J. Mol. Biol. (1998) **277**, 419–448

JMB

Fold Assembly of Small Proteins Using Monte Carlo Simulations Driven by Restraints Derived from Multiple Sequence Alignments

Angel R. Ortiz[1], Andrzej Kolinski[1,2] and Jeffrey Skolnick[1]*

[1]*Department of Molecular Biology, TPC-5, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla CA 92037, USA*

[2]*Department of Chemistry University of Warsaw ul. Pasteura 1, 02-093 Warsaw Poland*

**Corresponding author*

The feasibility of predicting the global fold of small proteins by incorporating predicted secondary and tertiary restraints into *ab initio* folding simulations has been demonstrated on a test set comprised of 20 non-homologous proteins, of which one was a blind prediction of target 42 in the recent CASP2 contest. These proteins contain from 37 to 100 residues and represent all secondary structural classes and a representative variety of global topologies. Secondary structure restraints are provided by the PHD secondary structure prediction algorithm that incorporates multiple sequence information. Predicted tertiary restraints are derived from multiple sequence alignments *via* a two-step process. First, seed side-chain contacts are identified from correlated mutation analysis, and then a threading-based algorithm is used to expand the number of these seed contacts. A lattice-based reduced protein model and a folding algorithm designed to incorporate these predicted restraints is described. Depending upon fold complexity, it is possible to assemble native-like topologies whose coordinate root-mean-square deviation from native is between 3.0 Å and 6.5 Å. The requisite level of accuracy in side-chain contact map prediction can be roughly 25% on average, provided that about 60% of the contact predictions are correct within ±1 residue and 95% of the predictions are correct within ±4 residues. Precision in tertiary contact prediction is more critical than absolute accuracy. Furthermore, only a subset of the tertiary contacts, on the order of 25% of the total, is sufficient for successful topology assembly. Overall, this study suggests that the use of restraints derived from multiple sequence alignments combined with a fold assembly algorithm holds considerable promise for the prediction of the global topology of small proteins.

Keywords: protein structure prediction; correlated mutations; side-chain contact prediction; lattice protein models; secondary structure prediction; threading

Introduction

At present, the prediction of protein structure from amino acid sequence remains one of the major unsolved problems in molecular biology. The solution to this problem demands the development of effective conformational search algorithms and the formulation of potentials capable of recognizing the native state from the manifold of misfolded structures. Reduced protein models having one or a few interaction centers per residue and statistical potentials extracted from protein structure databases offer a reasonable way to address both issues (Godzik *et al.*, 1994; Kolinski *et al.*, 1995a,b; Kolinski & Skolnick, 1994a; Park & Levitt, 1995; Skolnick *et al.*, 1993). Using such approaches, several successful *ab initio* predictions of simple topologies have been reported. For example, the conformation of short peptides such as melittin (Ripoll & Scheraga, 1990), pancreatic polypeptide inhibitor (Sun, 1993; Wallqvist & Ullner, 1994), apamin (Sun, 1993), and PthrP (Wallqvist & Ullner, 1994) have been predicted with a backbone RMSD ranging from 1.7 Å to 4.5 Å. Lattice folding simulations (Kolinski & Skolnick, 1994b) of the B domain of Protein A (Gouda *et al.*, 1992) have

Abbreviations used: RMSD, root-mean-square deviation; cRMSD, coordinate RMSD; PDB, Protein Data Bank.

0022–2836/98/120419–30 $25.00/0/mb971595

yielded structures with a C^{α} RMSD from native on the order of 3.0 Å. The most accurate predictions to date are those of the GCN4 leucine zipper, whose final predicted C^{α} RMSD is 0.8 Å from the native crystal structure (O'Shea *et al*., 1991; Vieth *et al*., 1994). In general, these methods have been mainly successful on helical proteins having simple global folds. For more complex structures of natural proteins, *ab initio* folding has failed thus far. Simplified protein representations and inaccuracies in the attendant potentials can conspire to yield an inability to identify the native topology from alternative non-native structures, especially those with a substantial similarity to the native fold. Furthermore, as protein size and topological complexity increases, conformational sampling becomes exponentially more problematic. Thus, alternative approaches are required that can surmount these difficulties.

