PROTEINS: Structure, Function, and Genetics 32:211–222 (1998)

Recognition of Native Structure From Complete Enumeration of Low-Resolution Models With Constraints

Banu Özkan[1,2] and Ivet Bahar[1,2]*
[1]*Polymer Research Center and Chemical Engineering Department, Bogazici University, Istanbul, Turkey*
[2]*TUBITAK Advanced Polymeric Materials Research Center, Bebek, Istanbul, Turkey*

ABSTRACT **Complete sets of low-resolution conformations are generated for eight small proteins by rotating the C_α-C_α virtual bonds at selected flexible regions, while the remaining structural elements are assumed to move in rigid blocks. Several filtering criteria are used to reduce the ensemble size and to ensure the sampling of well-constructed conformations. These filters, based on structure and energy constraints deduced from knowledge-based studies, include the excluded volume requirement, the radius of gyration constraint, and the occurrence of sufficiently strong attractive inter-residue potentials to stabilize compact forms. About 8,000 well-constructed decoys or "probable folds" (PFs) are constructed for each protein. A correlation between root-mean-square (rms) deviations from X-ray structure and total energies is observed, revealing a decrease in energy as the rms deviation decreases. The conformation with the lowest energy exhibits an rms deviation smaller than 3.0 Å, in most of the proteins considered. The results are highly sensitive to the choice of flexible regions. A strong tendency to assume native state rotational angles is revealed for some flexible bonds from the analysis of the distributions of dihedral angles in the PFs, suggesting the formation of foldons near these locally stable regions at early folding pathway. Proteins 32:211–222, 1998.** © **1998 Wiley-Liss, Inc.**

Key words: sequential folding; local structure formation; coarsed-grained simulations; knowledge-based potentials; virtual bond rotations; misfolded structures

INTRODUCTION

Two major difficulties faced during the computational search for the most favorable state of proteins are (1) the exponentially large number of possible conformations, and consequently the low probability of generating a sufficient number of compact conformations having native-like packing density, and (2) the lack of effective criteria for differentiating between correct and incorrect folds. A possible way of overcoming the first difficulty is to adopt coarse-grained, or low-resolution, models. These allow for extensive coverage of the conformational space. After generating coarse-grained conformations, an important question is to assess whether an energy function exists that is able to discriminate between the native fold and the misfolded structures.[1,2] Several sets of empirical energy functions have been derived for this purpose using Protein Data Bank (PDB)[3] structures, as recently reviewed.[4,5]

One aim of the present study is to test the possibility of recognizing the native fold among a set of well-constructed decoys, using a recently proposed low-resolution model and energy parameters.[6] Complete sets of conformations will be generated for eight test proteins, using a number of structure and energy constraints that ensure the optimization of the computational time and memory requirements, and also the extraction of the relatively more probable folds. The resulting set of well-constructed decoys will be analyzed with the objectives of (1) characterizing the energies of the conformations as a function of their root-mean-square (rms) deviation from X-ray structures, and (2) identifying stretches of residues, or protein segments, accurately folded in a large number of low-energy conformers. An effort in this direction is motivated by recent studies[7–9] that emphasize the usefulness of examining ensembles of energetically favorable conformations as a realistic means of approaching the protein folding problem.

Our conformation generation approach is similar in spirit to that recently adopted by Park and Levitt.[2] Basically, the packing of rigid structural elements connected by flexible strings of amino acids is explored. Some segments of the proteins are therefore implicitly assumed to possess sufficient stability on a local scale to maintain their structure during the three-dimensional organization of the molecule. The

*Correspondence to: I. Bahar, Polymer Research Center and Chemical Engineering Department, Bogazici University, and TUBITAK Advanced Polymeric Materials Research Center, Bebek 80815, Istanbul, Turkey. E-mail: bahar@prc.bme.boun.edu.tr

Received 16 December 1997; Accepted 19 March 1998

TABLE I. Proteins Considered in the Present Study

PDB code	Protein	Resolution (Å)	Size (n)	Structural class	Reference
4RXN	Rubredoxin	1.2	54	—	47
4PT1	Bovine pancreatic trypsin inhibitor	1.5	58	$\alpha + \beta$	48
1R69	Phage 434 repressor (N-terminal domain)	2.0	69	α	49
2CRO	434 Cro protein	2.35	71	α	50
1SN3	Scorpion neurotoxin	1.8	65	$\alpha + \beta$	46
1CTF	L7/L12 ribosomal protein (C-domain)	1.7	74	α/β	51
3ICB	Calbindin D_{9k} (vitamin D-dependent)	2.3	75	α	42
1UBQ	Ubiquitin	1.8	76	$\alpha + \beta$	52

set of accessible conformations is further reduced to about 8,000 well-constructed decoys for each protein upon resort to multiple screening criteria controlling the global size and energy characteristics of the generated folds.

A sequential scheme for structure formation goes back to the original proposals of Ptitsyn and Rashin[10] and Cohen et al.[11] for α-helices, and Cohen et al.[12] for β-strands, while distance-constraint approaches to protein folding may be traced back to the studies of Wako and Scheraga.[13,14] Folding algorithms utilizing secondary structure assignments or a number of distance restraints proved useful in several coarse-grained simulations.[15–26] Obviously, such hierarchical folding algorithms are useful if (1) a sequential folding mechanism is valid for the investigated protein, and (2) information about the preformed secondary structure elements, or more precisely the identity of relatively rigid segments formed at early folding stage and maintained throughout the folding pathway, is available.

The present study is composed of three parts: (1) generating a complete set of conformations for each protein by rotating the virtual C_α-C_α bonds located in the flexible regions, (2) sorting out the most "native-like" conformations, referred to as probable folds (PFs), on the basis of the database extracted criteria for the overall molecular dimensions and inter-residue potentials, and (3) analyzing the PFs in comparison with X-ray or nuclear magnetic resonance (NMR) structures.

In addition to an assessment of the discriminative ability of the particular coarse-grained method and knowledge-based potentials, the present analysis may give insights into other issues related to protein folding. For example, is the energy minimum at the native state deep and broad enough to be recognized despite moving in relatively large steps over the energy landscape? Such a trait was discerned in our previous enumeration of the conformations accessible to ROP monomer,[26] but we needed further evidence for clarification. Another point of interest is to understand the limitations of a sequential folding mechanism, and to identify, if any, chain segments that invariably recognize the correct tertiary fold in all generated PFs and might therefore act as nuclei or core regions. The present method and results will therefore be analyzed from these perspectives.

MODEL AND METHOD

Dataset

The proteins considered in the present study are presented in Table I. The same set was previously considered by Park and Levitt[2] in their simulation of protein conformations. The PDB identifiers, the size, and the structural class of each protein are listed in the table. These are structurally distinct, except for 434 cro protein and 434 repressor, which are homologous but distant in sequence.

Model

The virtual bond model originally proposed by Flory and collaborators[27] is adopted for representing the backbone. A backbone of a protein of n residues is therefore represented by $n - 1$ virtual bonds. l_i is the length of the virtual bond connecting the ith α-carbon, C_{α_i}, to the (i + 1)st, $C_{\alpha_{i+1}}$. θ_i is the bond angle between l_i and l_{i+1}, and ϕ_i is the torsional angle defining the rotation about bond l_i. The sidechains are each represented by a single interaction site, S_i, specific to the type of the amino acid. S_i is determined from the centroid of either all sidechain atoms, or a few specific ones, depending on the type of the amino acid, as previously described.[6,28] The sidechain virtual bond of length l_{S_i} connects C_{α_i} to S_i. θ_{si} is the bond angle between l_i and l_{S_i}, and ϕ_{S_i} is the sidechain dihedral angle defined by the three consecutive bonds l_{i-1}, l_i and l_{S_i}. The set of geometric variables $\{l_{i-1}, \theta_{i-1}, \phi_{i-1}, l_{S_i}, \theta_{S_i}, \phi_{S_i}\}$ completely describes the position of the ith residue, provided those of the preceding two α-carbons are known.

