol.* *3*, 414–423.

Svensson, L. A., Thulin, E. & Forsen, S. (1992). Proline *cis-trans* isomers in calbindin D_{9k} observed by X-ray crystallography. *J. Mol. Biol.* *223*, 601–606.

Teeter, M. M., Roe, S. M. & Heo, N. H. (1993). Atomic resolution (0.83 Å) crystal structure of the hydrophobic protein crambin at 130 K. *J. Mol. Biol.* *230*, 292–311.

Vijay-Kumar, S., Bugg, C. E. & Cook, W. J. (1987). Structure of ubiquitin refined at 1.8 Å resolution. *J. Mol. Biol.* **194**, 531–544.

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

Wilson, C. & Doniach, S. (1989). A computer model to dynamically simulate protein folding with crambin. *Proteins. Struct. Funct. Genet.* *6*, 193–209.

Wlodawer, A., Walter, J., Huber, R. & Sjolin, L. (1984). Structure of bovine pancreatic trypsin inhibitor. Results of joint neutron and X-ray refinement of crystal form II. *J. Mol. Biol.* **180**, 301–329.

Wonnacott, T. H. & Wonnacott, R. J. (1984). *Introductory Statistics for Business and Economics*. John Wiley & Sons, New York.

Edited by B. Honig

(Received 31 January 1994; accepted 29 July 1994)