Introduction of secondary structure restraints obtained from secondary structure prediction algorithms is a natural extension of pure, restraint free *ab initio* folding. This is a particularly appealing idea since knowledge of the native secondary structure elements enormously reduces the conformational space that must be searched. Recently, the accuracy of secondary structure prediction has improved from about 65% to 72%, on average (Rost & Sander, 1996b). This development lends credence to the idea that secondary structural elements can be identified with reasonable accuracy. Some progress is also being made in algorithms that can predict regions where the chain reverses global direction, *viz*., U-turns (Kolinski *et al*., 1997). The feasibility of biasing search algorithms with secondary structure knowledge was first explored using ''exact'' secondary structure, as observed in the experimental native conformation. In this regard, using an off-lattice model and exact knowledge of the native secondary structure, Friesner *et al*. (1996) have obtained very encouraging results, successfully folding two four-helix bundles (cytochrome *b*562 (B256) and myohemerythrin (2MHR)), a large α-helical protein (myoglobin (1MBO)), and a relatively complicated α/β fold, the C-terminal domain of the L7/L12 50 S ribosomal protein (1CTF) (Friesner & Gunn, 1996; Gunn *et al*., 1994). Similarly, Dandekar & Argos (1994, 1996), using a genetic algorithm to search conformational space, obtained encouraging results for a test set of 19 small proteins, including all α and some α/β proteins. In their studies, they have succeeded in predicting a significant proportion of these small proteins at ~5 Å resolution. However, Dandekar & Argos have also observed that use of predicted secondary structure information produces a substantial deterioration in the performance of their prediction algorithm. As a practical matter, however, any successful tertiary structure assembly algorithm must be able to successfully handle predicted secondary structural information whose accuracy is at the current state-of-the-art, and tests using predicted secondary structure should be made. Along these lines, an early example of where predicted secondary structure was incorporated as restraint information in a subsequent topology assembly algorithm is due to Kolinski & Skolnick (1994b). The resulting predicted structure of crambin had a backbone C^{α} RMSD of 3.2 Å. A more recent example is due to Simons *et al*. (1997), who instead of secondary structure predictions used an interesting technique to derive short-range conformational preferences from multiple sequence alignments. Some α-helical proteins could be assembled using this approach, although the potential function used by the authors did not allow the native-like topology to be discriminated from alternative answers. However, it was apparent from these and other studies that knowledge of secondary structure preferences alone does not entirely eliminate competing misfolded states. Furthermore, secondary structure bias is local in nature and, therefore, does not provide a gradient in the conformational energy landscape that can funnel the conformation towards the native state. Thus, problems with both potentials and conformational search protocols still remain.

Funneling can be efficiently obtained through the use of long-range (in sequence) distance restraints. A number of workers have begun to examine the feasibility of such an approach. For example, Smith-Brown *et al*. (1993) have attempted to predict several protein folds by assuming exact knowledge of the secondary structure and a subset of interresidue distance restraints encoded as a biharmonic potential. They find that a considerable number of restraints per residue is required to assemble the fold, making the approach impractical for most prediction purposes. Another interesting study is due to Aszodi & Taylor (1996), who assumed correct native secondary structure and a set of simulated tertiary restraints (Aszodi *et al*., 1995). Here in an attempt to build the protein core, restraints were supplemented by a set of interresidue distances based on patterns of conserved hydrophobic residues obtained from a multiple sequence alignment. Folds were then assembled using distance geometry with a simplified protein chain model. Aszodi & Taylor (1996) were able to assemble structures below 5 Å RMSD when at least $N/4$ restraints are used, where N is the number of protein residues. However, with their force field, they have excessive difficulties selecting the correct fold from competing alternatives. Along similar lines, Mumenthaler & Braun (1995) developed an interesting self-correcting distance geometry method that tries to automatically eliminate wrongly predicted contacts derived from multiple sequence alignments. Again, the correct secondary structure is assumed, but now totally predicted tertiary restraints, based on the conservation patterns of hydrophobic residues in multiple sequence alignments, are used to assemble the fold. With this method, encouraging results have been obtained with the successful folding to the native topology in six out of eight helical proteins stu-

died. Still, a significant number of restraints are required in all these approaches. This poses a problem because no prediction technique is available that can provide both the requisite number and accuracy of secondary and tertiary restraints that these approaches demand for more complex folds.

