*Proc. Natl. Acad. Sci. USA*
Vol. 95, pp. 1020–1025, February 1998
Biophysics

# Nativelike topology assembly of small proteins using predicted restraints in Monte Carlo folding simulations

ANGEL R. ORTIZ*, ANDRZEJ KOLINSKI*†, AND JEFFREY SKOLNICK*‡

*Department of Molecular Biology, The Scripps Research Institute, La Jolla, CA 92037; and †Department of Chemistry, University of Warsaw, Pasteura-1, 02–093, Warsaw, Poland

*Communicated by H. A. Scheraga, Cornell University, Ithaca, NY, November 26, 1997 (received for review October 24, 1997)*

**ABSTRACT** **By incorporating predicted secondary and tertiary restraints derived from multiple sequence alignments into *ab initio* folding simulations, it has been possible to assemble native-like tertiary structures for a test set of 19 nonhomologous proteins ranging from 29 to 100 residues in length and representing all secondary structural classes. Secondary structural restraints are provided by the PHD secondary structure prediction algorithm that incorporates multiple sequence information. Multiple sequence alignments also provide predicted tertiary restraints via a two-step process: First, *seed* side chain contacts are selected from a correlated mutation analysis, and then an inverse folding algorithm expands these seed contacts. The predicted secondary and tertiary restraints are incorporated into a lattice-based, reduced protein model for structure assembly and refinement. The resulting native-like topologies exhibit a coordinate root—mean–square deviation from native for the whole chain between 3.1 and 6.7 Å, with values ranging from 2.6 to 4.1 Å over ≈80% of the structure. Overall, this study suggests that the use of restraints derived from multiple sequence alignments combined with a fold assembly algorithm is a promising approach to the prediction of the global topology of small proteins.**

The relationship of a protein sequence to its native structure is commonly referred to as the protein structure prediction problem. This is a subset of the protein folding problem that is also concerned with the mechanism of native structure assembly (1). It is widely accepted that proteins obey the "thermodynamic hypothesis," which asserts that the native conformation corresponds to a global free energy minimum (2). However, because of the complexity of the interactions and the structure and high dimensionality of the energy landscape, the task of finding this free energy minimum by theoretical methods is extremely difficult (3).

One way to partially surmount the multiple minimum problem is to create a funnel in the potential energy landscape by using restraint information in the conformational search. Such restraints might include known or predicted secondary structure and/or tertiary contacts. In the past, this approach has been explored by a number of authors (4–8). For example, by using a genetic algorithm to explore conformational space in which a known secondary structure is assumed, Dandekar and Argos (4) have described encouraging results for some simple helical and $\beta$ proteins. There also have been studies that incorporate known secondary structure and a limited number of known tertiary restraints (5–7). For example, Mumenthaler and Braun (8) developed a self-correcting distance geometry method that incorporates known secondary structure and uses tertiary restraints predicted from multiple sequence alignments. In 6 of 8 helical proteins, this approach successfully identified the native topology. In this spirit, Aszodi and coworkers (6) have developed an approach based on distance geometry. A set of experimentally derived tertiary distance restraints is supplemented by predicted interresidue distances obtained from conserved hydrophobic amino acid patterns extracted from multiple sequence alignments. To assemble structures whose backbone coordinate root—mean–square deviation (cRMSD) is below 5 Å on average, this approach requires more than N/4 tertiary restraints, where N is the number of protein residues. More recently, Skolnick and coworkers have reported encouraging results when loosely defined exact secondary structure and N/4 exact tertiary restraints are used (7) in their MONSSTER (MOdeling of New Structures from Secondary and TErtiary Restraints) algorithm and have tested the method on a variety of different topologies. In what follows, we extend the MONSSTER algorithm to explore whether use of predicted secondary structure and tertiary restraints is adequate to assemble tertiary structure from sequence information alone.

## METHODS

Fig. 1 presents a schematic overview of the entire approach. Our tertiary structure prediction scheme is divided logically into two parts: derivation of restraints from multiple sequence alignment information and global structure assembly/refinement using the MONSSTER protein structure assembly algorithm (7, 9) as modified to reflect the expected accuracy and precision of the predicted tertiary restraints. In what follows, we describe each aspect of the protocol.

**Secondary Structure Prediction.** Multiple sequence alignments as obtained from the HSSP database (10) are inputted into the PHD secondary structure prediction algorithm (11, 12). Elements predicted as U-turns (regions in the structure where the chain reverses its global direction) by our *LINKER* algorithm (13) override the secondary structure predictions of PHD. Thus, residues are assigned to one of five secondary structural states: strand, helix, U-turn, extended state/loop, and nonpredicted. The set of predicted secondary elements (helix or strand) between U-turns comprises the putative core region of the molecule.