Conformational Sampling Technique

Conformations are generated by rotating the backbone virtual bonds at fixed intervals within their full range ($-180° \leq \phi_i \leq 180°$), while the remaining geometric variables are held fixed at their native state values. Clearly, complete enumeration of all bond rotations is not technically feasible. Intervals of 30°, for example, lead to $\nu = 12$ rotational states per residue, and consequently $\nu^{n-2} \approx 12^{52}$ states to be enumerated even for the smallest protein of 54

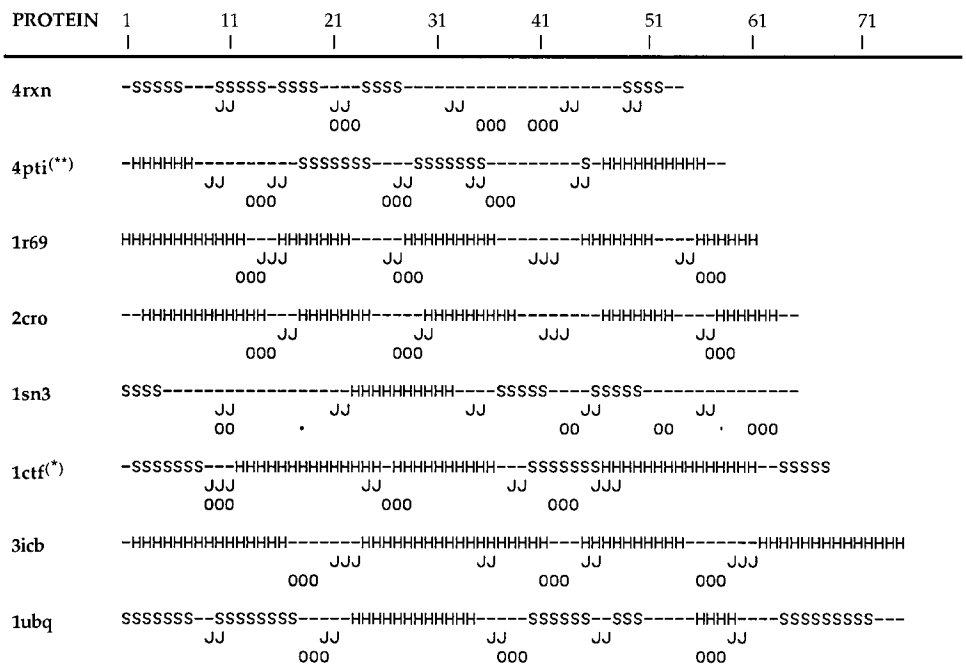

Fig. 1. Secondary structure and flexible segments of the proteins considered in the present study (Table 1). The secondary structure is indicated by S: β-strand, H: α-helix, or −: coil or turn. Flexible residues marked with the symbol "J" below the secondary structure are taken from the study of Park and Levitt.[2] Those presently determined with the Gaussian network model are marked with an "0" on the third row for each protein. *The sequence number in the PBD file is assigned as 1–33, 86–107. Here we used sequential numbers between 1 and 68. **Secondary structure assignment conforms with the refined crystal structure reported by Wlodawer et al.[53]

residues in our set. A significant reduction in conformational space is achieved by assuming that some segments are preformed at early stages of folding and possess sufficient internal stability to be held rigid in simulations. This approach has proved useful in recent simulations.[2,17,20,21,23,26] A subset of ten rotatable bonds, generally located in loop regions, may, for example, be conveniently chosen to enumerate $\nu^{10} \approx 10^{11}$ conformations for each protein. With this simplification, each protein is divided into five or six rigid segments, separated by four or five hinges composed of two or three "flexible" residues.

The proper selection of flexible residues is critically important for the success of the method, as will be elaborated below. In view of this feature, we performed our simulations using two sets of flexible residues for each protein. The former is taken to be identical to that previously proposed by Park and Levitt.[2] Figure 1 displays the partitioning of the test proteins into such regions. The symbols H and S refer to helix or β-strand regions, while the coiled regions are indicated by the dashed lines. The flexible residues previously identified[2] using a dynamic programming algorithm for aligning sequences are indicated by the letters J on the second line for each protein.

An alternative set of flexible residues, indicated by the letters O, is given on the third line for each protein. The flexible residues in this set are determined using the Gaussian network model (GNM) of proteins.[29,32] The GNM allows expression of the dynamics of the folded protein in terms of a collection of vibrational modes. The most flexible residues are identified by extracting the slowest (or largest amplitude) modes and examining the residues most strongly affected by these modes. These are generally found to occupy the loop regions or turns between α-helices or β-strands, or the helix termini. In the following, calculations performed with the flexible residues defined by Park and Levitt[2] will be referred to as set (a) and those using the residues identified by the GNM as set (b). Thus, 16 different cases will be analyzed, i.e., eight proteins with two sets of flexible residues each, which will be shortly designated with the PDB identifiers followed by the suffix (a) or (b).

Screening Criteria for Generating Native-Like Folds

To reduce the ensemble size and to obtain native-like structures, we applied two screening tests before proceeding with the energetic evaluation of the conformations: (2) conformations that violate the excluded volume principle are discarded; a threshold value of 2.0 Å is adopted for the closest distance of approach between two interaction sites; and (2) the

TABLE II. Properties of Probable Folds (PFs) Generated for Each Protein†

	Set (a)			Set (b)		
PDB code	No. of PFs	⟨E/nRT⟩	⟨d-rmsd⟩ (Å)	No. of PFs	⟨E/nRT⟩	⟨d-rmsd⟩ (Å)
4RXN	8,104	−2.98	5.41	7,237	−2.09	6.20
4PT1	7,955	−2.42	6.02	8,003	−2.09	3.81
1R69	8,024	−3.37	5.38	7,910	−4.56	2.11
2CRO	8,003	−3.22	4.98	8,069	−4.49	2.52
1SN3	8,077	−2.47	5.51	8,069	−3.10	5.01
1CTF	7,870	−3.28	6.18	7,980	−4.30	5.44
3ICB	3,845	−2.28	6.93	7,904	−2.17	6.23
1UBQ	1,925	−1.79	6.10	1,904	−1.62	6.75

†Set (a) refers to the simulation results obtained using the flexible residues identified by Park and Levitt and set (b) to those determined by the Gaussian network model.

radius of gyration R_g of a protein segment of n residues is required to obey the empirical expression[6]

$$\log R_g^2 = (2/3) \log n + 0.92 \quad (1)$$

within an error limit of $\Delta[\log R_g^2] = \pm 0.2$. In simulations, pairs of rigid segments connected by a middle flexible region, referred to as subchains, were considered separately, at the first step, and the local conformations satisfying the above two criteria were sorted out. An optimization of the flexible dihedral angles up to 10° resolution was performed at this step. At the next step, the resulting optimized subchains were combined by adding one subchain at a time to a growing group of subchains. The newly generated conformations at each step were again filtered on the basis of the above two criteria.