One step towards addressing these problems was made recently by Skolnick *et al*. (1997b). They developed a new program, called MONSSTER (MOdeling of New Structures from Secondary and TErtiary Restraints), that is able to successfully fold small proteins using a considerably smaller number of distance restraints than previous approaches demanded. It was found that when ''exact'' restraints are available, helical proteins can be folded with roughly $N/7$ restraints, while all β and α/β proteins require about $N/4$ restraints, where N is the number of residues in the protein chain. Of course, for any particular case, the accuracy depends on restraint distribution and fold complexity. MONSSTER employs a lattice-based reduced representation of the protein chain. In addition to secondary and tertiary restraints, there is a potential that incorporates statistical preferences for secondary structure, side-chain burial and pair interactions, together with a hydrogen bond potential. We term these non-restraint contributions inherent interactions. The resulting fold accuracy is substantially degraded when these inherent contributions to the potential are eliminated. Thus, these simulations indicated that there is a complementarity between the inherent contributions to the potential and the supplementary but crucial information provided by secondary and tertiary restraints.

The encouraging results obtained with this model prompted us to attempt the next logical step. Namely, we explored the possibility of assembling global protein topologies using entirely predicted secondary and tertiary restraints. Predicted restraints are noisy in nature and it is unclear whether an algorithm that works within the limit of a very small number of correct restraints is robust enough to successfully handle the unavoidable presence of incorrect predictions. Extant secondary structure prediction schemes provide a logical jumping off point for the incorporation of predicted secondary structure information. The protocol for predicting tertiary restraints is less obvious. Following on the ideas of Göebel *et al*. (1994), predicted tertiary contacts are extracted on the basis of evolutionary information contained in multiple sequence alignments, complemented with threading calculations (our unpublished results). In sequence alignments, since some pairs of positions appear to exhibit a covariation in their mutational behavior consistent with their physical and chemical properties, it has been suggested that spatially close neighbors might be more likely to exhibit such behavior. Using statistical techniques, this effect has been quantified by a method known as correlated mutation analysis (Göebel *et al*., 1994). It has been shown that, by applying a stringent significance cut-off in the prediction of contacts by correlated mutations, a small number of contacts can be predicted that are a factor of 1.4 to 5.1 times better than random. Previously, the number of correct contacts obtained this way has been either too small to permit successful tertiary structure assembly if a high significance cut-off was used to avoid false positives, or too noisy if the number of contacts selected was that demanded by existing assembly algorithms (Rost & Sander, 1996a). Here, it is shown that, for a representative set of proteins, a modification of the correlated mutation analysis approach (our unpublished results), when coupled to secondary structure prediction and followed by structure assembly using a version of MONSSTER updated to handle incorrect predictions, is able to bridge the gap between sequence analysis and folding simulations. This permits the *ab initio* folding of some complex topologies.

The outline of the remainder of the paper is as follows. In Methods, we describe the approach followed in this work, which can be logically divided into two parts: secondary and tertiary restraint derivation, and fold assembly/refinement using MONSSTER. In the Results, we describe its application to a set of 20 representative single domain proteins. This is followed by the Discussion, which examines possible reasons why the current approach can be successful and delineates the improvements required, in both the restraint derivation procedure and in the fold assembly/refinement protocol, to make this approach generally applicable. Finally, in the Conclusions, we summarize the current state-of-the-art of protein structure prediction as provided by MONSSTER.

Methods

Overview

A flow chart of the tertiary structure prediction protocol is depicted in Figure 1. The procedure presented in this work can be logically divided into two parts: restraint derivation and structure assembly/refinement using the MONSSTER algorithm. With respect to restraint derivation, the first objective is to predict the number, location, and identity of the dominant secondary structural elements that will comprise the protein. These consist of helices and β-strands, termed here the core topological elements of the molecule. In addition, U-turns between these secondary structure elements are predicted (Kolinski *et al*., 1997). Next, we try to predict the secondary structure elements in contact. This is attempted by obtaining the most reliable set of predicted contacts between core elements using correlated mutation analysis. We denote the contacts obtained in this way as ''seeds''. This protocol allows us to maximize the signal-to-noise ratio of predicted contacts. However, the number of seeds is still insufficient to allow successful fold assembly to occur. Hence, we exploit the fact that packing patterns between sec-

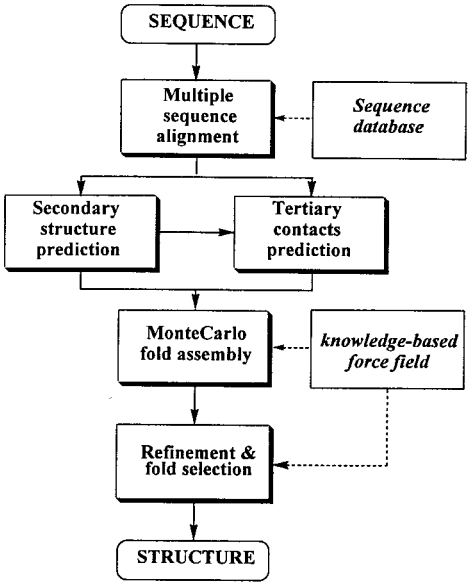


Figure 1. Flow chart of the protein fold prediction method.

