



Determining Protein Topology from Skeletons of Secondary Structures

Yinghao Wu¹, Mingzhi Chen², Mingyang Lu³, Qinghua Wang³ and Jianpeng Ma^{1,2,3*}

¹Department of Bioengineering
Rice University, Houston
TX 77005, USA

²Graduate Program of
Structural and Computational
Biology and Molecular
Biophysics, Baylor College of
Medicine, One Baylor Plaza
Houston, TX 77030, USA

³Verna and Marrs McLean
Department of Biochemistry
and Molecular Biology, Baylor
College of Medicine, One Baylor
Plaza, Houston, TX 77030
USA

We report a novel computational procedure for determining protein native topology, or fold, by defining loop connectivity based on skeletons of secondary structures that can usually be obtained from low to intermediate-resolution density maps. The procedure primarily involves a knowledge-based geometry filter followed by an energetics-based evaluation. It was tested on a large set of skeletons covering a wide range of protein architecture, including one modeled from an experimentally determined 7.6 Å cryo-electron microscopy (cryo-EM) density map. The results showed that the new procedure could effectively deduce protein folds without high-resolution structural data, a feature that could also be used to recognize native fold in structure prediction and to interpret data in fields like structure genomics. Most importantly, in the energetics-based evaluation, it was revealed that, despite the inevitable errors in the artificially constructed structures and limited accuracy of knowledge-based potential functions, the average energy of an ensemble of structures with slightly different configurations around the native skeleton is a much more robust parameter for marking native topology than the energy of individual structures in the ensemble. This result implies that, among all the possible topology candidates for a given skeleton, evolution has selected the native topology as the one that can accommodate the largest structural variations, not the one rigidly trapped in a deep, but narrow, conformational energy well.

© 2005 Elsevier Ltd. All rights reserved.

Keywords: secondary-structural skeleton; secondary structure assignment; topology; protein fold; geometry scoring

*Corresponding author

Introduction

In the field of structural biology, the mission of tackling increasingly complicated cellular systems has led to a reality that many structures, at least at early stages, can be obtained only at low to intermediate resolutions, at which only incomplete structural information can be obtained. Typical examples are seen in the measurements of cryo-electron microscopy (cryo-EM)^{1–10} and low-resolution protein crystallography. Therefore, advanced computational methods that can aid in

interpretation of structural information at intermediate resolutions are urgently needed.

In previous computational studies, major α -helices and β -strands can be efficiently located in intermediate-resolution density maps by methods such as Helixhunter,¹¹ Sheetminer,¹² and Sheettracer.¹³ The outputs of these methods outline the skeletons of secondary structures that describe the location of α -helices and β -strands. But, the information of the directionality of the secondary structures and loop connectivity among them is not available.

For a given skeleton of protein secondary structures, there are a large number of possible ways to connect the secondary structures by loops, i.e. there exist multiple topology candidates that can be assigned to a given skeleton (Figure 1). Among them, only one is the native topology selected during evolution for folding, stability and function. Therefore, the problem of determining native

Abbreviations used: cryo-EM, cryo-electron microscopy; MC, Monte Carlo; PDB, Protein Data Bank; GA, genetic algorithm.

E-mail address of the corresponding author:
jpma@bcm.tmc.edu

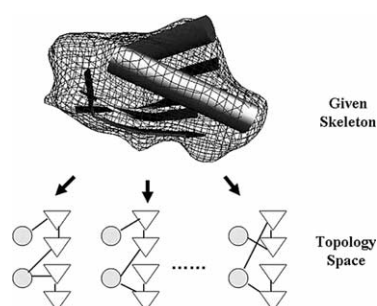


Figure 1. Schematic representation of protein topology space. For a given secondary-structural skeleton, there are a large number of possible topology candidates associated with it. Together they form a topology space. In the Figure, the skeleton is depicted in such a way that helices are drawn as cylinders and strands as ribbons. In the schematic diagrams of the topology space, the circles are for α -helices and triangles for β -strands.

protein topology, or fold, based on intermediate-resolution structural data becomes a problem of how to derive protein topology from secondary-structural skeletons.

To achieve the above goal, we have developed an energetics-based procedure for evaluating topology candidates. The method assigns protein sequence to the modeled C^α traces and energy is calculated using a knowledge-based pair-wise potential function.¹⁴ Moreover, to accelerate the screening, a complementary geometry-based analysis was also developed. It focuses on utilizing geometry scoring functions, based on knowledge extracted from a high-resolution protein structure database, to narrow down the number of possible topology candidates that a given protein skeleton is most likely to adopt. The output of geometry screening is used as input for final energetic screening.

The method has been extensively tested on secondary-structural skeletons of 50 medium-sized single-domain protein structures obtained from the Protein Data Bank (PDB). Among them 25 were all-helical proteins and 25 were sheet-containing proteins. They were also tested on skeletons in which one or more secondary structure was purposely removed and an eight-stranded skeleton obtained from an experimental 7.6 Å cryo-EM density map using Sheetminer¹² and Sheettracer.¹³ In most cases, the native topology was identified as the most energetically favorable one for a given skeleton. These results strongly suggest it is indeed plausible to derive reliably protein native topology from secondary-structural skeletons.

In our study, a major working hypothesis is that, for a given protein skeleton, its native topology is chosen by evolution to accommodate the largest structural variation, not merely the one trapped in a deep, but narrow, energy well. A natural implication of such an hypothesis is that the average energy of an ensemble of structures varying in the

vicinity of native skeleton would most likely be the lowest, and the standard deviation of the average energy would be the smallest. Our results seem to support the hypothesis very well and they also seem to indicate that the ensemble-averaging scheme is an effective way of compensating the inevitable errors in the artificially constructed structures and in empirical potential functions.

The sections are arranged here in such a way that the Introduction is followed by Methods for ease of understanding the new computational procedure. Some lengthy details are given in Supplementary Data. Readers who are only interested in the overall performance of the procedure can jump directly to Results.

Methods

Generation of topology candidates

Generation of all possible topology candidates was the first step for the new computational procedure and required two kinds of inputs: (1) the secondary-structural skeleton in three-dimensional space including the axes for α -helices and pseudo-traces for β -strands without information of connectivity (referred to hereinafter as skeleton); and (2) the secondary structure assignment on the primary amino acid sequence (referred to hereinafter as assignment).

To obtain skeletons, we first used 50 high-resolution crystal structures deposited in the PDB. From each crystal structure, sequence identity and directionality of all secondary structures and all the loops were eliminated altogether. In the case of β -sheets, an additional smoothing procedure was applied to minimize the potential bias from high-resolution structures (for the detailed procedure, see Supplementary Data). In real applications, skeletons can be obtained from intermediate-resolution density maps using advanced computational packages such as Helixhunter,¹¹ Sheetminer,¹² and Sheettracer.¹³ One such application presented here is an eight-stranded skeleton obtained from an experimental 7.6 Å cryo-EM density map using Sheetminer¹² and Sheettracer.¹³

Secondary structure assignment in sequence was obtained using the PSIPRED server[†].^{15,16} The number of predicted α -helices and β -strands was compared with that from the corresponding skeleton. A mismatch between them would incur a further evaluation using a consensus secondary-structure prediction server[‡]. A predicted assignment closest to the skeleton would be picked. Here, we gave higher credibility to skeletons because, even for skeletons derived from intermediate-resolution density map, they were

[†] <http://bioinf.cs.ucl.ac.uk/psipred>

[‡] http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_seccons.html

expected to contain more correct structural information than assignments predicted using most of the advanced secondary structure prediction algorithms currently available. However, we did test our procedure on skeletons in which a small number of secondary structures were misplaced.

Once we had skeleton in space and assignment in sequence in hand, we generated all possible topology candidates. It was done by aligning the predicted secondary structures in assignment with those in skeleton, i.e. α -helices against α -helices and β -strands against β -strands (Figure 2). It is easy to see that, for any skeleton with N^α α -helices and N^β β -strands, the total number of ways to thread the polypeptide through the skeleton was:

$$N_T = (N_\alpha! \times 2^{N_\alpha}) \times (N_\beta! \times 2^{N_\beta}) \quad (1)$$

The factorial terms came from permutation of all helices and all strands. The terms of 2^N came from the forward and backward directions of any secondary structure.