**Prediction of Tertiary Contacts.** When any two side chain heavy atoms in a pair of residues lie at a distance ≤5 Å, these two residues are said to be in contact. Different authors have suggested that multiple sequence information can be used to predict such contacts in native protein structures on the basis of residue conservation (6, 8) or covariation (14–16). Here, we focus on residue covariation and slightly adapt the approach of Goebel *et al.* (14) by restricting the sequence covariation calculation to those residues predicted to be in the putative protein core. In these regions, the assumption of spatial closeness might be better. In practice, a cutoff of 0.5 for the correlation coefficient for pairwise mutations is used for contact prediction. We term those contacts extracted from the correlated mutation analysis as

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.



Abbreviation: cRMSD, coordinate root—mean–square deviation.

‡To whom reprint requests should be addressed. e-mail: skolnick@scripps.edu.

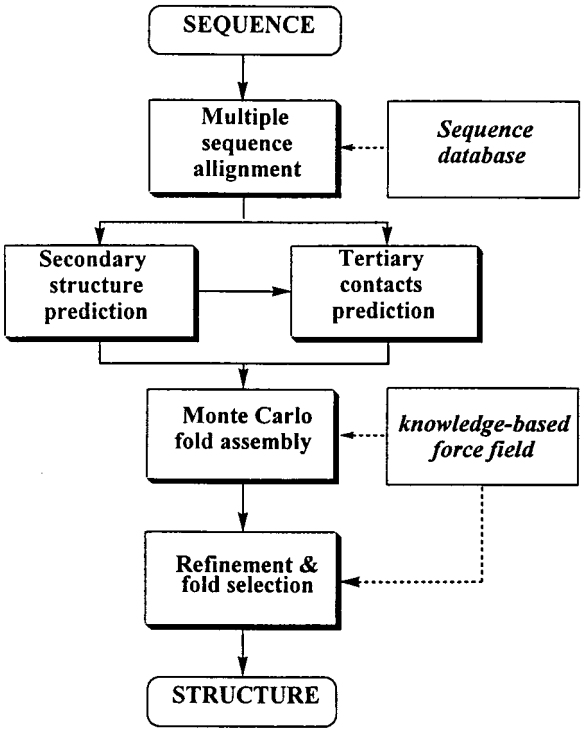


FIG. 1. Schematic overview of the tertiary structure prediction procedure.

"seeds." Unfortunately, there are too few to assemble a protein from the unfolded state by using MONSSTER.

To enrich the number of predicted contacts, a combined structural fragment search and inverse folding procedure (9, 17) is used. First, a structural database is searched to identify all pairs of secondary structural elements compatible with the predicted seed contact (within ±1 residue). Next, the top 10 scoring fragment pairs based on their secondary structural propensities and burial energy are examined, and the mutual pairwise cRMSDs of the Cα atoms between all extracted fragments are calculated. If there is no clear structural clustering (with an upper limit of 5.5 Å from the centroid of the cluster for the most divergent pair), additional side chain contact restraints are not derived; rather, only the seed contact is used. If the fragments cluster, the fragment closest to the average is selected, and additional tertiary restraints are extracted. The final outcome of the prediction protocol is a set of noisy secondary and tertiary restraints, whose accuracy and precision is discussed below in *Results*.

**Fold Assembly/Refinement/Selection.** Each residue in the protein consists of a Cα confined to a high coordination number lattice plus an off lattice, single-ball representation of its side chain (8). The potential incorporates terms reflecting statistical preferences for secondary structure, side chain burial, pair interactions, and hydrogen bond contributions. Furthermore, the predicted secondary and tertiary restraints are incorporated into an improved version of MONSSTER (9) that takes into account their accuracy and precision. A residue pair-dependent, flat bottom harmonic function is used, with a width 50% larger than the average contact distance extracted from a representative protein database. To accommodate inconsistencies between restraints operating between different secondary structural blocks, the restraints act on smoothed representations of the protein backbone (9). In addition, the predicted U-turn regions experience an energetic bias to lie at the protein surface. To improve the packing of putative β strands, an interstrand hydrogen bond cooperativity term is introduced in which β-type residues having hydrogen bonds to residues in two different strands are favored energetically.