Beyond the addition of the fourth subchain, large numbers ($\geq 10^6$) of accessible conformations were encountered, which were then subjected to a third screening criterion based on energetics: only those conformations whose overall non-bonded energies were more favorable than −1.0 RT per residue were retained at this stage. This criterion, consistent with our analysis of potentials of mean forces stabilizing protein structures,[6] eliminated the conformations subject to repulsive or weakly attractive potentials. When the number of accepted folds was still substantially large, the energy requirements was rendered more severe so as to end up with a set of about 8,000 PFs for each protein (Table II).

Conformational energies were evaluated by using the knowledge-based potentials recently extracted[6] from PDB structures. These account for the interactions between sites separated by at least five virtual bonds along the chain sequence. They include three contributions determined as a function of the separation r_{ij} at 2.0 Å distance intervals[6]

$$E = \sum_{i=1}^{n-3} \sum_{j=i+3}^{n} E_{SS}(r_{ij}) + \sum_{i=1}^{n-4} \sum_{j=i+4}^{n} E_{SB}(r_{ij}) + \sum_{i=1}^{n-5} \sum_{j=i+5}^{n} E_{BB}(r_{ij}). \quad (2)$$

Here $E_{BB}(r_{ij})$ is the potential between backbone (B) sites $C_{\alpha i}$ and $C_{\alpha j}$, $E_{SS}(r_{ij})$ is the one between sidechains S_i and S_j, and $E_{SB}(r_{ij})$ refers to sidechain (S) and backbone (B) sites of residues i and j. The former is invariant with respect to residue type, whereas the latter two are residue specific. The complete set of energy parameters is available on the internet (http://klee.bme.boun.edu.tr/supplementarydata.html).

RESULTS AND DISCUSSION

Energy and Geometry Characteristics of the PFs

The number of PFs generated for each protein are presented in Table II, along with their mean energies and rms deviation with respect to X-ray structure. Columns 2–4 refer to the results obtained using the set (a) of flexible residues, and columns 5–7 refer to those from set (b) (Fig. 1). < E/nRT > values represent the mean energies of the PFs. The angular brackets designate the average over all PFs found for a given protein with a given choice of flexible residues. The large negative values (per residue) demonstrate that the PFs are highly favorable from an energetic point of view. The quantities ⟨d-rmsd⟩ in columns 4 and 7 refer to the distance rms deviation with respect to X-ray structure, again averaged over the set of PFs generated for each case. For a given PF, the distance rms deviation is found from $\Sigma_i \Sigma_j [r_{ij} - r_{ij}^0]^2/m)^{1/2}$ where r_{ij} and r_{ij}^0 are the distances between sites i and j in the PF and the corresponding X-ray structure, respectively, and m is the total number of pairs included in the double summation. We note that the distance rms deviations are lower than the coordinate rms deviations by a factor of 1–3. As illustrated below for a few cases, distance rms deviations of the PFs lie in the range 0.5 Å ≤ d-rms ≤ 10 Å, in general, leading to the tabulated <d-rmsd> values of 5–7 Å in general.

Energies Vs. rms Deviations With Respect to Native Fold

A detailed analysis of the potential energies of individual PFs as a function of rms deviations reveals a tendency for the energy to decrease, in general, as the rms deviation decreases, i.e., as the native fold is approached. This trend is not necessarily a smooth decrease, however. Figure 2 illustrates the results for four example cases, L7/L12 ribosomal

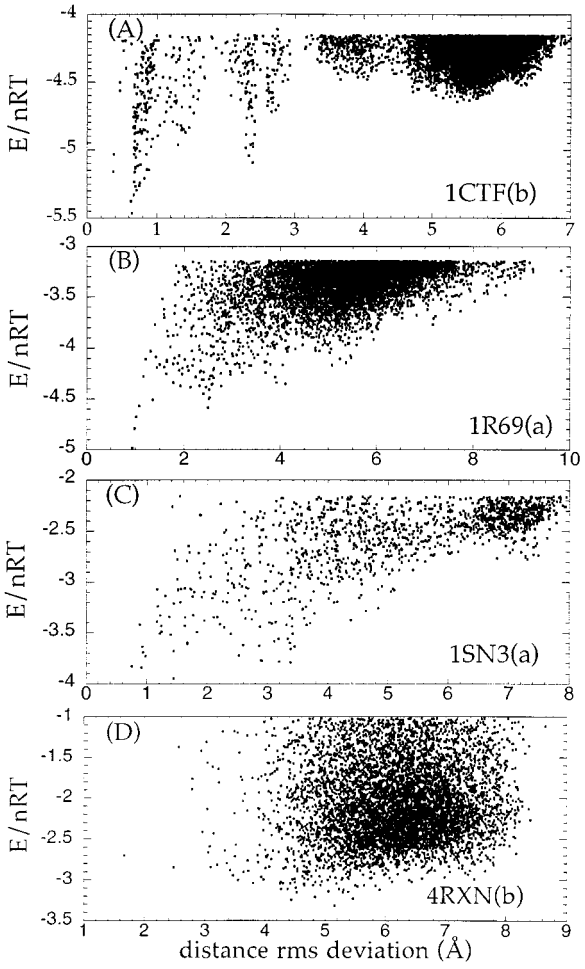

Fig. 2. Energies of the generated probable folds (PFs) plotted as a function of distance rms deviations from X-ray structure illustrated for (**A**) 1CTF(b), (**B**) 1R69(a), (**C**) 1SN3(a), and (**D**) 4RXN(b). See Table 1 for description of the proteins. There exists a trend towards a decrease in energy with decreasing rms deviation from X-ray structure.

protein sequence analyzed with the set (b) of flexible residues, 1CTF(b); N-terminal domain of phage 434 repressor with set (a), 1R69(a); scorpion neurotoxin with set (a), 1SN3(a); and finally rubredoxin with set (b), 4RXN(b).

Interestingly, multiple minima are observed in Figure 2A at different d-rms deviations. Such multiple minima were also observed in 4PTI(a-b), 1R69(b), and 1SN3(b). Except for 4PTI(a), the deepest minimum was observed to occur at the lowest d-rms region in all cases. The occurrence of multiple minima is consistent with the presence of local wells on the highly structured energy landscapes typically attributed to proteins' conformational space. The energy traps such as that occurring near the d-rms deviation of 2.5 Å in 1CTF(b) indicates the possible formation of stable but misfolded structures during the folding pathway.

In Figure 2B and C, on the other hand, a more uniform distribution of energies as a function of d-rms is observed. In general, there is a decrease in energy as the rms deviation from the native fold is diminished. A similar behavior was observed in 4RXN(a), 3ICB(a-b), 2CRO(a-b), and 1UBQ(a-b). The only case that exhibited a completely different trend was 4RXN(b), where rather scattered data were observed (Fig. 2D). This and 4PTI(a) were, in fact, the only two among the 16 cases examined that failed to recognize the native fold, as will be described below.

We note that in previous examinations of well-constructed decoys, there was not a perceptible trend towards a decrease in energy with decreasing rms deviation from X-ray structure.[2] To our knowledge, this is the first time a clear decrease in energy of well-constructed decoys has been observed at lower rms values. The energetics of a large number of competitive plausibly misfolded structures of myoglobin, 1CTF, and 1R69 were analyzed by Monge et al. using both low-resolution and all-atom models.[21] The rms deviation versus total energy plots for the low-resolution models were rather scattered, like that in Figure 2D, exhibiting a hardly distinguishable trend for the energy to decrease with decreasing rms deviation from X-ray structure. In particular, the energy of the native fold for 1R69 was found to be very high relative to misfolded structures. The present approach, however, permits a reasonable classification of the set of highly favorable, compact structures as to their increasing similarity to native fold on the basis of their conformational energies.