In cases where there were mismatches in the number of secondary structures between assignment and skeleton, the total number of ways to thread the polypeptide through skeleton was:

$$N_T = C_{N_{\alpha,Max}}^{N_{\alpha,Dif}} \times (N_{\alpha,Min}! \times 2^{N_{\alpha,Min}}) \times C_{N_{\beta,Max}}^{N_{\beta,Dif}} \times (N_{\beta,Min}! \times 2^{N_{\beta,Min}}) \quad (2)$$

in which:

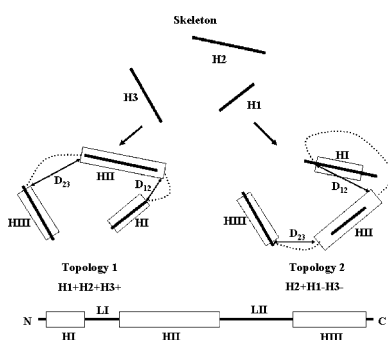


Figure 2. Alignment of secondary structures in assignment to skeleton in order to generate all the possible topology candidates. The skeleton shown consists of three helices; H1, H2 and H3. The corresponding sequence has been predicted to have three helices, HI, HII, and HIII, with the inter-helix loops named as LI and LII. In topology 1, HI, HII, HIII were aligned to H1, H2, H3, respectively, with correct directionality, yielding the topology as H1+H2+H3+. In topology 2, HI, HII, HIII were aligned to H2, H1, H3, respectively, with the direction of H1 and H3 being inverted, giving the topology as H2+H1-H3-. In each case, the length of the vectors D12 and D23 is compared with the maximal length of loops LI and LII in extended conformation in the initial screening in order to assess whether a particular topology is accessible for the given skeleton.

$$N_{\beta}^{\alpha, Dif} = \left| N_{\beta}^{\alpha, Sk} - N_{\beta}^{\alpha, Seq} \right|$$

$$N_{\beta}^{\alpha, Min} = \text{Min} \left(N_{\beta}^{\alpha, Sk}, N_{\beta}^{\alpha, Seq} \right)$$

$$N_{\beta}^{\alpha, Max} = \text{Max} \left(N_{\beta}^{\alpha, Sk}, N_{\beta}^{\alpha, Seq} \right)$$

$N^{\alpha, Sk}$, $N^{\beta, Sk}$, $N^{\alpha, Seq}$ and $N^{\beta, Seq}$ stand for the number of α -helices and β -strands present in skeleton and assignment, respectively. If there were one or more α -helices or β -strands missing in assignment, i.e. $N^{Sk} > N^{Seq}$, the combination term in equation (2) was applied to the skeleton so that all possible ways of connecting secondary structures in skeleton were fit to secondary structures in assignment. On the other hand, if there were one or more redundant α -helices or β -strands in assignment, i.e. $N^{Sk} < N^{Seq}$, the combination term in equation (2) was applied to assignment so that all possible means of connecting secondary structures in assignment were fit to secondary structures in skeleton.

Once all possible topology candidates were generated, to check whether a particular topology was reasonable, an initial screening was carried out by following one of the three criteria. First, the length of the loop vector V2 (as defined in Figure 3(a)) cannot be longer than the length of a completely extended loop (L_{loop}) (Figure 2), which was estimated using $L_{loop} = N_{Laa} \times 3.8 \text{ \AA}$. Here, N_{Laa} is the number of residues in the loop and 3.8 \AA the average $C^\alpha-C^\alpha$ distance in extended loops in proteins. Second, a maximum of 50% length variation was allowed for α -helices and 33.3% for β -strands when threading against the skeleton. The rationale for the differently allowed length variation was based on the observation that β -strands are generally more sensitive to length variation due to its more extended conformation (each residue extends a β -strand by 3.3 \AA and an α -helix by 1.5 \AA). Any topology was eliminated if it contained a predicted secondary structure that was either shorter than a half, or longer than twice the length of its matching secondary structure in skeleton. This tolerance of length variation was introduced to account for errors in the determination of the length of secondary structures, both in skeleton from experimental data and in assignment from sequences. Note, any mismatch between predicted length of secondary structure from sequence and from skeleton was equally distributed over the two ends. Third, for sheet-containing proteins, since the smallest β -sheet would have two β -strands, any topology candidate with a β -sheet of only one

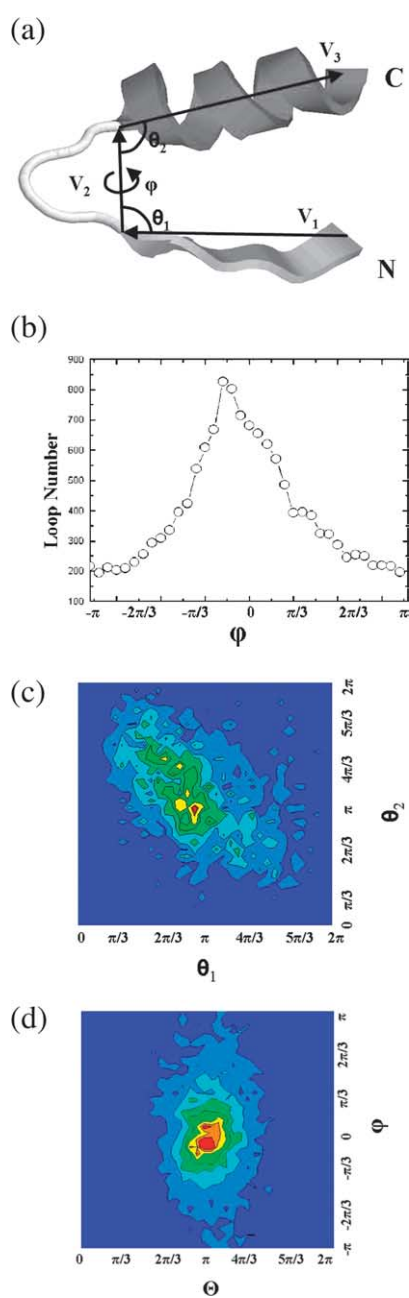


Figure 3. Geometry of two consecutive secondary structures connected by a loop (method I). (a) Three parameters, θ_1 , θ_2 and ϕ , were used to describe the relative arrangement of the two consecutive secondary structures connected by a loop. For an α -helix, it is represented by a vector of the axis of the cylinder directed from the N terminus to the C terminus. For a loop or β -strand, the vector runs from the first C^α atom to the last C atom of the loop or strand. Based on these three vectors, we defined the packing angle θ_1 between vectors V_1 and V_2 , packing angle θ_2 between vectors V_2 and V_3 , and the dihedral angle ϕ formed by the three vectors. (b) The distribution of loops as a function of the dihedral angle ϕ . The curve resembles a Gaussian distribution with a peak at near zero. (c) Two-dimensional contour representation of the distribution of angles θ_1 and θ_2 . The ridge is along the diagonal line. The loops included in this calculation are within the dihedral values between $-\pi/6$ to $\pi/6$ around the peak of the Gaussian profile shown in (b). (d) Two-dimensional contour representation of the

β -strand was eliminated. The application of these criteria removed a number of topology candidates that were inaccessible to a given sequence and skeleton. The remaining topology candidates were subject to geometry-based screening described in the following section.

Geometry-based screening

Packing geometry of two consecutive secondary structures and geometry scoring functions

In order to analyze statistically the packing geometry of secondary structures around loops, we utilized a structural database of 1084 non-homologous (less than 20% homology) proteins with resolutions better than 1.8 Å in PISCES.¹⁷ A total of 14,190 loops were extracted, together with the information about their immediately neighboring secondary structures. Three simple degrees of freedom were first defined. As shown in Figure 3(a), a loop and its neighboring secondary structures were simplistically described by three vectors; V_1 , V_2 and V_3 . For an α -helix, the vector was represented by the axis of the cylinder, and directed from N terminus to C terminus. For a loop or β -strand, the vector started from the first C^α atom and ended at the last C^α atom. Based on these three vectors, we defined the packing angle θ_1 between vectors V_1 and V_2 , packing angle θ_2 between vectors V_2 and V_3 , and the dihedral angle ϕ describing the torsional rotation among three vectors. Here, θ_1 and θ_2 were between 0 and π , while ϕ was between $-\pi$ and π . These three degrees of freedom, together with the loop end-to-end distance, i.e. the length of vector V_2 , uniquely determined the packing geometry around a particular loop.

Three methods were used to analyze the packing geometry. In method I, the values of θ_1 , θ_2 and ϕ were determined for each loop and its neighboring secondary structures without considering the types of secondary structures. They were collectively used to derive the preferred packing geometry represented in Figure 3(b)–(d).

On the basis of the distribution shown in Figure 3(d), we defined a topology scoring function to evaluate the topology candidates for a given skeleton. Considering the fact that the profile of the distribution was similar to a symmetric two-dimensional Gaussian distribution, we defined the scoring function as:

$$S(\Theta, \varphi) = A \exp \left[- \left(\frac{\Theta - a_2}{a_1} \right)^2 - \left(\frac{\varphi - a_4}{a_3} \right)^2 \right] \quad (3)$$

where A was a normalization factor. An advantage

distribution of Θ and ϕ . The dihedral angle is clearly centered at approximately zero and the sum of the two packing angles is centered at around π .