All predictions use the identical parameter set and folding protocol. For each protein sequence, 10–40 independent, simulated annealing simulations from a fully extended initial conformation are carried out. The structures are then clustered. If no clustering is apparent (e.g., if, in each and every simulation, a different global fold is obtained), then the algorithm would be considered to have failed, and no tertiary structure prediction is made. For proteins considered below, this did not happen. If,

Table 1. Summary of prediction accuracy for tertiary contacts and results from the folding simulations

| Prot | Type | *n* | $Q_3$ | $N_p$ | $N_w$ | δ = 0 | δ = 2 | $rms_n$ | $E_n$ | σ | $rms_w$ | $E_w$ | σ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3cti | small | 29 | 82.4 | 6 | 0 | 83.3 | 100 | **3.8** | −107 | 7.4 | 6.7 | −103 | 7.8 |
| 1ixa | small | 39 | 97.4 | 5 | 0 | 100 | 100 | **5.6** | −130 | 7.0 | **7.7** | −131 | 8.0 |
| 1gpt | small | 47 | 72.3 | 13 | 0 | 46.1 | 100 | **5.9** | −276 | 12.4 | 6.6 | −142 | 7.0 |
| 1tfi | small | 50 | 78.0 | 37 | 0 | 21.6 | 88.8 | **5.9** | −202 | 7.3 | 7.0 | −191 | 9.4 |
| prtA* | α | 47 | 83.0 | 17 | 0 | 0.0 | 70.5 | **3.1** | −246 | 6.6 | 9.4 | −240 | 5.5 |
| 1ftz | α | 56 | 71.4 | 12 | 1 | 25.0 | 58.3 | **5.1** | −277 | 8.0 | 10.1 | −270 | 7.9 |
| 1c5a | α | 66 | 93.8 | 43 | 1 | 24.4 | 73.3 | **4.2** | −194 | 4.0 | 9.8 | −182 | 5.2 |
| 1pou | α | 71 | 84.5 | 49 | 0 | 28.6 | 89.8 | **3.5** | −418 | 3.4 | 11.9 | −364 | 4.0 |
| 3icb | α | 75 | 89.3 | 25 | 0 | 28.0 | 68.0 | **4.5** | −406 | 7.0 | 12.6 | −342 | 3.9 |
| 1hmd | α | 85 | 85.0 | 20 | 2 | 10.0 | 65.0 | **4.6** | −458 | 5.4 | **9.3** | −460 | 7.2 |
| 1shg | β | 57 | 64.9 | 39 | 0 | 28.2 | 100 | **4.5** | −420 | 4.5 | 6.7 | −397 | 5.5 |
| 1fas | β | 61 | 90.2 | 25 | 1 | 26.3 | 78.9 | **6.2** | −330 | 6.3 | 9.4 | −284 | 7.6 |
| 6pti | αβ | 56 | 80.4 | 19 | 0 | 68.4 | 100 | **4.7** | −410 | 10.0 | 9.7 | −397 | 10.0 |
| 1cis | αβ | 66 | 86.4 | 23 | 0 | 8.6 | 78.2 | **6.4** | −240 | 8.2 | 7.6 | −232 | 6.6 |
| 1lea | αβ | 73 | 87.5 | 41 | 2 | 9.7 | 75.6 | **6.1** | −136 | 7.7 | 9.4 | −115 | 7.5 |
| 1ubi | αβ | 76 | 77.6 | 17 | 0 | 23.5 | 94.1 | **6.1** | −238 | 6.3 | 11.5 | −203 | 5.7 |
| 1poh | αβ | 85 | 74.1 | 36 | 3 | 8.3 | 55.5 | **6.5** | −336 | 9.5 | 11.7 | −299 | 8.6 |
| 1ego | αβ | 85 | 71.8 | 33 | 0 | 15.1 | 93.9 | **5.7** | −417 | 8.9 | 9.0 | −396 | 15.0 |
| 1ife | αβ | 100 | 70.0 | 21 | 3 | 14.2 | 38.0 | **6.7** | −419 | 7.3 | **8.2** | −482 | 10.8 |

Prot, the Protein Data Bank access number; *n*, number of residues in the protein in the Protein Data Bank file; $Q_3$, percentage of correctly predicted secondary structure. All proteins have a $Q_3$ within 1 SD of the average; $N_p$, number of predicted contacts; $N_w$, number of contacts that are incorrect when δ = 5; δ = 0 and δ = 2, % of predicted contacts within δ residues of a native contact; $rms_n$, average cRMSD deviation in Å from the native structure for the family of structures having the lowest average energy; $E_n$, lowest average energy (in kT) after refinement for the native-like topology; σ, SD of the energy of the lowest average energy native-like structure; $rms_w$, average cRMSD deviation from native in Å of the alternative topology of lowest average energy; $E_w$, lowest average energy (in kT) in the alternative topology after refinement runs; σ, SD of the energy of the lowest average energy alternative topology structure.

*The B domain of protein A (21).