Comparison of the Lowest Energy Decoys With X-Ray Structures

The lowest energy decoy generated for each protein, referred to as the most probable fold (MPF), is compared with the X-ray structure in Table III. The third and fifth columns are the potential energies of the MPFs obtained using the sets (a) and (b) of flexible bonds. These may be compared with the potential energies of the X-ray structures (2nd column), found using the same model and energy parameters.[6,30] We note that the MPFs have generally higher energies by about 1.0 RT compared with X-ray structures. This is a consequence of the coarse-grained sampling of the conformational space (at 10° interval torsional angles) in our approach, and the ensuing absence of optimization of interactions as efficient as in the native fold. The fourth and sixth columns are the coordinate rms (c-rms) deviations of the MPFs from X-ray structures. These are found from c-rms $= (\Sigma_i |\mathbf{r}_i - \mathbf{r}_i^0|^2/n)^{1/2}$ where $\mathbf{r}_i$ is the position vector of the ith backbone site in the MPF, and $\mathbf{r}_i^0$ is its counterpart in the X-ray structure, provided that the two have been optimally superimposed.[31]

TABLE III. Comparison of the Lowest Energy Decoys With X-Ray Structures†

		Set (a)		Set (b)	
PDB name	E/nRT (X-ray)	E/nRT (simulations)	c-rms deviation (Å)	E/nRT (simulations)	c-rms deviation (Å)
4RXN	−4.37	−3.86	1.91	−3.32	8.08
4PTI	−4.05	−3.21	6.63	−3.45	1.53
1R69	−5.58	−4.98	0.99	−5.30	0.63
2CRO	−5.35	−4.59	1.17	−4.98	0.75
1SN3	−4.45	−3.94	1.77	−3.48	2.92
1CTF	−5.51	−4.58	1.05	−5.46	0.89
3ICB	−4.97	−4.36	0.76	−3.97	2.26
1UBQ	−5.08	−4.02	1.40	−3.66	1.36
Average‡	−5.62	−4.18	1.29	−4.33	1.48

†The lowest energy decoy obtained in simulations is referred to as most probable fold (MPF) for each protein.
‡Excluding the outliners 4PTI in set (a) and 4RXN in set (b).

The c-rms deviations of the MPFs from X-ray structures are below 3.0 Å, for all proteins, except for 4RXN(b) and 4PTI (a). On average, the MPFs exhibit an rms deviation of 1.4 Å from the X-ray structures, and their intramolecular potential amounts to −4.2 RT per residue, as presented in the last row of Table III. The c-rms deviations are much lower than those obtained by Park and Levitt[2] for the same set. Thus rms deviation values above 5.7 Å were reported for the lowest energy conformations. This departure between the two results may be attributed to the following differences in the model and method: (2) the present model contains two sites per residue, one on the sidechain and the other on the backbone, while that adopted previously contained only the backbone α-carbons; the existence of a sidechain rigidly appended to each α-carbon effectively constrains the conformational space and selects relatively more favorable conformations; (2) a more complete scanning of the conformational space is performed here by varying the dihedral angles at 30° intervals, originally, and then optimizing each at 10° intervals, compared with the assignment of one of four isomeric states to each rotatable bond; and (3) distance-dependent inter-residue potentials extracted at 0.4 Å resolution are used to evaluate the conformational energies, as opposed to the approximate distance-dependent versions of "on-off" contact potentials.

Figure 3 illustrates the comparison of the backbone structures of the MPFs with the corresponding X-ray structures, for a few representative examples among the 14 cases that closely reproduce the X-ray structure. The two cases that fail to recognize the X-ray structure are 4RXN(b), and 4PTI (a). Their respective c-rms deviations are 8.08 and 6.63 Å. The results for 4PTI are drastically improved (c-rms deviation = 1.53 Å), when the set (b) of flexible residues deduced from the GNM approach[29,32] are adopted. Conversely, the use of set (a) for 4RXN reduces the c-rms deviation to 1.91 Å. The strong dependence of the results for these two proteins on the choice of flexible residues is illustrated in Figure 4.

This analysis suggests that fixing an inappropriate set of residues in a hierarchical simulation algorithm leads to misfolded structures. These structures exhibit quite favorable, but non-native-like inter-residue interactions. In our simulations, the conformations of small portions of sequentially contiguous segments were optimized and gradually packed together. Improper choice of such segments, like incorrect structure formation on a local scale in experiments, could thus lead to misfolded structures, unless a mechanism for the dissociation of these originally misfolded regions operates at later stages of folding. Such a mechanism being absent in our simulations, the correct folded states were not captured in two cases, 4RXN(b) and 4PTI(a).

Decoys Exhibiting the Lowest c-rms Deviations From X-Ray Structures

As a further test, we concentrated on the PFs exhibiting the lowest rms deviation from the crystal structure. The aim was to visualize their energy rank among the set of PFs generated for each protein. The results are presented in Table IV. Here, the lowest c-rms deviations attained in the generated PFs, their energies expressed in RT units, per residue basis, and their rank on the basis of total energies, are listed for the two sets (a) and (b) of flexible residues. The ranks of 1R69, 3ICB, 1CTF, and 1UBQ in set (a), and 1UBQ and 2CRO in set (b) are 1. Therefore, the most native-like fold insofar as the structural similarity to crystal fold is concerned, is also the one having the lowest total energy, which supports the view that the native structure is also a thermodynamic equilibrium state. In the case of 3ICB (b), the energy rank is 2, and those of 2CRO and 4RXN set (a) are 3.

These results support the use of the present model and energy parameters for recognizing the native fold. The most native-like fold always lies in the top