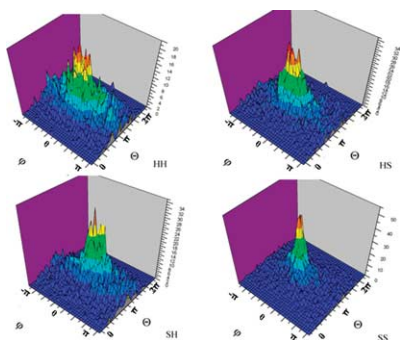


Figure 4. Geometry of two consecutive secondary structures connected by a loop (method II). The distributions of angles Θ and ϕ for loop motifs of helix-helix (HH), helix-strand (HS), strand-helix (SH) and strand-strand (SS) are shown separately.

of the analytical scoring function, $S(\Theta, \phi)$, was that it could be evaluated continuously against packing angles Θ and dihedral angle ϕ instead of the discrete data extracted from the structure database. In total, there were four parameters (a_1, a_2, a_3 and a_4), which were determined to be ($a_1 = 0.40\pi, a_2 = 1.05\pi, a_3 = 0.63\pi, a_4 = -0.02\pi$). These values of the parameters were used in the score calculation for each accessible topology candidate. For any secondary-structure packing configuration around a loop, a score was calculated to gauge the packing propensity against the known database. Then, the sum of scores for packing geometry around all loops gave the total score of that particular topology. The underlining assumption was that the native topology, selected by evolution, should have a high score in order to form a stable and compact tertiary structure.

Alternatively, in method II, we counted the packing angles and dihedral angle for four different loop motifs, helix-helix, helix-strand, strand-helix and strand-strand, separately, with both packing angle Θ and dihedral angle ϕ divided into bins of $\pi/18$ in our statistics. Packing propensities (Figure 4) were found to be similar to those using method I, although the peak for strand-strand motif was sharper and the helix-helix motif gave a wider and rougher distribution.

Instead of using continuous scoring function like equation (3), a discrete doublet probability function was used to evaluate the loop with geometrical parameter, Θ_i and ϕ_j for packing motif k . The scoring function was represented as:

$$S_k(\Theta_i, \phi_j) = N_k(\Theta_i, \phi_j) / \sum_m \sum_n N_k(\Theta_m, \phi_n) \quad (4)$$

in which $N_k(\Theta_i, \phi_j)$ was the statistical counts of all the loops belonging to motif type k from the database within the bin Θ_i and ϕ_j . The summations in the denominator were performed over all the bins accessible to Θ and ϕ .

The strand vector in methods I and II, which

was defined to connect the first and last C^α atoms of a β -strand, is expected to be very sensitive to the length of a strand and may introduce systematic errors in calculation. Thus, in method III we used a consensus vector to represent a β -strand that was defined as the weighted summation over all the vectors connecting any two C^α atoms within the strand pointing from the N terminus to the C terminus. The packing angles Θ and dihedral angle ϕ were similarly calculated as in method II for four different loop motifs, helix-helix, helix-strand, strand-helix and strand-strand, and the discrete scoring function in equation (4) was used to evaluate the packing geometry. Similar distributions as those found in method II were derived (Figure 4).

To compare the effects of these different methods on geometry screening, the statistics of packing geometry of two consecutive secondary structures using methods I and II were used in geometry-based screening of all-helical proteins (Table 1), while that of methods I, II, and III were all used in geometry-based screening of sheet-containing proteins (Table 3).

Sheet motif filter

In order to improve the efficiency of the geometry analysis in dealing with sheet-containing proteins, we adopted a probability distribution of sheet motifs with two-edge strands (open sheets) derived from a non-homologous protein database that contains 4603 sheets with different sizes (please refer to Ruczinski *et al.*¹⁸ for details). The probability distribution of a specific sheet motif provides information on how likely this sheet motif is favored in nature.

In our application, since each topology of a given sheet-containing skeleton determined a particular sheet motif, the sheet motif score of a specific topology was calculated as the occurrence of the corresponding sheet motif in the database (including the small pseudo-count) divided from the total count of β -sheets with the given strand number and helix status (see Supplementary Data for an explanation of these terms). We found that the sheet motif filter was only needed for sheet-containing proteins without mismatch between assignment and skeleton and when the total accessible topologies exceeded 1000. We used a sheet motif score of 20% of the maximal value as a cutoff, and any topology candidate with a sheet motif score higher than the cutoff was kept. For each of them, the geometry score was calculated and a composite score was derived by adding the normalized sheet motif score(s) to the normalized geometry score with a weight of 15:1 given to the geometry score for its higher reliability.

Application of geometry filter

Since the energetics calculation is relatively computationally costly, to reduce the unnecessary

Table 1. Results on 25 single-domain all-helical proteins

PDB ID	Architecture ^a	Topology ^a	Total residues	$N^{\alpha b}$	Possible topologies ^c	Accessible topologies ^d	Native rank geometry (method I) ^e	Native rank geometry (method II) ^f	Topologies used for energetics	Native rank energetics (mean) ^g	Native rank energetics (median) ^h
1erc*	Up-down bundle	Pheromone ER-1	40	3	48	30	5th	5th	26	4th	4th
1mbg	Orthogonal bundle	Arc repressor mutant	40	3	48	20	1st	1st	20	1st	1st
2ezh	Orthogonal bundle	Arc repressor mutant	65	4	384	28	5th	4th	28	2nd	2nd
1a32	Orthogonal bundle	Helix hairpins	85	4	384	2	2nd	1st	2	1st	1st
1utg	Orthogonal bundle	Uteroglobin	70	4	384	6	5th	3rd	6	1st	1st
1mho	Orthogonal bundle	Recoverin	88	4	384	64	7th	26th	22	1st	1st
1no1	Not classified	Not classified	66	4	384	22	10th	14th	22	1st	1st
1i2t	Not classified	Not classified	61	4	384	4	1st	1st	4	1st	1st
1eo0	Up-down bundle	Transcription elongation factor	76	4	384	148	18th	25th	40	1st	1st
1lpe	Up-down bundle	4-Helix bundle	144	5	3840	8	4th	4th	8	1st	1st
1vls	Up-down bundle	4-Helix bundle	146	5	3840	12	2nd	1st	12	1st	1st
1aep	Up-down bundle	4-Helix bundle	153	5	3840	15	15th	5th	15	1st	1st
1bz4	Up-down bundle	4-Helix bundle	144	5	3840	4	2nd	2nd	4	1st	1st
1nkl	Orthogonal bundle	NK-lysin	78	5	3840	49	5th	1st	26	1st	1st
3icb	Orthogonal bundle	Recoverin	75	5	3840	48	2nd	1st	24	1st	1st
2psr*	Orthogonal bundle	Recoverin	95	5	1920	960	2nd	4th	40	5th	7th
1l0i*	Orthogonal bundle	NRPSP carrier	77	5	23040	58	6th	9th	34	8th	7th
2cro	Orthogonal bundle	434 repressor	65	5	3840	1544	184th	227th	200	12th	10th
2asr	Up-down bundle	Hemerythrin (Met, subunit A)	142	5	3840	21	2nd	6th	21	1st	1st
1g7d	Not classified	Not classified	101	5	3840	15	6th	6th	15	1st	1st
1abv	Orthogonal bundle	Peroxidase	105	6	46,080	166	8th	8th	56	2nd	2nd
1a1w	Orthogonal bundle	Death domain	83	6	46,080	113	8th	9th	35	1st	3rd
1c15	Orthogonal bundle	Death domain	94	6	46,080	12	4th	4th	12	1st	1st
1ngr	Orthogonal bundle	Death domain	74	6	46,080	32	5th	5th	24	2nd	2nd
1bvc	Orthogonal bundle	Globin-like	153	8	10,321,920	14	1st	2nd	14	1st	1st

An asterisk (*) indicates the presence of mismatches between assignment and skeleton.

^a Architecture and topology are according to the classification in CATH.²²

^b N^{α} is the total number of α -helices in the crystal structures.

^c The total possible topologies were calculated according to equation (1) or (2).

^d The total accessible topologies were the number of topologies surviving through the initial screening.

^e Rank of the native topology among all accessible topologies using geometry analysis method I.

^f Rank of the native topology among all accessible topologies using geometry analysis method II.

^g Rank of the native topology among all accessible topologies by the energetics approach and ranked according to arithmetic mean.

^h Rank of the native topology among all accessible topologies by the energetics approach and ranked according to median.

cost, the geometry analysis was used to rank all the accessible topologies, so that an appropriate cutoff was applied to select a fraction of them for energetics the calculation. However, the ranking from the geometry analysis was not used in the final ranking of accessible topologies by the energetics calculation.