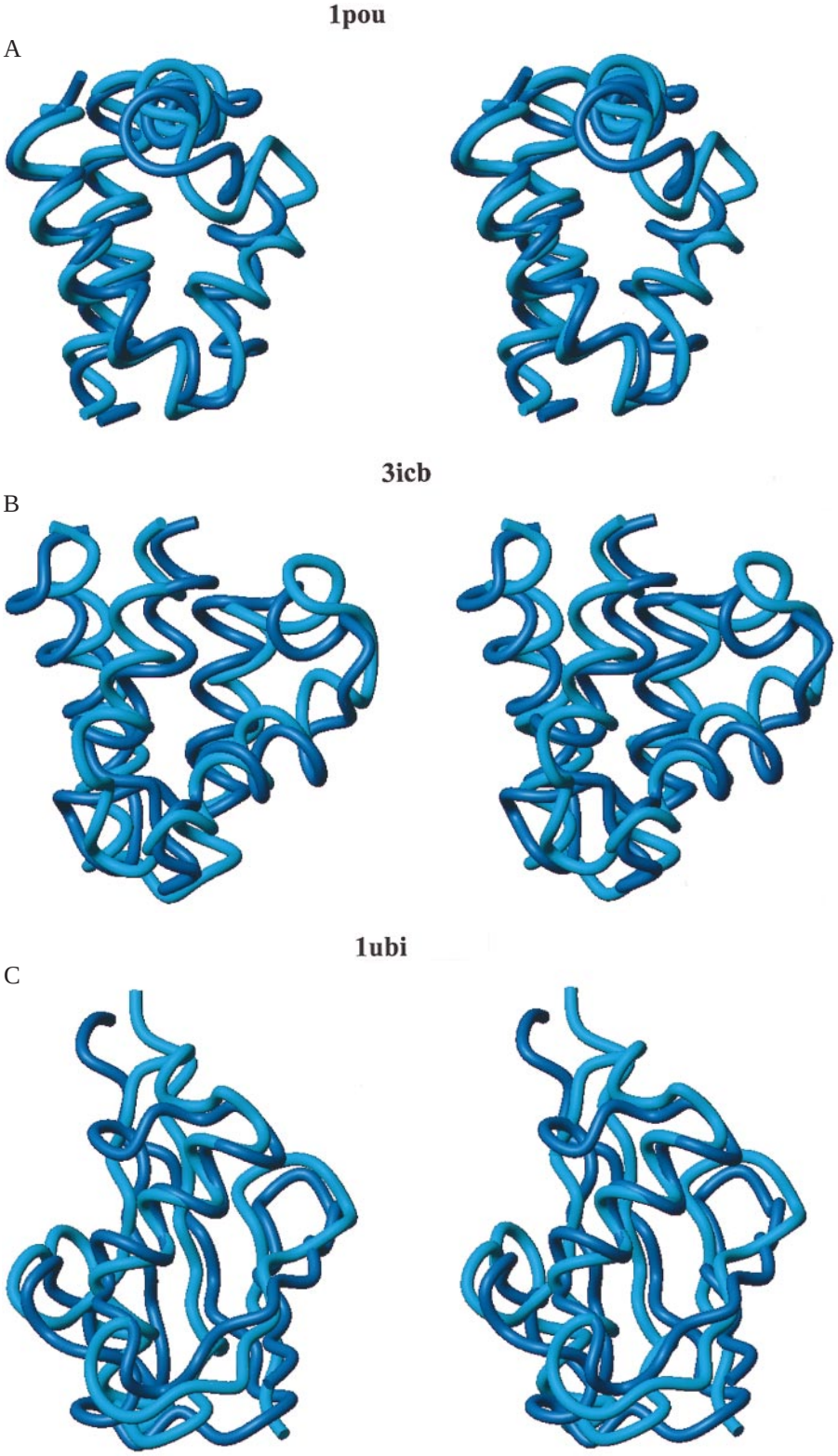


FIG. 2. Superimposition (in stereo) of the experimentally observed and predicted structure for two $\alpha$ proteins, 1pou and 3icb; two $\alpha/\beta$ proteins, 1ubi and 1ego; and two $\beta$ proteins, 1fas and 1shg. The structural superimposition was obtained by using DALI (18, 19). Blue indicates the experimental structure, and cyan shows the predicted structure. The structures are displayed by using MOLMOL (22). (*Figure continues on the opposite page*.)

however, a handful of global topologies occur, i.e., the structures cluster, then the assembled conformations are subject to low temperature, isothermal refinement. The predicted structure is the one having the lowest average (roughly 5 kT per residue) energy. If no specific topology can be unambiguously selected on an energetic basis (i.e., it is 2 SD lower in average energy then the best alternative fold), then the prediction consists of the handful of distinct lowest energy topologies that result.