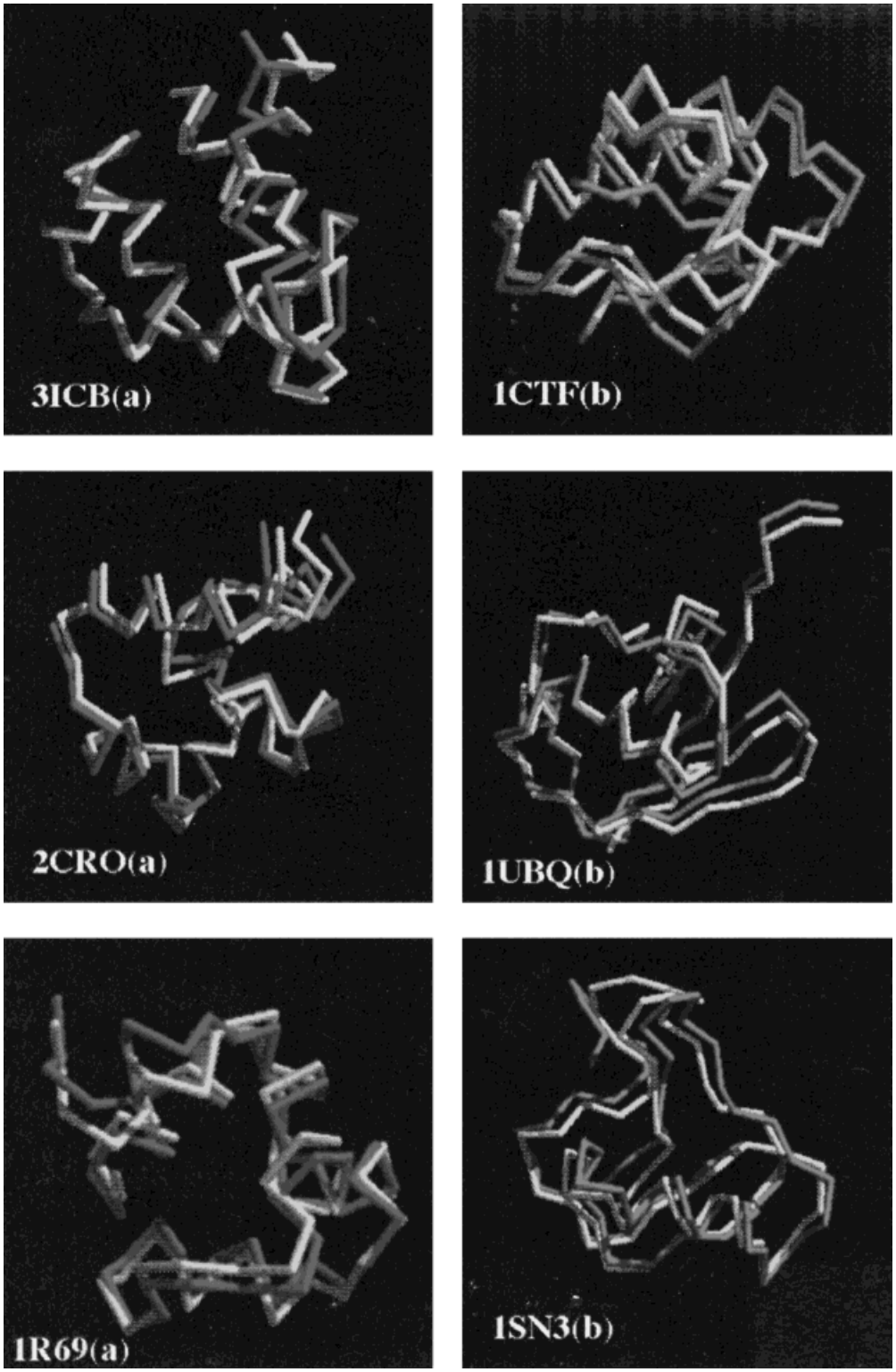

Fig. 3. Comparison of the lowest energy conformations obtained in present simulations with corresponding X-ray structures. Results are displayed for 3ICB(a), 2CRO(a), 1R69(a) on the left, and 1CTF(b), 1UBQ(b), and 1SN3(b) on the right. The α-carbon trace of the X-ray structure is shown in white in all cases, while the backbone structures predicted with the set of flexible residues (a) and (b) are shown in gray.

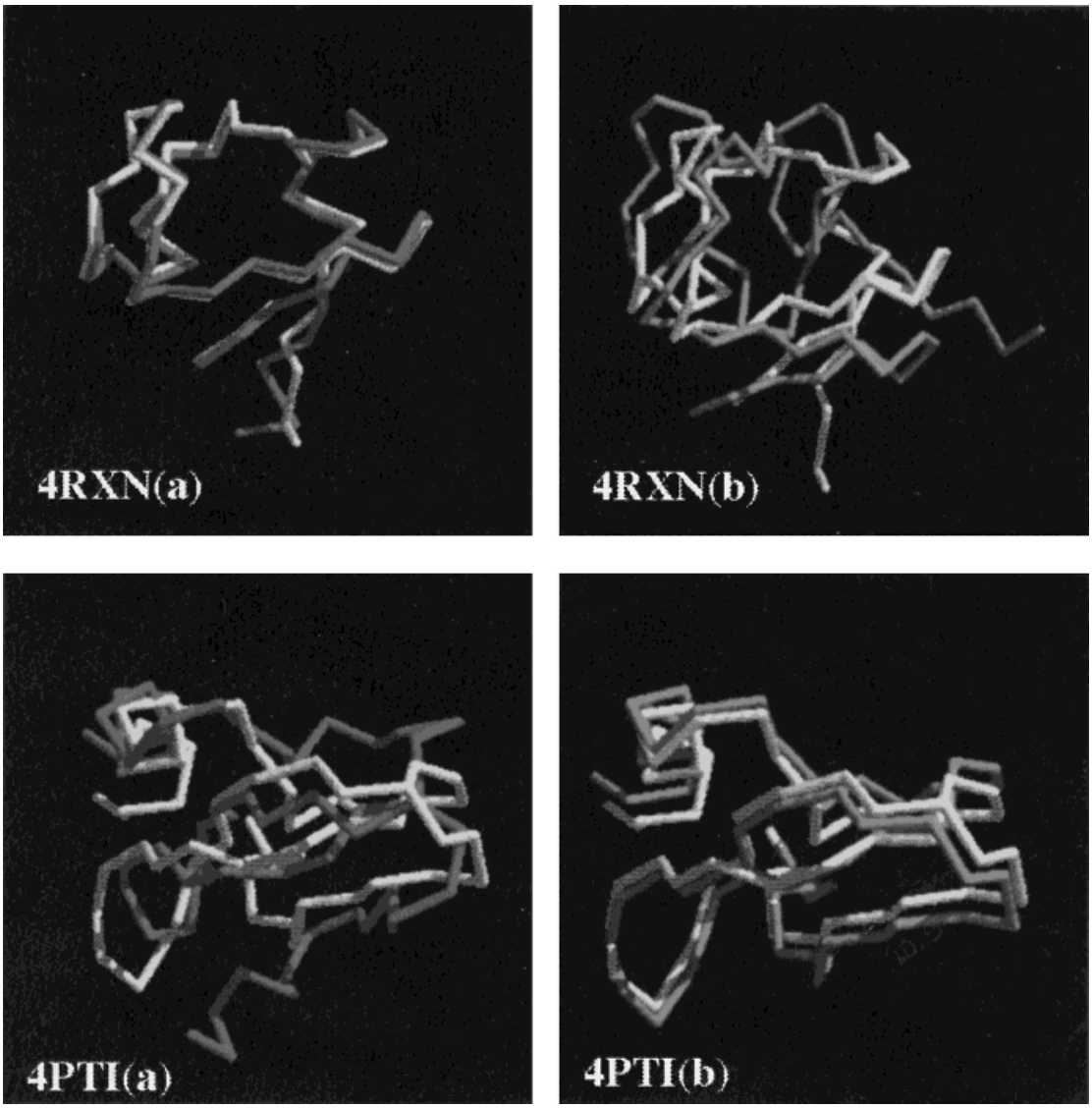

Fig. 4. Backbone structures of the lowest energy conformations generated for 4RXN **(upper panels)** and 4PTI **(lower panels),** using the set (a) and (b) of rotatable bonds, all of them optimally superposed on the corresponding X-ray structures shown in white. The darker backbone traces refer to the results from set (a) and (b).

~600 among the approximately 8,000 PFs generated for each protein, sorted with respect to their energies. It is to be noted that prior to the energy screening criterion, the number of generated folds was above 10^6 for each protein.

No energy gap was observed between the most native-like fold and the other PFs in our simulations. The same observation was made by Covell and Jernigan[33] in their simulations of conformations on lattices with predefined distribution of cells. In that study, which is one of the earliest examples of threading simulations, the native structure was found to lie within the top 1% of the energy-sorted conformations.