In selecting the appropriate cutoff, the absolute value for the geometry filter S_{cutoff} was set as:

$$S_{\text{cutoff}} = S_{\text{Max}} - |S_{\text{Max}} - S_{\text{Min}}| \times f(N_A) \quad (5)$$

in which S_{Max} and S_{Min} were the maximal and minimal values of the geometry scores for all accessible topology candidates, respectively. $f(N_A)$ was a percentage function of the total number of the accessible topology candidates N_A , expressed as:

$$f(N_A) = (1 - k) \times \exp\left(-\frac{N_A}{\alpha N_0}\right) + k \quad (6)$$

α was a parameter to control the decay of the exponential term and set as 0.5 in our study. N_0 and k were set as 50 and 0.3 for all-alpha proteins, 50 and 0.25 for alpha-beta-mixed proteins, and 100 and 0.1 for all-beta proteins, respectively.

Energetics-based screening

Construction of structure ensemble for each topology candidate

For each of the accessible topology candidates that passed the geometry filter, an ensemble of slightly perturbed structures was generated to represent the distribution of that topology in conformational space. All the protein structures generated were coarse-grained, i.e. each residue was treated as one point and represented by the position of its C^α atom. Each structure was constructed by secondary-structure placement (α -helix and β -strand) and loop construction (see Supplementary Data for details). Loop construction was achieved by allowing Monte Carlo (MC) relaxation of loops that were initially built in an extended conformation. Any perturbed structure was discarded if a potential spatial clash (less than 2 Å between any two C^α atoms) was present.

The C^α -based energy function that guided the MC relaxation in loop construction included three parts: (i) within the loop, the short-range bonded energy terms were described by the potential functions:¹⁹

$$E_{\text{SR}} = \sum_{i=2}^N E(l_i) + \sum_{i=2}^{N-1} E(\theta_i) + \sum_{i=2}^{N-1} [E(\varphi_i^-)/2 + E(\varphi_{i-1}^+)/2 + \Delta E(\varphi_i^-, \varphi_i^+)] + \sum_{i=2}^{N-1} [\Delta E(\theta_i, \varphi_i^-) + \Delta E(\theta_i, \varphi_i^+)] \quad (7)$$

where N is the number of C^α atoms in the loop. The first summation was the potential energy associated with stretching the virtual bonds between neighboring C^α atoms, which was approximated by a harmonic form, $k(l_i - l_0)^2$, in which k was the force constant of $10RT/\text{Å}^2$ and l_0 was set to 3.8Å. The second and the third summations accounted for the distortion of bond angles and bond torsions, respectively, and were expressed as knowledge-based energetic parameters. The last summation accounted for the pair-wise coupling of bond torsion and bond angle. These bonded energy terms enforced the loop to adopt a reasonable local stereochemical configuration. (ii) The long-range interactions were set to a form of hard-core potential (2.5 Å in radius) to avoid the intramolecular clashes. Given that the structures were generated very crudely at this stage, the more accurate long-range interactions were not included here and would be used in the next global optimization step. (iii) Finally, an additional energy term was introduced between the C terminus of the constructed loop and the N terminus of the following secondary structure. It was also modeled by a harmonic potential with the form of $K(r - l_0)^2$, where r was the distance between the C terminus of the loop and the N terminus of the following secondary structure, l_0 was 3.8Å, and K was the force constant with a value of $20RT/\text{Å}^2$. This additional energy term ensured the smooth loop closure during construction.

Global optimization

With all the secondary structures and loops in place, the randomly perturbed structures were globally optimized to be energetically and stereochemically more favorable. Global optimization included the genetic algorithm (GA)^{20,21} optimization for rotations of the helices, MC relaxation for loop regions and MC optimization for β -sheet regions. By operating this process iteratively, the whole structure was guided globally to a better-defined state.

In the rotational optimization of the helices, the total energy function used included the short-range terms expressed in equation (7) and the long-range terms evaluated using a residue-specific distance-dependent coarse-grained potential extracted from the structural database by Bahar & Jernigan.¹⁴ The long-range terms can be written as:

$$E_{LR} = \sum_{i < j} \bar{u}(i, j, r) \quad (8)$$

in which i and j stand for different residue type and r was the distance between i and j residues $\bar{u}(i, j, r)$, was the energy parameter derived by Bahar & Jernigan.¹⁴

All helices in a structure were allowed to rotate around their helical axes by a GA to search for lower energy configurations. The rotation around their helical axes was the most sensitive movement for α -helices. The loops were relaxed again using an MC simulation similar to that used for loop construction with the exception that the total energy function used in this step was the same as the one used in helical rotational optimization. For β -sheet regions, we used the method similar to that used in the loop relaxation with the exception that tangential elastic restraints were added between movable C^α atoms in the β -strands and the original pseudotraces (see Supplementary Data for details). This effective potential assured that each optimized structure would not deviate too much from the given skeleton. Iterative application of GA optimization of helical rotation and MC optimization of loop and sheet conformation guided the whole structure to a globally optimized state.

Energetics-based topology ranking

Within the ensemble of slightly perturbed structures for a given topology, each structure was subject to iterative global energy optimization of α -helices, β -sheets and loops. Finally, the average energy for the whole ensemble was evaluated, which was used to rank the topology candidate in the accessible topology space, the topology with the lowest energy (first) is considered as the most favorable. The basic assumption of this approach is that, although the individual structure of a non-native topology may be lower in energy than that of the native topology, due to inevitable errors in artificial construction and optimization of structures, the native topology is in general more robust

against structural variations so that the average energy of the entire ensemble of structures for the native topology is most likely to be the lowest.

Symbols used in this study

In the whole ensemble of N_T topology candidates for a given skeleton, only one is the native topology, while all others are non-native. Figure 5(a) and (b) show the sketches of the native topology and one non-native topology for the protein ligd. Here, H1, S1–S4 are the indices of the secondary structures that are numbered in the order of their spatial arrangement, and positive or negative signs refer to the direction: positive if it is the same as appeared in the crystal structure and negative if opposite. Therefore, the expression of the native topology (from N→C terminus) is S2+S1+H1+S4+S3+ while the non-native topology is S2+S1+H1–S3+S4+.

In dealing with each specific topology candidate, every secondary structure and connecting loop was converted into vector representation, so that a connecting vector chain for geometric packing was generated. As shown in Figure 5(c) and (d), the vectors representing loops are drawn as dotted lines and vectors representing helices and strands are continuous lines.

CPU requirement

The geometry method is very computationally efficient. It normally takes less than five minutes for any of the systems we tested here. For the energetics method, the required time is dependent on the size and the number of secondary structures of the protein as well as the number of accessible topologies used in the energetics calculation. For a simple three-helical protein, it takes 20 minutes to generate the structure ensemble and another 80 minutes to perform the global minimization for each accessible topology candidates. For an eight-helical protein, on the other hand, it takes 240 minutes to generate the structure ensemble and another 720 minutes for global minimization for each accessible topology. All CPU estimates are based on the SGI Origin 2000.

Results

Secondary structure prediction based on protein primary sequence

When applying secondary structure prediction algorithms to the 50 proteins in our current study, 12 showed mismatches in the number of secondary structures between skeletons and assignments. Additional consensus evaluation was therefore applied and correctly identified the assignment in three cases (3icb, 1a1w and 1d1l). The other nine proteins with mismatches are indicated with an asterisk (*) in Tables 1 and 3.

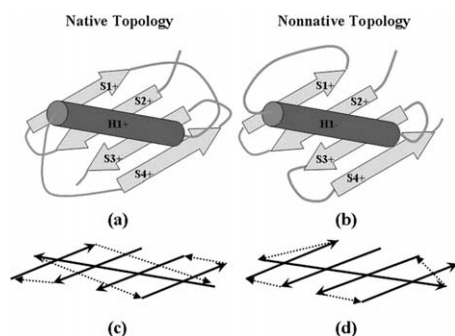


Figure 5. Vector expression of topology candidates, using ligd as an example. (a) The native topology; (b) one non-native topology; (c) the vector representation of the topology in (a); (d) the vector representation of the topology in (b).

To obtain secondary structure assignment for an eight-stranded sheet of the $\lambda 2$ protein of reovirus, we first used PSIPRED that gave a significant mismatch with the skeleton model based on an experimental 7.6 Å cryo-EM density map. Therefore, a consensus approach was utilized. Among all the methods provided by the server, DSC²² yielded an assignment that matched with the skeleton from cryo-EM data. This assignment was used to align with the skeleton.