In practice, this approach is found to produce a subset of structures with a cRMSD often $\approx$6 Å from native. At this level of cRMSD, the structures might or might not have gross topological errors. Thus, an additional objective assessment of the ability of the method to assemble these native-like struc-

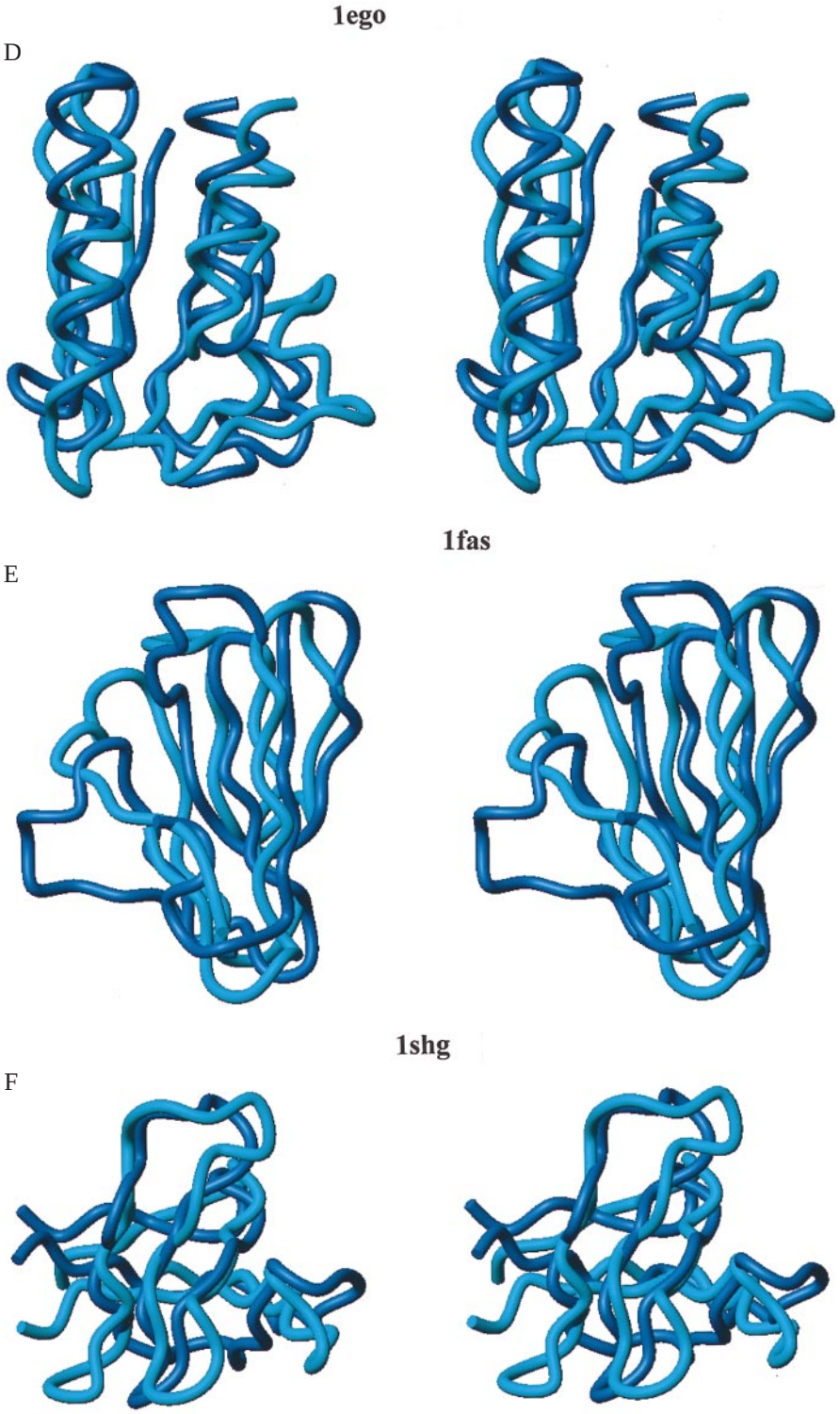


FIG. 2. (*Continued.*)

tures was performed as follows: The predicted folded conformation was subjected to a structural similarity search over a representative subset of the protein database by using the structure superimposition program DALI (18, 19). The topology of the hit found with the highest score was then analyzed and compared with that of the native structure. If both structures shared the same topology, then the predicted fold was more similar to the native fold than to any other alternative fold in the protein structural database. In such cases, our tertiary structure prediction protocol is considered to have successfully predicted the native topology. A similar approach has been applied to assess the quality of structures predicted by threading methods (20).

## RESULTS AND DISCUSSION

The above protocol was applied to the 19 proteins listed in Table 1, demonstrating that the approach can handle a wide variety of folds and different secondary structure types. Table 1 shows the accuracy of the predicted secondary structure and tertiary contacts as well as the results from the folding simulations. The test