ondary structure elements are degenerate. Seed contacts are ''enriched'' by finding the most compatible contact map in the structural database given the predicted seed and the secondary structure elements involved by using a fragment clustering/inverse folding protocol (Godzik *et al.*, 1992). Full details of this procedure will be given in a forthcoming publication. The key point is that the resultant restraints are not particularly accurate and that approaches incorporating such restraints must be adjusted to account for their ambiguity. This is accomplished in an updated version of MONSSTER designed to accommodate the inherent inaccuracies of such restraints, as is described below. The general features of the force field and representation are the same as those described earlier (Skolnick *et al.*, 1997b). Thus, we focus here only on those aspects that differ from the previous implementation in order to adapt the procedure to incorrect and/or ambiguous restraints.

Restraint derivation method

Secondary structure prediction

Multiple sequence alignments for each of the proteins studied were obtained from the HSSP database (Sander & Schneider, 1991). This alignment is used as input for the PHD (Rost & Sander, 1993) secondary structure prediction method. For the purposes of hydrogen bond assignment, all predicted strand elements are assumed to correspond to a strand in the real secondary structure. For helices, only those elements with a reliability index higher than three are used. Chain reversals are predicted by the U-turn prediction algorithm LINKER developed by Kolinski *et al.* (1977). Because of their reliability, elements predicted as U-turns override PHD predictions (Kolinski *et al.*, 1997). In practice, each residue can be assigned to one of five conformational states: a predicted extended state, a predicted helix, a predicted U-turn, a β-(strand) state or a non-predicted state. The set of predicted helices and strands comprise the putative core elements of the protein.

Side-chain contact prediction

The prediction of residue contacts is performed in two stages. First, a correlated mutation analysis (Göebel *et al.*, 1994) of the multiple sequence alignment is done to identify the seed contacts. The procedure is based on defining an exchange matrix or other similarity measure at each sequence position in a multiple sequence alignment. One then calculates the correlation coefficient between exchange matrices at any two positions. In the calculation of the covariance matrix, regions containing deletions and insertions are not considered. Here, residue comparison is carried out using the McLachlan (1971) matrix. In this work, the same multiple sequence alignment is used for secondary structure and correlated mutation analysis. Only correlations between elements predicted to be in core regions (and not U-turns) are considered. The rationale is that by restricting the predictions to rigid elements of the putative core, the assumption of closeness in space for positions showing covariance in their mutational behavior might be more valid. Correlation is measured by a Pearson-type correlation coefficient:

$$r_{ij} = \frac{1}{N^2} \sum_{kl} \frac{(s_{ikl} - \langle s_i \rangle)(s_{jkl} - \langle s_j \rangle)}{\sigma_i \sigma_j} \qquad (1)$$

Here, i and j are two different positions in a multiple sequence alignment, and the indices k and l run from 1 to the number N of sequences in the family. The parameter s_{ikl} (s_{jkl}) is the comparison score (according to the McLachlan (1971) mutation matrix) of the amino acids of sequences k and l at position $i(j)$ of the alignment. Average values over all aligned sequences at positions i and j are given by $\langle s_i \rangle$ and $\langle s_j \rangle$. The parameters σ_i and σ_j correspond to the standard deviation of the scores in the positions i and j, respectively. A correlation coefficient cut-off threshold of 0.5 is used for contact prediction. At most, only one contact per each pair of core secondary structure blocks is used. Thus, this analysis delineates predicted secondary structure elements in contact. In addition, the maximum number of seeds allowed is equal to the maximum number of expected contacts (n_s) between the L of predicted secondary structure elements. This value is obtained from a representative database of small proteins and is roughly given by $n_s = L\ (L-1)/4$ (our unpublished results). However, correlated

mutation analysis only provides a few seed side-chain contacts; their number is generally insufficient to assemble a protein from the unfolded state using MONSSTER. Thus, the number of side-chain contacts needs to be increased.