Packing geometry of two consecutive secondary structures

In order to investigate if there was any preference of packing geometry between two consecutive secondary structures (α -helices or β -strands) connected by a loop, we started searching the structural database and analyzing the distribution of packing geometry of the secondary structures around loops. We examined 1084 non-homologous protein structures with better than 1.8 Å resolution compiled in PISCES.¹⁷ Three parameters, θ_1 , θ_2 and ϕ , were used to describe the relative arrangement of two secondary structures and their connecting loop (method I) (Figure 3(a)). Interestingly, the distribution of dihedral angle ϕ resembled a Gaussian distribution with a peak at near zero (Figure 3(b)), which suggests that the majority of two consecutive secondary structures are arranged in a plane with a *cis* configuration. Moreover, the ridge of the distribution of two packing angles, θ_1 and θ_2 , was along the diagonal line from the lower-right corner to upper-left corner, with a sum $\Theta = \theta_1 + \theta_2$ of π (Figure 3(c)). This indicates that the two secondary structures connected by a loop have a strong tendency to be antiparallel. When the distribution was plotted against Θ and ϕ (Figure 3(d)), the dihedral angle was clearly centered at approxi-

mately zero and the sum of the two packing angles was centered around π . These results suggest that the packing geometry of the consecutive secondary structures exhibits a high preference of being antiparallel and in a *cis* configuration, which seems to be necessary for constructing an overall compact protein structure. Two additional methods were also used to analyze packing geometry of these secondary structures (methods II and III), and similar statistics were obtained (Figure 4). These statistics were used as the basis of the geometry scoring function (equations (1) and (2)) in the geometry filter.

All-helical proteins

A total of 25 all-helical proteins were included in the test set (Table 1). These proteins belong to two major types of architecture of all-helical proteins with a single domain: up-down bundle and orthogonal bundle,¹⁹ and represent 14 types of topology (three proteins do not have a classified architecture and topology).

Geometry approach

In the geometry analysis using method I, three proteins have their native topologies ranked as first, and 19 other proteins have their native topologies ranked within the top ten, and only one (PDB code, 2cro) has its native topology ranked as 184th (Table 1, eighth column). Close inspection of this particular protein revealed that it has a highly globular structure and almost all the helices are of similar length. These features resulted in a quite large number of accessible topology candidates that survived the initial screening. In sharp contrast, apomyoglobin (PDB code, 1bvc) has wider variations in the length of helices and loops, which

Table 2. Results of energetics analysis for all accessible topology candidates of protein 1g7d

Index	Spatial alignment	Average long-range energy (RT)	SD of long-range energy (RT)	Average total energy (RT)	SD of total energy (RT)	No. randomized structures in the ensemble
1	H1 + H2 + H3 + H4 + H5 +	-107.27	31.1	-433.04	39.13	1067
2	H1 + H2 + H3 + H4 + H5-	-98.39	30.4	-423.09	41.37	1027
3	H2 - H1 - H3 + H4 + H5 +	-74.54	31.76	-413.26	42.88	240
4	H2 - H1 - H3 + H4 + H5 -	25.023	40.54	-315.99	46.36	164
5	H2 - H1 - H4 - H3 - H5 +	268.73	96.3	-29.28	118.45	45
6	H2 - H1 - H4 - H3 - H5 -	632.65	81.32	336.12	96.85	16
7	H3 + H4 + H1 + H2 + H5 -	-82.08	27.79	-427.24	116.17	2401
8	H4 - H3 - H2 - H1 - H5 +	235.34	53.98	-115.27	59.98	107
9	H4 - H3 - H2 - H1 - H5 -	442.12	63.35	85.37	67.39	77
10	H4 - H3 - H1 + H2 + H5 +	48.52	38.06	-309.14	49.27	301
11	H2 - H1 - H4 - H5 - H3 +	274.64	79.41	-49.43	82.53	50
12	H5 + H4 + H1 + H2 + H3 +	-83.067	26.23	-422.36	45.38	2406
13	H4 - H5 - H2 - H1 - H3 +	259.58	68.31	-79.94	78.69	138
14	H4 - H5 - H2 - H1 - H3 -	450.7	70.99	80.49	68.96	62
15	H4 - H5 - H1 + H2 + H3 +	-52.17	34.88	-407.34	51.34	357

The native topology is highlighted in bold. Energy is in the unit of RT. Average long-range energy (third column) and average total energy (fifth column) are listed together with their standard deviation (SD) (fourth and sixth columns, respectively). The total number of perturbed structures in the ensemble for each topology is listed in the seventh column.

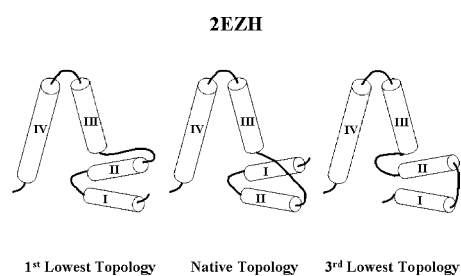


Figure 6. Comparison of the three lowest-energy topology candidates for an all-helical protein 2ezh whose native topology was ranked the second-lowest.

drastically narrowed the accessible topology to 14 in the initial screening out of the total of 10^7 possible topologies. The use of method II in the geometry analysis yielded similar results (Table 1, ninth column) to method I.

Energetics approach

The results of the energetics analysis on these 25 proteins are listed in Table 1 (11th to 12th columns). The energetics-based evaluation was performed after geometry filter. An appropriate cutoff was applied so that all topology candidates above the cutoff were input for energetics analysis. The native topologies of 18 proteins were successfully recognized as of the lowest average energy (ranked as first). Such a successful rate is reasonably high. As an example, Table 2 lists all accessible topology candidates of protein 1g7d together with their energy values. Note that the number of randomly perturbed structures in the ensemble of each accessible topology is quite different. This is because, for some of the energetically unfavorable topology candidates, it was much harder to generate perturbed structures around the given skeleton that satisfied all the criteria. In Table 2, there is a good correlation between the number of perturbed structures and the average energy, especially the average long-range non-bonded energy. The ones with a larger number of randomized structures tend to have a lower average energy and a smaller standard deviation of average energy.

Among the other seven proteins, three (2ezh, 1abv and 1ngr) have their native topology candidates recognized as the second lowest in average energy (ranked as second in the 11th column of Table 1). The topology of the lowest energy (first) is very similar to the native topology (second) in all cases. As an example, the schematic sketches of the three lowest energy topology candidates for protein 2ezh are illustrated in Figure 6, the difference between the first and the native topology (second) was a swap of two helices that are very similar in length and nearby in space.

The native topologies of two other proteins, 2psr and 110i, were ranked fifth and eighth, respectively.

As shown in Table 1, mismatches were found between the assignment and skeleton for both cases: two helical regions were merged as one in the assignment of 2psr, and one extra helical region was predicted for 110i. These original errors obviously affect the ranks of the native topologies of these two proteins.

One special case worth mentioning is a three-helical up-down bundle protein, 1erc. In the prediction, one α -helical region and one β -strand region were predicted (instead of three helices). Thus, the mismatch of the prediction results was out of tolerance, and the equation for generating all possible topology (equation (2)) was no longer applicable. In such a case, we did our calculation based on the skeleton information only. Since the three helices in 1erc are highly symmetric, obeying a pseudo-3-fold symmetry, in order to preclude any bias from the skeleton, we evenly divided the amino acid sequence into three helical regions of ten residues each, intervened by two loops of five residues. In other words, helices 1, 2 and 3 are composed of residues 1–10, 15–25, and 30–40, respectively. The analysis identified the native topology as fourth-lowest in energy (Table 1).

In the case of 2cro, the native topology was ranked as the 12th-lowest in average energy. This is presumably because of the length similarity of all five helices of this protein that gave rise to the largest number of accessible topology candidates in the initial screening. However, despite that, our methods were very effective in narrowing the searching space of possible topologies (the native topology was ranked as the 12th among all 1544 accessible candidates).

Finally, to cross-check the potential errors in our ranking procedure resulting from the non-Boltzmann random sampling in generating the perturbed structure ensemble, we also used the median energy value of the ensemble, instead of the arithmetic mean, to rank the topology candidates. The median could give a faithful indication of true average in the absence of *a priori* knowledge of data distribution. The results, shown in the 12th column of Table 1, are very consistent with those ranked according to the arithmetic mean (11th column), indicating the fidelity of our ranking procedure.