Identification of Local Stabilization Regions

In a recent insightful review, Finkelstein[9] pointed out that structural features common to energetically favorable conformations can be identified by the statistical examination of the ensemble of predicted tertiary structures, instead of concentrating on a single conformation. Following this approach, we examined the probability distribution of the rotatable dihedral angles (ϕ_i) and determined the most favorable torsional states selected in the energetically most favorable conformations. Interestingly, some flexible bonds exhibited a very pronounced preference for native torsional angles, which may be

TABLE IV. Ranks and Energies of the PFs Exhibiting the Lowest c-rms Deviation From X-Ray Structure†

	Set (a)		Set (b)	
PDB name	Lowest c-rms deviation (Å) (a)	Energy rank (a)	Lowest c-rms deviation (Å) (b)	Energy rank (b)
4RXN	0.77	3	1.76	385
4PTI	1.65	35	1.05	271
1R69	0.99	1	0.52	17
2CRO	1.1	3	0.48	1
1SN3	0.47	25	0.90	596
1CTF	1.05	1	0.70	10
3ICB	0.76	1	1.41	2
1UBQ	1.40	1	1.36	1

†Among the 8,000 PFs generated for each protein, with each set of flexible residues.

considered as evidence for local stability near these residues. This interpretation is only tentative, in view of the coarse-grained nature of our analysis, yet some agreement between theoretical results and experiments is observed, which suggests the possible utility of such statistical analyses of local conformational stability.

Figure 5 illustrates the results for 4PTI and 3ICB, both of which contain flexible bonds with sharp preferences for certain torsional states, far above those expected from random probability distributions. Here the PFs generated with the set (a) of flexible residues are considered. The angle distributions are presented for a few bonds only, for clarity. For example, in 4PTI, among the ten rotatable bonds of set (a), Lys26-Ala27 and Ala27-Gly28 are distinguished by their strong preference for the torsional angles 40° and 230°, respectively, as may be observed in Figure 5A. The corresponding torsional angles in the native state are 39.3° and 227.5°. Together with the segments that are held fixed on both sides of these residues, a locally stabilized region extending between Arg17 $\leq$ i $\leq$ Thr32 is implied. The latter is in black in the α-carbon trace shown in the inset. This region includes the first β-strand of the protein, which is the most stable region of the inhibitor as indicated by experiments[34–38] and theoretical[39,40] studies.

Likewise, examination of Figure 5B reveals the strong preference of bonds Thr34-Glu35 and Glu35-Phe36 of calbindin D_{9k} for the torsional angle 20°. These again coincide with the torsional angles (19.6° and 22.2°) of the respective virtual bonds in the native state. The segment Ser24 $\leq$ i $\leq$ Thr45, limited by the flexible bonds Gln22-Leu23 and Leu46-Asp47, which do not exhibit a strong preference for any well-defined torsional state (Fig. 5B), therefore appears as a stable region. This region includes the helix II and the single helical turn (Ser38 $\leq$ i $\leq$ Lys41) of calbindin $D_{9k,}$ which stabilize the mobile linker between helices II and helices III.[41,42] This linker connects the N-terminal and C-terminal EF-hand motifs of the protein. The present analysis thus indicates that the N-terminal EF-hand is more stable than the C-terminal EF hand. This supports previous NMR observations of the solution structure of apo calbindin D_{9k}, in which the Ca^{2+} binding of the two EF-hands was noted to be highly asymmetric: the backbone rms deviation for the C-terminal hand was nearly double that of the N-terminal hand, which was attributed to the preformation of the N-terminal hand.[43]

A similar analysis was done for all proteins presently examined. No significant preference for native state dihedral angles could be observed in 1R69 and 1UBQ, using either set of flexible residues; all rotatable bonds were found to obey relatively broad or multimodal distributions. Hydrogen exchange experiments indicate the absence of formation of any specific, stable secondary structure in the folding pathway of ubiquitin (1UBQ),[44] which supports our observation. The remaining three proteins revealed the following regions with observable conformational preferences: Thr18 $\leq$ i $\leq$ Lys40 in 2CRO, Glu23 $\leq$ i $\leq$ Thr55 in 1SN3, and Asn12 $\leq$ i $\leq$ Ala37 (Ala 90 in PDB) in 1CTF. The indicated region in phage 434 Cro includes two helices buried in the hydrophobic core.[45] That in neurotoxin (1SN3) consists of two β-strands and an α-helix linked to the strands in the hydrophobic core,[46] whereas in 1CTF, an α-turn-α motif is identified. More experimental data and detailed atomic analyses are needed for an assessment of the validity of the proposed regions of enhanced local stability.

CONCLUSIONS

The following conclusions are drawn from the present study:

Kinetically Favorable Shape of the Conformational Energy Landscape Near the Native State

The energy minimum associated with the native structure is broad and deep enough to be recognized upon sampling of conformational space at 30° interval rotations of C_α-C_α virtual bonds, conforming with our previous results for ROP monomer.[26] Due to computational limitations, a finite number of bonds were rotated here similarly to the approach adopted by Park and Levitt.[2] Two alternative sets of rotatable bonds (a) and (b), -one proposed in an earlier study,[2] and another based on the Gaussian network model of proteins,[29,32] were considered for an assessment of the validity of the results irrespective of the choice of the flexible segments. The conformations having rotational angles close to native state dihedrals passed the screening tests in most proteins examined by either set of rotatable bonds, i.e., the majority of