Table 2. Results of the structural search* of folds most similar to the predicted fold in a database of representative protein structures by using the structure superimposition program DALI (19) (column 2) and results of the structural alignments of the predicted conformation with the experimental conformation (columns 3–6)

| Prot | HIT | $Z_{scr}$ | rms | LA[a] | Structural alignment |
|---|---|---|---|---|---|
| 3cti | ---- | --- | --- | -- | ---------- |
| | | | | | ---------- |
| 1ixa | 1edm | 0.6 | 3.0 | 30 | 6-11, 13-22, 24-37 |
| | | | | | 6-11, 14-23, 24-37 |
| 1gpt | 1sco | 0.2 | 2.9 | 33 | 6-12, 16-27, 28-36, 38-42 |
| | | | | | 8-14, 16-27, 29-37, 39-43 |
| 1tfi | ---- | --- | --- | -- | ---------- |
| | | | | | ---------- |
| prtA | 1edl | 2.7 | 2.5 | 42 | 2-15, 17-23 |
| | | | | | 2-15, 17-23 |
| 1ftz | 1lfb | 0.9 | 3.2 | 45 | 3-6, 7-15, 16-19, 22-37, 42-45, 47-50, 52-55 |
| | | | | | 3-6, 8-16, 19-22, 23-38, 42-45, 46-49, 50-53 |
| 1c5a | 2adk | 2.0 | 3.0 | 48 | 3-8, 12-15, 20-23, 28-38, 42-48, 49-54, 56-65 |
| | | | | | 3-8, 10-13, 23-26, 28-38, 42-48, 50-55, 56-65 |
| 1pou | 1oct-C | 5.0 | 2.7 | 65 | 1-24, 26-40, 42-46, 50-59, 61-71 |
| | | | | | 2-25, 26-40, 41-45, 51-60, 61-71 |
| 3icb | 1wde | 3.7 | 3.3 | 64 | 1-19, 25-31, 32-36, 38-58, 64-75 |
| | | | | | 1-19, 25-31, 33-37, 38-58, 64-75 |
| 1hmd | 1cei (2hmr) | 5.0 | 4.1 | 76 | 1-18, 22-28, 29-45, 46-67, 72-83 |
| | | | | | 4-21, 22-28, 30-46, 49-70, 72-83 |
| 1shg | 1abo-A | 0.9 | 4.0 | 43 | 4-7, 8-16, 25-28, 29-32, 36-57 |
| | | | | | 5-8, 10-18, 25-28, 30-33, 36-57 |
| 1fas | 3ebx | 0.4 | 3.8 | 41 | 5-8, 17-28, 30-37, 40-56 |
| | | | | | 6-9, 17-28, 32-39, 41-57 |
| 6pti | ---- | --- | --- | -- | ---------- |
| | | | | | ---------- |
| 1cis | 2sec-I | 1.0 | 2.6 | 40 | 2-7, 8-11, 14-19, 27-32, 36-39, 45-48, 49-53, 56-60 |
| | | | | | 1-6, 9-12, 13-18, 28-33, 35-38, 44-47, 49-53, 60-64 |
| 1lea | ---- | --- | --- | -- | ---------- |
| | | | | | ---------- |
| 1ubi | 1ubq | 2.5 | 3.4 | 58 | 1-7, 9-18, 21-37, 41-46, 56-59, 61-64, 67-76 |
| | | | | | 1-7, 11-20, 21-37, 39-44, 45-48, 59-62, 66-75 |
| 1poh | 1rth_A | 1.9 | 3.2 | 57 | 2-5, 15-29, 31-34, 42-45, 53-56, 58-61, 63-69, 70-80, 82-85 |
| | | | | | 1-4, 15-29, 33-36, 48-51, 55-58, 59-62, 63-69, 71-81, 82-85 |
| 1ego | 1grx | 4.3 | 3.0 | 68 | 1-7, 10-26, 27-38, 40-45, 60-65, 66-85 |
| | | | | | 2-8, 10-26, 29-40, 48-53, 58-63, 66-85 |
| 1ife | 1tig (1ife) | 2.8 | 3.7 | 69 | 1-20, 23-28, 31-43, 45-63, 75-80, 83-87 |
| | | | | | 2-21, 23-28, 30-42, 49-67, 71-76, 83-87 |

Prot, Protein Data Bank access number; HIT, first hit (according to the Z score value) of the structural alignment of the predicted conformation against the set of DALI representative folds of the protein data base. Bracketed names represent second hits; $Z_{scr}$, statistical significance (Z score) of the structural alignment of the predicted and experimental structures, as defined in the DALI method; rms, coordinate RMSD between the predicted and experimental structure for the structural alignment shown in column 6; LA, number of residues found in the structural alignment; Structural alignment, residues aligned. The residue numbering scheme refers to the sequential numbering from the N to C terminus and not to the numbering scheme in the Protein Data Bank file. For each entry, the first row corresponds to the predicted conformation and the second row to the experimental one.
*In the case of 6pti, 1lea, 1tfi and 3cti no match was obtained using DALI.

set of proteins used in this work has an average fraction of a correctly predicted secondary structure, $Q_3$, of ≈82%. This is ≈10% higher than the average performance of the PHD method, mainly because, here, in most cases, all of the secondary structural elements present in the native protein are more or less correctly predicted. On the other hand, for the set of examined proteins, the contact map prediction method provides predicted contacts ≈25% of which are exactly correct and ≈75% of which are correct within ±2 residues. Approximately 20–25% of the total number of contacts seen in the native conformation are obtained by our combined seed derivation–enrichment approach to contact prediction.