Sheet-containing proteins

Geometry approach

There were 25 sheet-containing proteins in the test set, 19 alpha-beta-mixed proteins and six all-beta proteins. The 19 alpha-beta-mixed proteins belong to three different types of architecture and seven types of topology (three proteins do not have classified architecture or topology in CATH¹⁹). The six all-beta proteins, on the other hand, belong to three types of architecture and three types of topology (one protein does not have classified architecture or topology). The results of a geometry analysis by method I are listed in the ninth column

Table 3. Results on 25 sheet-containing proteins

PDB ID	Architecture ^a	Topology ^a	Total residues	$N^{\alpha b}$	$N^{\beta c}$	Possible topologies ^d	Accessible topologies ^e	Native rank geometry method I ^f	Native rank geometry method II ^g	Native rank geometry method III ^h	Topologies used for energetics	Native rank energetics (mean) ⁱ	Native rank energetics (median) ^j
<i>A. Alpha-beta-mixed proteins</i>													
1igd	Roll	Ubiquitin-like	61	1	4	768	22	2nd	3rd	3rd	22	1st	1st
1em7	Roll	Ubiquitin-like	56	1	4	768	40	1st	1st	2nd	13	1st	1st
1h0y	Not classified	Not classified	89	2	4	3072	36	2nd	2nd	2nd	17	1st	1st
1ctf*	2-Layer sandwich	Alpha-beta plaits	68	3	3	4608	23	4th	2nd	2nd	23	15th	16th
1d1l	2-Layer sandwich	CRO repressor	61	3	3	2304	6	3rd	1st	2nd	6	1st	2nd
1p1l	Not classified	Not classified	102	3	4	18,432	7	1st	1st	1st	7	1st	2nd
1cm ²	2-Layer sandwich	Alpha-beta plaits	85	3	4	18,432	18	1st	2nd	1st	5	1st	1st
1h75	3-Layer (aba) sandwich	Glutaredoxin	76	3	4	18,432	109	9th	13th	2nd	21	1st	1st
1lba	3-Layer (aba) sandwich	Lysozyme	145	3	5	184,320	640	27th	23rd	23rd	31	1st	1st
3fx2*	3-Layer (aba) sandwich	Rosmann fold	147	4	5	737,280	3476	54th	300th	220th	44	1st	1st
1rlk	Not classified	Not classified	116	4	5	1,474,560	22	1st	3rd	3rd	12	1st	1st
1thx	3-Layer (aba) sandwich	Glutaredoxin	108	4	5	1,474,560	97	7th	3rd	3rd	12	1st	1st
1rb*	Roll	Ubiquitin-like	76	2	5	7680	1712	4th	1st	1st	14	1st	1st
1orc*	2-Layer sandwich	CRO repressor	59	3	3	1152	8	1st	1st	1st	8	1st	1st
2hpr	2-Layer sandwich	Alpha-beta plaits	87	3	4	18,432	76	1st	1st	1st	5	2nd	2nd
1eof	2-Layer sandwich	Potassium channel 1.1	100	5	4	1,474,560	28	6th	6th	16th	20	2nd	2nd
1ubq	Roll	Ubiquitin-like	76	2	5	30,720	56	1st	1st	1st	7	3rd	2nd
1ck2	2-Layer sandwich	Alpha-beta plaits	104	5	4	1,474,560	276	42nd	42nd	2nd	44	3rd	3rd
1aba	3-Layer (aba) sandwich	Glutaredoxin	87	3	3	2304	640	87th	87th	53rd	88	6th	3rd
<i>B. All-beta proteins</i>													
1e0n	Not classified	Not classified	27	0	3	48	16	1st	1st	1st	8	1st	1st
1mjc	Barell	OB fold	69	0	5	3840	1872	9th	7th	7th	13	1st	1st
1fna	Sandwich	Imunoglobulin-like	91	0	7	645,120	74	7th	11th	12th	7	1st	1st
1tpm*	Ribbon	Complement module	50	0	5	3840	192	18th	14th	9th	20	8th	9th
1ten	Sandwich	Imunoglobulin-like	89	0	7	645,120	1728	10th	10th	10th	19	1st	1st
3ait*	Sandwich	Imunoglobulin-like	74	0	6	23,040	2352	3rd	11th	23rd	40	1st	1st

There are a total of 25 sheet-containing proteins tested, 19 of which are alpha-beta-mixed proteins, and six are all-beta proteins. An asterisk (*) indicates the presence of mismatch between assignment and skeleton.

^a The classification of architecture and topology are according to CATH.²²

^b N^{α} is the total number of α -helices in the crystal structures.

^c N^{β} is the total number of β -strands in the crystal structures.

^d Possible topologies were calculated according to equation (1) or (2).

^e Accessible topologies were the number of topologies surviving through the initial screening allowing a maximal 50% variation for α -helices and 33.3% for β -strands.

^f Rank of the native topology among all accessible topologies using geometry analysis method I.

^g Rank of the native topology among all accessible topologies using geometry analysis method II.

^h Rank of the native topology among all accessible topologies using geometry analysis method III.

ⁱ Rank of the native topology among all input accessible topologies by the energetics approach and ranked according to arithmetic mean.

^j Rank of the native topology among all input accessible topologies by the energetics approach and ranked according to median.

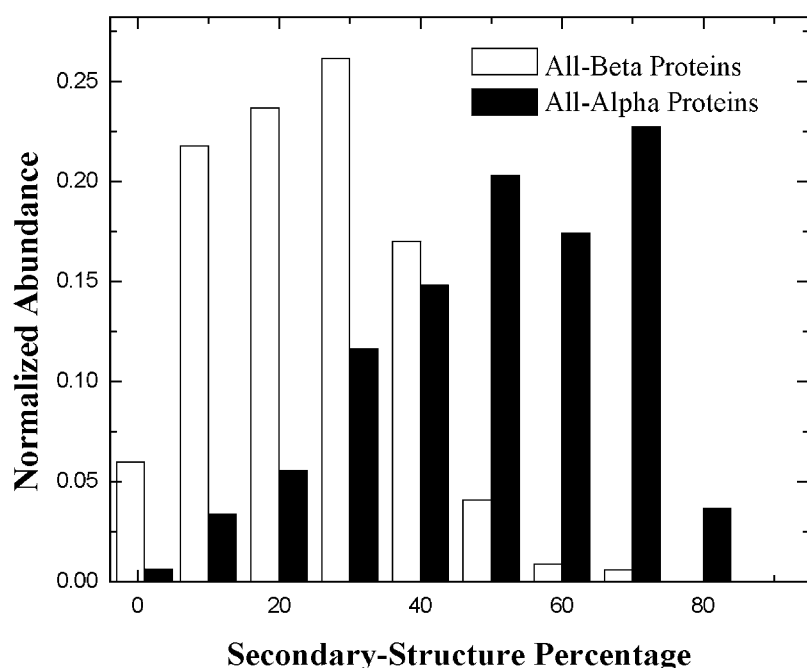


Figure 7. Comparison of secondary-structural content in all-helical proteins *versus* all-beta proteins. It is clear that all-beta proteins have a much lower secondary structural content than all-helical proteins.

of Table 3, in which the native topology was ranked as the highest scores (first) in seven cases and within the top ten in eight other cases for alpha-beta-mixed proteins. For all-beta proteins, the native topology of one protein was ranked as the first and four others within the top ten. The geometry analysis using methods II (Table 3, tenth column) and III (Table 3, 11th column) yielded qualitatively similar results.

There are several reasons that make the search for sheet-containing topology more difficult. Comparing with all-helical proteins, all-beta proteins have an overall lower percentage of secondary structures and a higher percentage of loop regions (Figure 7). This increases the complexity of the topology space, and fewer topology candidates can be filtered out in the initial screening process. Moreover, α -helices have more rigid structures with strong local interactions while β -strands have the intrinsic bending and twisting that involve long-range stabilizing interactions.

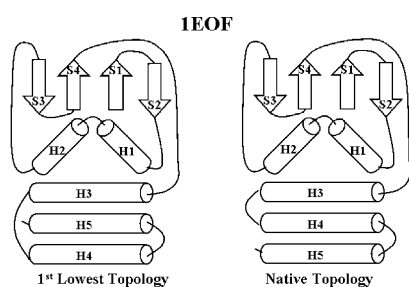


Figure 8. Comparison of the lowest-energy topology with the native topology for sheet-containing protein 1e0f whose native topology was ranked the second-lowest.

Energetics approach

The results of applying the energetics approach to the 25 sheet-containing proteins are shown in Table 3. All the accessible topologies were first subject to geometry analysis. An appropriate cutoff was applied so that all topology candidates above the cutoff were input for the energetics analysis. In order to get rid of sampling bias, we also used both the arithmetic mean and the median of the energy as ranking criteria. The final ranks of the 25 proteins by these ranking methods are listed in the 13th and 14th columns of Table 3, respectively. Again, the two ranking methods yielded quite similar results. The native topologies of 18 out of 25 proteins have been successfully recognized as of the lowest energy among all candidates by arithmetic mean. Among the other seven proteins, two have their native topologies ranked as the second-lowest average energy and two others as the third. In all these four cases, the difference between the lowest-energy topology (first) and the native topology (second) was the switch of two secondary structures with similar length and symmetric spatial location, or the shift of the direction of certain strands in the skeleton. As an example, the lowest-energy topology (first) and the native topology (second) of protein 1e0f are compared in Figure 8.