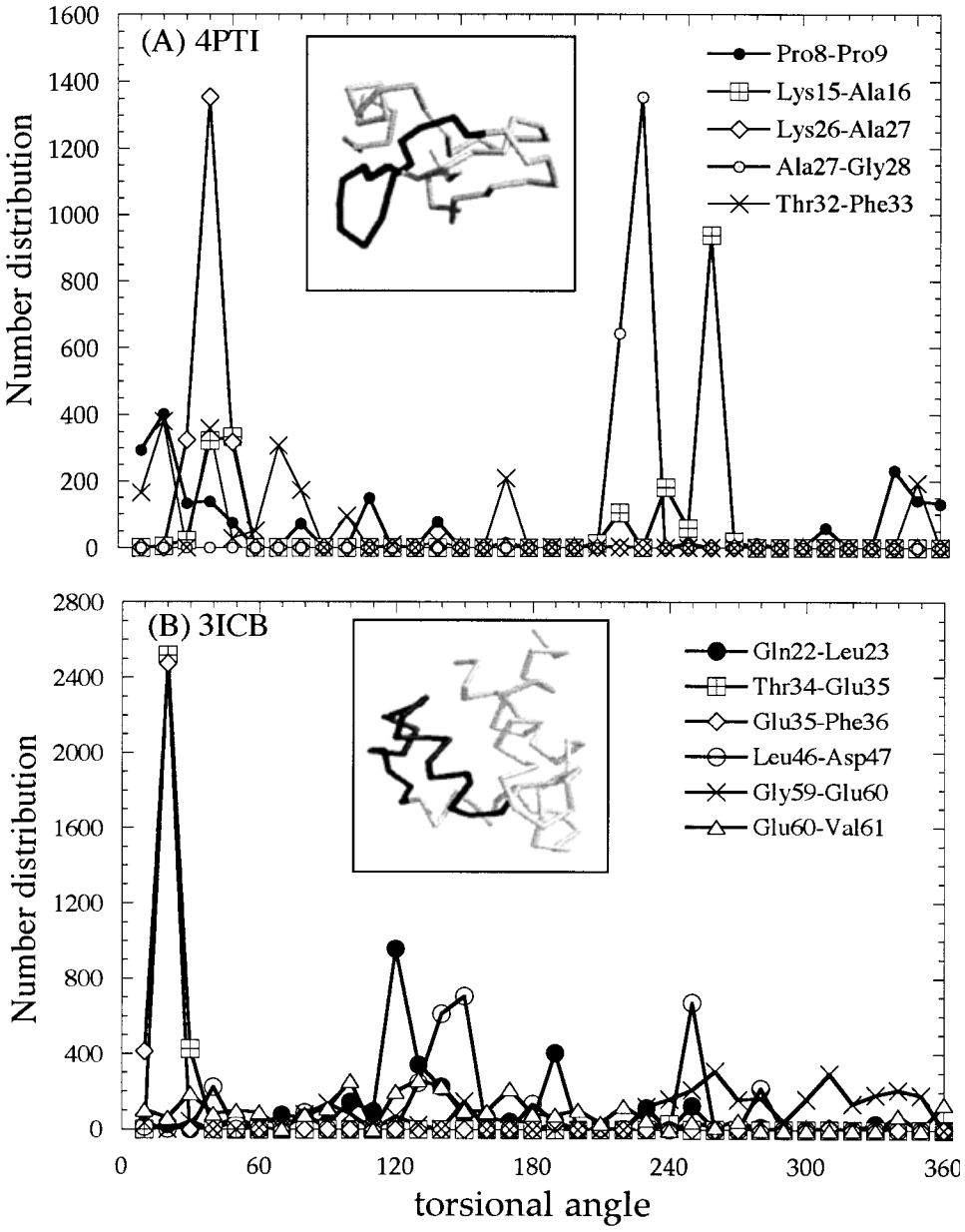

Fig. 5. Distribution of rotational angles (ϕ_i) for a subset of flexible bonds of (**A**) bovine pancreatic trypsin inhibitor (4PTI) and (**B**) calbindin D_{9k}(3ICB). The bonds are indicated on the figure. Pronounced preferences for native state rotational angles are distinguished for bonds Lys26-Ala27 and Ala27-Gly28 in A, and bonds Thr34-Glu35 and Glu35-Phe36 in B. These bonds and the rigidly held segments on both sides are shown in black in the inset. The remaining portions are shown in gray.

native-like conformations were included in the reduced set of well-constructed decoys.

Suitability of Database Extracted Inter-Residue Potentials and Screening Criteria for Discriminating the Native Structure

The energy parameters, as well as the size and energy constraints adopted in screening tests, were recently derived[6] from databank structures using the same virtual bond model as the one presently adopted. Using either set (a) or (b) of flexible bonds, seven of the eight proteins examined are shown to recognize the native fold with an rms deviation of about 1.4 Å (Table III), which is considerably lower than those observed in previous coarse-grained simulations. One protein in each set was observed, on the other hand, to be misfolded into an alternative energetically favorable structure.

Improper Choice of Preformed Segments Leads to Misfolded Structures

To reduce the ensemble size, chain portions composed of flexible bonds flanked by rigidly held seg-

ments were combined gradually, using screening criteria that ensure native-like geometry and energy characteristics. The conformational states closest to the X-ray structure in our coarse-grained space were overlooked in this hierarchical procedure, in two cases: 4PTI(a) and 4RXN(b). The fact that for the same proteins, the use of the alternative sets of flexible bonds satisfactorily led to the recognition of the X-ray structure, as illustrated in Figure 4, invites attention to the important role of the correct formation of foldons in early folding phases, for the proper evolution towards the native structure. Restriction of the conformational space, or more precisely the folding pathways, by fixing a number of degrees of freedom may thus totally obstruct the passage to the correctly folded state. A sequential folding mechanism is useful in computer simulations only if the preformed structural elements are accurately identified.

Conformational Energy Decreases as the rms Deviation From X-Ray Structure Decreases

This pattern, observed in a distinct way for the first time in coarse-grained simulations, is illustrated in Figure 2.

A more critical test reveals that in 6 of the 16 cases examined, the structure with the lowest rms deviation from X-ray structure coincides exactly with the lowest energy fold (Table IV), while it lies in the top 600 of the energy-sorted list of about 8,000 probable folds extracted for each protein starting from more than 10^6 original conformations in each case. In previous low-resolution simulations, there was little or no obvious correlation between energy and rms deviation.[2,21] Park and Levitt suggested that the best energy functions will only start to discriminate effectively when structures close to the correct conformation can be examined. The present set of structures indeed exhibit highly favorable energetics and native-like geometry in general (Table II). Therefore, the fact that a correlation between energy and rms deviation could be observed in the present set supports the original conjecture of Park and Levitt[2].

A Pronounced Preference for Native Dihedral Angles Is Observed in Some Rotatable Bonds

This enhanced preference emerges from the analysis of the probability distribution of dihedral angles in the ensemble of well-constructed decoys, as recently suggested.[7–9] Segments with enhanced propensity for conformational states that coincide with the native structure may be viewed as relatively stable regions possibly correctly folded at early folding stage. This hypothesis is supported by comparison of experimental results for bovine pancreatic inhibitor and calbindin D_{9k}. Further data, both theoretical and experimental, are needed, however, to strengthen this hypothesis.

REFERENCES

1. Godzik, A., Kolinski, A., Skolnick, J. Topology fingerprint approach to the inverse protein folding problem. J. Mol. Biol. 227:227–238, 1992.
2. Park, B., Levitt, M. Energy functions that discriminate X-ray and near-native folds from well-constructed decoys. J. Mol. Biol. 258:367–392, 1996.
3. Abola, E.E., Bernstein, F.C., Bryant, S.H., Koetzle, T.F., Weng, J. Protein Data Bank: A computer-based archival file for macromolecular structures. J. Mol. Biol. 112:535–542, 1987.
4. Jernigan, R.L., Bahar, I. Structure-derived potentials and protein simulations. Curr. Opin. Struct. Biol. 6:195–209, 1996.
5. Sippl, M.J. Knowledge-based potentials for proteins. Curr. Opin. Struct. Biol. 5:229–235, 1995.
6. Bahar, I., Jernigan, R.L. Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation. J. Mol. Biol. 266:195–214, 1997.
7. Finkelstein, A.V., Gutin, A.M., Badretdinov, A.Y. Perfect temperature for protein structure prediction and temperature. Proteins 23:142–150, 1995.
8. Shortle, D., Wang, Y., Gillepsie, J.R., Wrable, J.O. Protein folding for the realists: A timeless phenomenon. Protein Sci. 5:991–1000, 1996.
9. Finkelstein, V.A. Protein structure: What is it possible to predict now? Curr. Opin. Struct. Biol. 7:60–71, 1997.
10. Ptitsyn, O.B., Rashin, A.A. A model of myoglobin self-organization. Biophys. Chem. 3:1–20, 1975.
11. Cohen, F.E., Richmond, T.J., Richards, F.M. Protein folding: Evaluation of some simple rules for the assembly of helices into tertiary structures with myoglobin as an example. J. Mol. Biol. 132:275–288, 1979.
12. Cohen, F.E., Sternbeg, M.J.E., Taylor, W.R. Analysis and prediction of protein β-sheet structures by a combinatorial approach. Nature 285:378–382, 1980.
13. Wako, H., Scheraga, H.A. Distance-constraint approach to protein folding. I. Statistical analysis of protein conformations in terms of distances between amino acid residues. J. Prot. Chem. 1:5–45, 1982.
14. Wako, H., Scheraga, H.A. Distance-constraint approach to protein folding. II. Prediction of the three-dimensional structure of bovine pancreatic trypsin inhibitor. J. Prot. Chem. 1:85–117, 1982.
15. Hinds, D.A., Levitt, M. A lattice model for protein structure prediction at low resolution. Proc. Natl. Acad. Sci. USA 89:2536–2540, 1992.
16. Seitoh, S., Nakai, T., Nishikawa, K. A geometrical constraint approach for reproducing the native backbone conformation of a protein. Proteins 15:191–204, 1993.
17. Sun, S., Thomas, P.D., Dill, K.A. A simple protein folding algorithm using a binary code and secondary structure constraints. Protein Eng. 8:769–778, 1995.
18. Sun, S. Reduced representation model of protein structure prediction: Statistical potential and genetic algorithms. Protein Sci. 2:762–785, 1993.
19. Gunn, J.R., Monge, A., Friesner, R.A., Marshall, C.H. Hierarchical algorithm for computer modeling of protein tertiary structure: Folding of myoglobin to 6.2 Angstrom resolution. J. Phys. Chem. 98:702–711, 1994.
20. Monge, A., Friesner, R.A., Honig, B. An algorithm to generate low-resolution protein tertiary structures from knowledge of secondary structure. Proc. Natl. Acad. Sci. USA 91:5027–5029, 1994.
21. Monge, A., Lathrop, E.J., Gunn, J.R., Shenkin, P.S., Friesner, R.A. Computer modeling of protein folding: Conformational and energetic analysis of reduced and detailed protein models. J. Mol. Biol. 247:995–1012, 1995.
22. Aszódi, A., Gradwell, M.J., Taylor, W.R. Global fold determination from a small number of distance restraints. J. Mol. Biol. 251:308–326, 1995.
23. Srinivasan, R., Rose, G.D. LINUS: A hierarchic procedure to predict the fold of a protein. Proteins 22:81–99, 1995.
24. Lund, O., Hansen, J., Brunak, S., Bohr, J. Relationship