In ≈10–30% of the assembly runs, the simulations yield native-like topologies as assessed by global cRMSD and DALI (18, 19), with the yield depending on the complexity of the global fold. In Table 1, we also present the cRMSD from native (based on the full sequence length) of structures having the native-like topology as well as the best (lowest average energy) alternative fold. The average cRMSD of the lowest average energy structures corresponding to the native topology ranges from about 3 Å for some helical proteins to roughly 6 Å for $\beta$ and $\alpha/\beta$ proteins. Because a structure with a cRMSD of 6 Å might contain errors in the global topology, to further illustrate the fidelity of the predictions, we present in Fig. 2 a representative set of six predicted structures alongside the experimentally determined conformation. The remaining 13 structures will be made available on our World Wide Web site (http://www.scripps.edu/skolnick/ORTIZ/ortiz.html). We emphasize that we do not report the lowest observed cRMSD but the values corresponding to the lowest average energy and the next excited state. As indicated in Table 1, native-like conformations can be obtained either as the best average energy in 16 of the 19 cases studied or as the next best energy structure in the remaining three cases. However, in most cases, the SD of the energy in a given structure is of the order of the energy difference between the average energy values, i.e., the energy spectra substantially overlap. Whether this is a physical

effect is uncertain, but it certainly complicates the selection of a particular low energy topology as being native and makes the structure selection unreliable. Additionally, there are three cases in which a misfolded or partly misfolded state is selected as being the lowest in average energy: 1ixa, 1hmd, and 1ife. The misfolded state of 1ixa results from the wrong placement of the C-terminal $\beta$ strand. In 1hmd, the topological mirror-image, four-helix bundle (where the chirality of the helices remains right-handed, but the chirality of the turns relative to the native fold is reversed) is essentially isoenergetic with the native fold. The final misidentified protein, 1ife, has the same global topology as native, but a strand is shifted from the edge of the front $\beta$ sheet to the back of the protein. However, it is worth mentioning that a structural alignment in the protein database (*vide infra*) using these partially (1ixa, 1ife) or globally (1hmd) misfolded conformations finds the native fold as the first (1ixa) or second (1ife, 1hmd) hit.

Results of the DALI structural comparison (18, 19) of the predicted folds are presented in Table 2. Two different types of computational experiments are presented. In the first one, we tried to establish whether the predicted conformation is more similar to the target fold than to any other alternative fold in a representative database of protein structures. Results are reported in the second column of Table 2, which shows the first hit found. In the cases of 6pti, 1lea, 1tfi, and 3cti, no hit was obtained, i.e., DALI did not find a significant structural relationship between the predicted and native structure, although in a number of these, the overall cRMSD was low. No match was found for 3cti. As for the rest of the structures, the first hit found corresponds to a close structural homologue of the target structure, except in the case of 1poh, in which a different fold was found (1rth_A). There are two other cases that are worth mentioning. The first case is 1hmd, for which a four-helix bundle is selected (1cei) but with a different angle between the helices. The target structure (2hmr) is selected as the second hit. The second case is 1ife, for which the correct fold is identified as the first hit (1tig), but the structural alignment between the two structures runs from the N to C terminus in one case and the C to N terminus in the other. This is because the contact map alignment method used by DALI cannot distinguish between the two possibilities. As the second best hit, the structure of 1ife is selected (Table 2). In our second computational experiment, we tried to establish whether it is possible to define a high resolution core in the predicted structures and whether some of the high global cRMSD values reported in Table 1 just reflect shifts in registration of the secondary structure elements and poor positioning of (mainly loop) regions in the structure having a low density of predicted restraints. Columns 4 and 5 of Table 2 indicate that, on average, it is possible to produce a structural alignment between the experimental and the predicted conformations covering $\approx$80% of the sequence length with a cRMSD between 2.5 and 4.1 Å. The structural alignment shown in column 5 of Table 2 shows that, to produce these cRMSD values, shifts in registration take place between the predicted and experimental structures. Thus, we conclude that this *ab initio* folding approach produces low resolution, native-like topologies of comparable quality to threading methods.