Application to incomplete skeletons

In all previous test cases, we assumed that the secondary structures in skeleton are correct and used them to judge the correctness of the predicted assignment on sequence. In reality, however, it is very likely that one or more secondary structures, especially short ones, fail to be identified in

Table 4. Results on incomplete secondary-structural skeletons of 1BVC

Missing helix index	N^a	Possible topologies	Accessible topologies	Native rank geometry (method I)	Topologies used for energetics	Native rank energetics (mean)
H3	7	5,160,960	14	1st	6	2nd
H4	7	5,160,960	33	1st	8	3rd
H3, H4	6	1,290,240	42	1st	8	1st

For an explanation, please refer to the footnotes to [Table 1](#).

skeletons based on experimental maps. We probed the efficiency of our new procedure in dealing with incomplete skeletons on protein 1bvc that has eight α -helices. In the test, we purposely used the skeletons of 1bvc that have one of the helices H3 or H4 or both missing ([Table 4](#)). H3 and H4 were chosen because they are relatively short and are more likely to be missed in structure determination and density interpretation.^{11–13} For the complete skeleton of 1bvc, our procedure successfully identified the native topology as the most preferable (first).

With H4 or both H3 and H4 missing, more accessible topologies were allowed for the skeleton ([Table 4](#)). The geometry approach, however, consistently identified the native topology as the most favorable topology (first) in all three cases. The use of the energetics approach identified the native topology as second, third and first when the missing component(s) was H3, H4, and H3 and H4, respectively. This simple test indicated that our current methods also have tolerance for small errors in skeleton.

We wish to reiterate that, in general, the success of our method does depend on the accuracy of secondary structures both in skeleton and in assignment. Usually, when those on skeleton are correct (our usual assumption), the predicted assignment is judged according to that; when the skeleton has some small ones missing, the dependence on the accuracy of the predicted assignment in sequence becomes stronger. In cases that both are drastically wrong, the likelihood for the method to fail will inevitably increase.

Application to real experimental data

Up to now, our new procedure has been extensively shown to be highly effective in identifying the native topology based on skeletons derived from heavily smoothed crystal structures (see Methods) and assignments from primary sequences. We now examine the effectiveness of our procedure in utilizing the skeletons that are derived from experimental low to intermediate-resolution structural data, which presumably contain larger errors.

The reovirion structure was solved to 7.6 Å by cryo-EM.²³ The $\lambda 2$ protein of reovirus is composed of 16 β -sheets, with one eight-stranded sheet located

at the tip of the structure. This region was selected to test our new procedure because of its continuity in sequence and comparable size to all other test cases we used. We first utilized Sheetminer¹² and Sheettracer¹³ to obtain the skeleton of this β -sheet. All eight strands were traced successfully, which are shown in [Figure 9](#) together with the corresponding crystal structure that was solved independently.²⁴ The corresponding secondary structure assignment has been obtained using the algorithm DSC.²²

In the initial screening, a large number of accessible topologies were identified. The geometry analysis identified the native topology as 116th. By applying the sheet motif filter together with the geometry filter, the native topology was ranked as fourth. When eight topologies were subject to the energetics calculation, the native topology was ranked as first. It is worth emphasizing that, despite the large deviations of main chains of the traced β -strands from the crystal structure, our analysis was still able to pinpoint correctly the native topology.

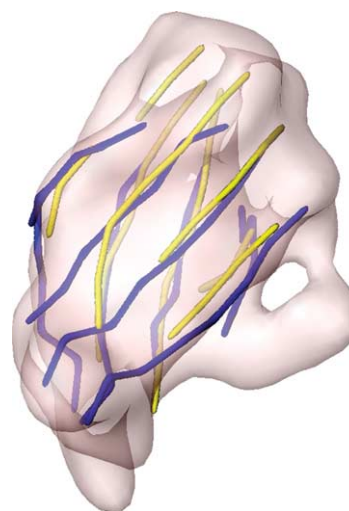


Figure 9. Superposition of the secondary-structural skeleton modeled by Sheetminer and Sheettracer (yellow) based on experimental 7.6 Å cryo-EM electron density maps (the transparent envelope) with that from the crystal structure (blue; PDB code 1ej6) of the $\lambda 2$ protein of reovirus.

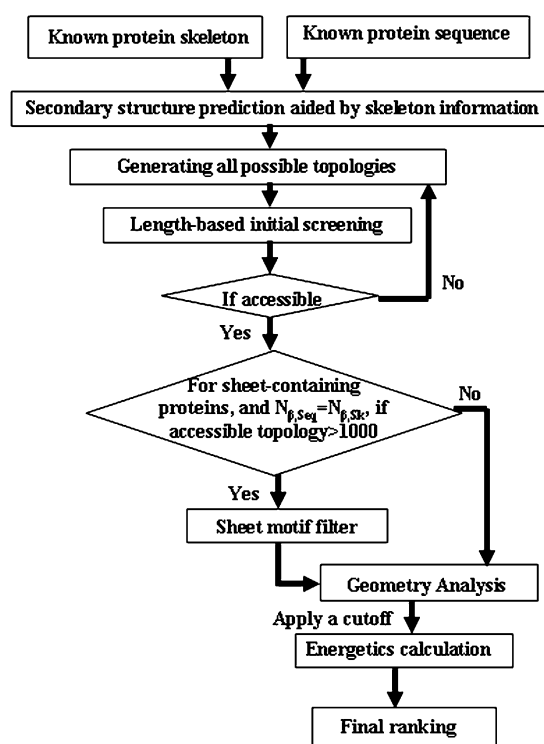


Figure 10. Flow-chart for the new computational procedure in identifying the native topology from protein secondary-structural skeleton in space and assignment in sequence.

Concluding Discussions

We have reported a new energetics-based computational approach for deriving protein topology, or fold, from secondary-structural skeleton in space and assignment in sequence (the software is available upon request). To improve computational efficiency, a complementary knowledge-based geometry filter was also developed. The new procedure was first tested on a set of 50 structurally unrelated single-domain protein structures obtained from the PDB that consist of 25 all-helical proteins and 25 sheet-containing proteins. The procedure correctly identified the native topology of 36 proteins out of the 50 soluble proteins as the first energetically favored topologies, and 12 other proteins within the top ten. Furthermore, the same procedure was applied to two more stringent test cases. One such test was on a skeleton modeled in an experimental 7.6 Å cryo-EM density map that carries significant deviations from the crystal structure. The native topology was identified successfully from the modeled skeleton (for a detailed flow-chart of the procedure, see Figure 10).

In the energetics approach, for all the structures in the ensemble generated for each topology candidate, although extensive optimization was employed to refine the structures, certain degrees of freedom were not optimized based on Boltzmann

sampling due to the limitation of computational resources. Examples include the randomization of rotational angles of helices around their centers of mass and translational shifts along their axes that were restricted to a narrow range. Therefore, the final ranking of energetics of topology candidates was compared for both arithmetic mean and median. The consistent results between the two ranking methods indicate that the ranking procedure using the arithmetic mean is reliable.

The empirical potential functions we used are very approximate. Also, the structures generated around the native skeleton evidently carry large errors regardless of the extensive optimization, which is particularly true for the loops that were essentially placed quite arbitrarily (although it was found that inclusion of loops was critical to cover a substantial portion of hydrophobic surfaces). As a consequence, the energy of an individual structure of a non-native topology can very often be lower than that of an individual structure of the native topology, which renders it impossible to distinguish the native topology by the energy value of a single constructed structure. To overcome this burden, we adopted a working hypothesis, which states, among all the possible topology candidates for a given skeleton, evolution has selected the native topology as the one that can accommodate the largest structural variation, not the one trapped in a deep, but narrow, energy well. According to this hypothesis, the average energy of an ensemble of structures varying in the vicinity of the native skeleton would most likely be the lowest, and the standard deviation of the average energy would be the smallest. Our results seem to support the hypothesis very well. Such an observation is also reminiscent of an earlier molecular dynamics study, which indicated that the overall averaged structure of multiple trajectories is significantly closer to the X-ray structure than any of the individual trajectory average structure.²⁵

The computational procedure reported here is fully applicable for determining the topology for skeletons of unknown structures. The protocol includes first using the initial screening to remove any inaccessible topologies, then using the geometry-based filter to obtain a ranking of all the accessible topology candidates, and finally using an appropriate cutoff to select a fraction of accessible topologies for the energetics analysis. The analysis will in many cases pinpoint the native topology as the most favorable on the final list. Furthermore, any additional knowledge about the structure under study can be used, together with the final ranking of all topologies, to help identify the native topology. For example, if one happens to know the identity of one or a few secondary structures in the density maps, such knowledge can enormously help filter out the non-native topology candidates if any of them happens to be of a better rank than the native topology or help confirm the native topology.