between protein structure and geometrical constraints. Protein Sci. 5:2217–2225, 1996.
25. Skolnick, J., Kolinski, A., Ortiz, A.R. MONSSTER: A method for folding globular proteins with a small number of distance restraints. J. Mol. Biol. 265:217–241, 1997.
26. Erman, B., Bahar, I., Jernigan, R.L. Equilibrium states of rigid bodies with multiple interaction sites: Application to protein helices. J. Chem. Phys. 107:2049–2059, 1997.
27. Brant, D.A., Flory, P.J. The configuration of random polypeptide theory. J. Am. Chem. Soc. 87:2791–2800, 1965.
28. Bahar, I., Jernigan, R.L. Angular distributions of nonbonded residues around central residues in globular proteins. Folding Design 1:357–370, 1996.
29. Haliloglu, T., Bahar, I., Erman, B. Gaussian dynamics of folded proteins. Phys. Rev. Lett. 79:3090–3093, 1997.
30. Bahar, I., Kaplan, M., Jernigan, R.L. Short-range conformational energies, secondary structure propensities, and recognition of correct sequence-structure matches. Proteins 29:292–308, 1997.
31. Kabasch, W. A discussion of the solution for the best rotation to relate two sets of vectors. Acta Crystallogr. A 34:827–828, 1978.
32. Bahar, I., Erman, B., Atilgan, A.R. Direct evaluation of thermal fluctuations in proteins using a single parameter harmonic potential. Folding Design 2:173–181, 1997.
33. Covell, D.G., Jernigan, R.L. Conformations of folded proteins in restricted spaces. Biochemistry 29:3287–3294, 1990.
34. Wagner, G., Wütrich, K. Amide proton exchange and surface conformation of basic pancreatic trypsin inhibitor in solution. Studies with two-dimensional nuclear magnetic resonance. J. Mol. Biol. 160:343–361, 1982.
35. Wagner, G., Stassinopolou, C.I., Wütrich, K. Amide-proton exchange studies by two-dimensional correlated ^{1}H-NMR in two chemically modified analogs of the basic pancreatic trypsin inhibitor. Eur. J. Biochem. 145:431–436, 1984.
36. Richarz, R., Sehr, P., Wagner, G., Wütrich, K. Kinetics of the exchange of individual amide protons in the basic pancreatic trypsin inhibitor. J. Mol. Biol. 130:19–30, 1979.
37. Kim, K.S., Tao, F., Fuchs, J.A., et al. Crevice-forming mutants of bovine pancreatic trypsin inhibitor: Stability changes and new hydrophobic surface. Protein Sci. 2:588–596, 1993.
38. Kim, K.S., Fuchs, J.A., Woodward, C.K. Hydrogen exchange identifies native-state motional domains important in protein folding. Biochemistry 32:9600–9608, 1993.
39. Wallqvist, A., Smythers, G.W., Covell, D.G. Identification of cooperative folding units in a set of native proteins. Protein Sci. 1627–1642, 1997.
40. Bahar, I., Wallqvist, A., Covell, D.G., Jernigan, R.L. Correlation between native state hydrogen exchange and cooperative residue fluctuations from a simple model. Biochemistry 36:13512–13523, 1997.
41. Groves, P., Linse, S., Thulin, E., Forsen, S. A calbindin D_{9k} mutant containing a novel structural extension: ^{1}H nuclear magnetic resonance studies. Protein Sci. 6:323–339, 1996.
42. Szebenyi, D.M.E., Moffat, K. The refined structure of vitamin D-dependent calcium-binding protein from bovine intestine molecular details, ion binding, and implications for the structures of other calcium-binding proteins. J. Biol. Chem. 261:8761, 1986.
43. Skelton, N., Kördel, J., Chazin, W.J. Determination of the solution structure of apo calbindin D_{9k} by NMR spectroscopy. J. Mol. Biol. 249:441–462, 1995.
44. Galdwin, S.T., Evans, P.A. Structure of very early folding intermediates: New insights through a variant of hydrogen exchange labelling. Folding Design 1:407–417, 1996.
45. Padmanabhan, S., Jimenez, M.A., Gonzales, C., Sanz, J.M., Gimenez-Gallego, Rico, M. Three-dimensional solution structure and stability of phage 434 Cro protein. Biochemistry 36:6424–6436, 1997.
46. Bugg, C.E. Structure of variant-3 scorpion neurotoxin from *Centruroides sculpturatus* Ewing, refined at 1.8 Å resolution. J. Mol. Biol. 170:493, 1983.
47. Watenpaugh, K.D., Sieker, L.C., Hensen, L.C. Crystallographic refinement of rubredoxin at 1.2 Å resolution. J. Mol. Biol. 138:615–633, 1980.
48. Marquart, M., Walter, J., Deisenhofer, J., Bode, W., Huber, R. The geometry of the reactive site and of the peptide groups in trypsinogen and its complexes with inhibitors. Acta Crystallogr. B 39:480, 1983.
49. Mondragon, A., Subbiah, S., Almo, C., Drottar, M., Harrison, S.C. Structure of the amino-terminal domain of phage 434 repressor at 2.0 Å resolution. J. Mol. Biol. 205:189, 1989.
50. Mondragon, A., Wolberger , C., Harrison, S.C. Structure of phage 434 Cro protein at 2.5 Å resolution. J. Mol. Biol. 205:179, 1989.
51. Leijonmarck, M., Liljas, A. Structure of the C-terminal domain of the ribosomal protein L7-L12 from *Escherichia coli* at 1.7 Å. J. Mol. Biol. 195:555, 1987.
52. Vijay, S., Kumar, Bugg, C.E., Cook, W.J. Structure of ubiquitin refined at 1.8 Å resolution. J. Mol. Biol. 194:531, 1987.
53. Wlodawer, A., Walter, H., Huber, R., Sjolin, L. Structure of bovine pancreatic trypsin inhibitor. Results of joint neutron and X-ray refinement of crystal form II. J. Mol. Biol. 180:301–329, 1984.