## CONCLUSIONS

In this work, we have explored the possibility of using predicted restraints derived from multiple sequence alignments in *ab initio* folding simulations aimed at the structure prediction of small proteins. On the basis of this study, the following conclusions can be drawn. First and most important, low resolution but native-like models of small proteins can be assembled from inaccurate tertiary contact predictions of a subset (20–25%) of the total number of tertiary contacts in the native protein. Here, $\approx$75% of the predicted contacts were correct within $\pm$2 residues. This is true provided that the number of totally wrong contacts is minimized (see the list of $N_w$ values compiled in Table 1). Second, at the level of secondary structural elements, the accuracy of contemporary secondary structure prediction schemes is adequate for successful structure assembly. As demonstrated for 1shg, 1 gpt, and 1ife, even when an entire element of secondary structure is missed, depending on its location in the native conformation, this failure does not necessarily prohibit the successful prediction of tertiary structure. Finally, when sequences with the same number of residues are considered, helical proteins are predicted with higher accuracy than $\alpha/\beta$ proteins and $\beta$ proteins (7, 9).

Based on our experience to date in these systems as well as in blind predictions of other proteins (unpublished results), one of two situations result when we attempt a tertiary structure prediction. In the first and worst case, the structures do not cluster on repeated simulations. This situation occurs in $\beta$ proteins larger than $\approx$100 residues in length. If so, the prediction has failed. If the resulting folds do cluster, then invariably the native conformation is among the handful of topologies. Thus far, this is typical of proteins whose tertiary contact prediction accuracy is on the level seen here and that are smaller than 100 residues in length. Turning to the nature of the folds that are observed, especially for helical proteins, either the native topology or its topological mirror image result. For $\alpha/\beta$ and $\beta$ proteins, typically a common (native-like) structural core is predicted, and the topologies differ in the placement of one or two secondary structural elements. The observation that the native fold is contained among a few possible conformations constitutes the robust and computationally rapid part of the prediction algorithm. The selection of the native conformation from this small number of possible folds is more difficult, more computationally intensive, and more unreliable. This finding clearly indicates that methodological improvements are required. Such improvements will entail modifications in the tertiary restraint derivation and the potentials, as well as the conformational sampling protocol. Nevertheless, the present study points to a promising methodology for the prediction of low resolution tertiary structures of small proteins, although additional investigation is required to establish the full extent of its generality.

This work was supported in part by National Institutes of Health Grant GM37408. A.R.O. also acknowledges partial support from the Spanish Ministry of Education. A.K. also acknowledges partial support from University of Warsaw Grant BST-34/97.

1. Ptitsyn, O. B. (1987) *J. Protein Chem.* **6,** 273–293.
2. Anfinsen, C. B. (1973) *Science* **181,** 223–230.
3. Piela, L., Kostrowicki, J. & Scheraga, H. A. (1989) *J. Phys. Chem.* **93,** 3339–3346.
4. Dandekar, T. & Argos, P. (1996) *J. Mol. Biol.* **256,** 645–660.
5. Smith Brown, M. J., Kominos, D. & Levy, R. M. (1993) *Protein Eng.* **6,** 605–614.
6. Aszodi, A., Gradwell, M. J. & Taylor, W. R. (1995) *J. Mol. Biol.* **251,** 308–326.
7. Skolnick, J., Kolinski, A. & Ortiz, A. R. (1997) *J. Mol. Biol.* **265,** 217–241.
8. Mumenthaler, C. & Braun, W. (1995) *Protein Sci.* **4,** 863–871.
9. Ortiz, A. R., Kolinski, A. & Skolnick, J. (1998) *J. Mol. Biol.*, in press.
10. Sander, C. & Schneider, R. (1991) *Proteins* **9,** 56–68.
11. Rost, B. & Sander, C. (1993) *J. Mol. Biol.* **232,** 584–599.
12. Rost, B., Schneider, R. & Sander, C. (1993) *Trends Biochem. Sci.* **18,** 120–123.
13. Kolinski, A., Skolnick, J., Godzik, A. & Hu, W. P. (1997) *Proteins* **27,** 290–308.
14. Goebel, U., Sander, C., Schneider, R. & Valencia, A. (1994) *Proteins* **18,** 309–317.
15. Thomas, D. J., Cesari, G. & Sander, C. (1996) *Protein Eng.* **11,** 941–948.
16. Olmea, O. & Valencia, A. (1997) *Folding Design* **2,** S25–S32.
17. Godzik, A., Skolnick, J. & Kolinski, A. (1992) *J. Mol. Biol.* **227,** 227–238.
18. Holm, L. & Sander, C. (1993) *J. Mol. Biol.* **233,** 123–138.
19. Holm, L. & Sander, C. (1997) *Nucleic Acids Res.* **25,** 231–234.
20. Miller, R. T., Jones, D. T. & Thornton, J. M. (1996) *FASEB J.* **10,** 171–178.
21. Gouda, H., Torigoe, H., Saito, A., Sato, M., Arata, Y. & Shimada, I. (1992) *Biochemistry* **40,** 9665–9672.
22. Koradi, R., Billeter, M. & Wuethrich, K. (1996) *J. Mol. Graph.* **14,** 51–55.