Obviously, the method is not perfect and naturally suffers from the errors in both structural

measurement and secondary structural prediction. Despite a high success rate, there will be cases that it fails to narrow down the native topology as top choices, particularly in cases that severe mismatch of secondary structures occurs between the skeleton modeled from density maps and the assignment predicted from sequence. However, it is important to realize that our new procedure demonstrated that it is possible to define native protein topology based on fairly limited structural data without solving the structure to much higher resolution. More importantly, the approach described here may be valuable in protein folding and structure prediction by allowing effective discrimination of non-native topology (fold) candidates from the native topology in the vast topology space.

Finally, we wish to point out that the successful use of the average energy of an ensemble of randomly perturbed structures as a robust parameter for comparing the relative stability of different topology candidates may have an important implication for threading studies.^{26–32} Since it is not uncommon that proteins belonging to the same topology (fold) differ significantly in length and exact positioning of secondary structures, one may obtain more accurate energy values if structural variations around the given template are taken into account.

Acknowledgements

The authors gratefully acknowledge the support from the National Institutes of Health (R01-GM067801). M.C. and M.L. are supported partially by a predoctoral fellowship from the W. M. Keck Foundation of the Gulf Coast Consortia through the Keck Center for Computational and Structural Biology. J.M. is a recipient of the Award for Distinguished Young Scholars Abroad from the National Natural Science Foundation of China. The authors also thank an anonymous referee for his or her careful and critical review that improved the paper substantially.

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at [doi:10.1016/j.jmb.2005.04.064](https://doi.org/10.1016/j.jmb.2005.04.064)

References

1. Bottcher, B., Wynne, S. A. & Crowther, R. A. (1997). Determination of the fold of the core protein of hepatitis B virus by electron cryomicroscopy. *Nature*, **386**, 88–91.
2. Conway, J. F., Cheng, N., Zlotnick, A., Wingfield, P. T., Stahl, S. J. & Steven, A. C. (1997). Visualization of a 4-helix bundle in the hepatitis B virus capsid by cryo-electron microscopy. *Nature*, **386**, 91–94.
3. Mancini, E. J., Clarke, M., Gowen, B. E., Rutten, T. & Fuler, S. D. (2000). Cryo-electron microscopy reveals the functional organization of an enveloped virus. Semliki Forest virus. *Mol. Cell*, **5**, 255–266.
4. Zhou, Z. H., Dougherty, M., Jakana, J., He, J., Rixon, F. J. & Chiu, W. (2000). Seeing the herpesvirus capsid at 8.5 Å. *Science*, **288**, 877–880.
5. Zhou, Z. H., Liao, W., Cheng, R. H., Lawson, J. E., Mcarthy, D. B., Reed, L. J. & Stoops, J. K. (2001). Direct evidence for the size and conformational variability of the pyruvate dehydrogenase complex revealed by three-dimensional electron microscopy The “breathing” core and its functional relationship to protein dynamics. *J. Biol. Chem.* **276**, 21704–21713.
6. Zhou, Z. H., Baker, M. L., Jiang, W., Dougherty, M., Jakana, J., Dong, G. *et al.* (2001). Electron cryomicroscopy and bioinformatics suggest protein fold models for rice dwarf virus. *Nature Struct. Biol.* **8**, 868–873.
7. Kuhn, R. J., Zhang, W., Rosmann, M. G., Pletnev, S. V., Corver, J., Lenches, E. *et al.* (2002). Structure of dengue virus: implications for flavivirus organization, maturation, and fusion. *Cell*, **108**, 717–725.
8. Li, H., DeRosier, D., Nicholson, W., Nogales, E. & Downing, K. (2002). Microtubule structure at 8 Å resolution. *Structure (Camb)*, **10**, 1317–1328.
9. Zhang, X., Shaw, A., Bates, P. A., Newman, R. H., Gowen, B., Orlova, E. *et al.* (2000). Structure of the AAA ATPase p97. *Mol. Cell*, **6**, 1473–1484.
10. Ludtke, S. J., Chen, D. H., Song, J. L., Chuang, D. T. & Chiu, W. (2004). Seeing GroEL at 6 Å resolution by single particle electron cryomicroscopy. *Structure (Camb)*, **12**, 1129–1136.
11. Jiang, W., Baker, M. L., Ludtke, S. J. & Chiu, W. (2001). Bridging the information gap: computational tools for intermediate resolution structure interpretation. *J. Mol. Biol.* **308**, 1033–1044.
12. Kong, Y. & Ma, J. (2003). A structural-informatics approach for mining β-sheets: locating sheets in intermediate-resolution density maps. *J. Mol. Biol.* **332**, 399–413.
13. Kong, Y., Zhang, X., Baker, T. S. & Ma, J. (2004). A structural-informatics approach for tracing β-sheets: building pseudo-C^α traces for β-strands in intermediate-resolution density maps. *J. Mol. Biol.* **339**, 117–130.
14. Bahar, I. & Jernigan, R. L. (1997). Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation. *J. Mol. Biol.* **266**, 195–214.
15. Jones, C. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292**, 195–202.
16. McGufin, L. J., Bryson, K. & Jones, D. T. (2000). The PSIPRED protein structure prediction server. *Bioinformatics*, **16**, 404–405.
17. Wang, G. & Dunbrack, R. L., Jr (2003). PISCES: a protein sequence culling server. *Bioinformatics*, **19**, 1589–1591.
18. Ruczinski, I., Kooperberg, C., Bonneau, R. & Baker, D. (2002). Distributions of beta sheets in proteins with application to structure prediction. *Proteins: Struct. Funct. Genet.* **48**, 85–97.
19. Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T.,

- Swindels, M. B. & Thornton, J. M. (1997). CATH—a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1109.
20. Davis, L. (1991). *Handbook of Genetic Algorithms*, Van Nostrand Reinhold, New York.
21. Goldberg, D. E. (1989). *Genetic Algorithms in Search Optimization and Machine Learning*, Addison-Wesley, Boston.
22. King, R. D. & Sternberg, M. J. E. (1996). Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. *Protein Sci.* **5**, 2298–2310.
23. Zhang, X., Walker, S. B., Chipman, P. R., Nibert, M. L. & Baker, T. S. (2003). Reovirus polymerase lambda 3 localized by cryo-electron microscopy of virions at a resolution of 7.6 Å. *Nature Struct. Biol.* **10**, 1011–1018.
24. Reinisch, K. M., Nibert, M. L. & Harison, S. C. (2000). Structure of the reovirus core at 3.6 Å resolution. *Nature*, **404**, 960–967.
25. Caves, L. S., Evanseck, J. D. & Karplus, M. (1998). Locally accessible conformations of proteins: multiple molecular dynamics simulations of crambin. *Protein Sci.* **7**, 649–666.
26. Miler, R. T., Jones, D. T. & Thornton, J. M. (1996). Protein fold recognition by sequence threading: tools and assessment techniques. *FASEB J.* **10**, 171–178.
27. Skolnick, J., Kolinski, A., Kihara, D., Betancourt, M., Rotkiewicz, P. & Boniecki, M. (2001). *Ab initio* protein structure prediction via a combination of threading, lattice folding, clustering, and structure refinement. *Proteins: Struct. Funct. Genet.* **5**, 149–156.
28. Elofson, A., Fischer, D., Rice, D. W., Le Grand, S. M. & Eisenberg, D. (1996). A study of combined structure/sequence profiles. *Fold. Des.* **1**, 451–461.
29. Kihara, D., Lu, H., Kolinski, A. & Skolnick, J. (2001). TOUCHSTONE: an *ab initio* protein structure prediction method that uses threading-based tertiary restraints. *Proc. Natl Acad. Sci. USA*, **98**, 10125–10130.
30. Lu, L., Lu, H. & Skolnick, J. (2002). MULTIPROSPECTOR: an algorithm for the prediction of protein-protein interactions by multimeric threading. *Proteins: Struct. Funct. Genet.* **49**, 350–364.
31. Jones, D. T., Miler, R. T. & Thornton, J. M. (1995). Successful protein fold recognition by optimal sequence threading validated by rigorous blind testing. *Proteins: Struct. Funct. Genet.* **23**, 387–397.
32. Jones, D. T. & Thornton, J. M. (1996). Potential energy functions for threading. *Curr. Opin. Struct. Biol.* **6**, 210–216.

Edited by J. Thornton

(Received 24 November 2004; received in revised form 24 April 2005; accepted 27 April 2005)