The set of seed restraints is enriched by a combined structural fragment search and threading folding procedure (Godzik *et al.*, 1992). All pairs of secondary structure elements compatible with the predicted secondary structure types and predicted contacts are extracted from a structural database. This structural database is the same used by Hu *et al.* (1997). For each tested sequence, homologues to the full target sequence found in the database were removed prior to the application of the growing procedure. To account for the inaccuracies in the correlated mutation analysis, a tolerance of a one-residue shift in each member of the contacting residue pair is allowed. Fragments are then scored by a statistical potential that considers local conformational propensities and the burial energy within the pair of fragments (Godzik *et al.*, 1992). Pair and higher order interactions are ignored (Godzik *et al.*, 1992; Hu *et al.*, 1997) to avoid the imbalance between intra and extra fragment interactions, which would result if such contributions were included. The top ten scoring fragments are superimposed in space by minimizing their coordinate RMSD, and then clustered on the basis of their pair-wise RMSD. If they do not show a clear clustering (with an upper limit of 5.5 Å for the most divergent fragment pair), then additional side-chain contact restraints are not derived. Conversely, if the fragments spatially cluster, then the fragment within this cluster whose RMSD is smallest, with respect to all other members, is selected and its side-chain contact map is projected onto the query sequence. Following this procedure, the number of predicted contacts usually increases by about a factor of five with respect to that predicted from correlated mutation analysis alone. Full details of this procedure will be given in a forthcoming publication (A.R.O. & J.S., unpublished results).

Assembly and refinement protocol

Protein model

Geometric properties. The C^{α} coordinates of the protein backbone are confined to a set of lattice points located on an underlying cubic lattice whose lattice spacing, $a = 1.22$ Å (Kolinski & Skolnick, 1994a). Successive C^{α} atoms are connected by a set of 90 virtual bond vectors $a \cdot \mathbf{v}$, with $\{\mathbf{v}\} = \{(\pm 3, \pm 1, \pm 1), \ldots (\pm 3, \pm 1, 0), \ldots (\pm 3, 0, 0), \ldots (\pm 2, \pm 2, \pm 1), \ldots (\pm 2, \pm 2, 0), \ldots\}$. The distance a is chosen so that the mean C^{α} virtual bond length is 3.8 Å. Side-chains are represented by a set of rotamers, each located at the side-chain center of mass. They are not restricted to lattice points. With the exception of Gly, Pro and Ala, there are multiple rotamers for each amino acid, chosen so that the center of mass of a side-chain in real proteins will be no farther than 1 Å from some member of the rotamer library. For more details, see Kolinski & Skolnick (1994a, 1997b).

Interaction scheme

Inherent contributions. This class of terms is independent of the restraint predictions and is designed to capture both generic (sequence independent) and sequence-specific protein-like features. Many of these contributions are identical to those described in MONSSTER (Skolnick *et al.*, 1997b). Such terms include an amino acid pair specific potential that describes the intrinsic secondary structural preferences, E_{14}, and a one-body centrosymmetric burial potential, E_1. Here, to avoid non-physical segregation of the subunits, we have added a packing density regularizer, E_{density} (Kolinski & Skolnick, 1997). This term is designed to ensure that the average overall density distribution of residues in native proteins is reproduced. This term provides a strong compressive force in the unfolded state, but contributes negligibly in compact states. A more sensitive side-chain pair contact potential, E_{pair}, that has been derived by a more careful analysis of the appropriate reference state is now used (Skolnick *et al.*, 1997a). Hydrogen bonds are C^{α}-based and very much in the spirit of Levitt & Greer (1977).

Restraint contributions. Predicted secondary structures and tertiary contacts are implemented into the model in the form of restraint contributions to the conformational energy. Furthermore, a set of somewhat refined knowledge-based restraints designed to reproduce the packing of supersecondary structural elements is used. The implementation of each type of restraint is discussed in turn.

Secondary structure dependent restraint contributions

Local secondary structure bias. Secondary structure bias is incorporated into the local secondary structure dependent terms with magnitude $E_{\text{target,sec}}$. As indicated above, a given residue can be in one of five conformational states assigned on the basis of the local chain geometry. For those residues having a predicted secondary structural type, energetic biases for the various allowed conformational states are assigned. Turns are encoded on a generic basis, i.e. their chirality is not specified. Rather, they behave as flexible joints between regular secondary structural elements (see Skolnick *et al.* (1997b) for additional details).

U-turn surface bias. Regions predicted as U-turns are assumed to lie at the protein surface. Thus, for these residues, a penalty of 0.5 kT (with k Boltzmann's constant and T the absolute temperature) per residue is added when they lie at or below the radius of gyration. This term of total magnitude,

$E_{\text{U-turn}}$, acts to reduce kinetic traps by segregating the different parts of the protein into its corresponding layers. Similarly, N and C-terminal residues are penalized by 4 kT if they are buried (i.e. at or below the radius of gyration) to account for their charged ends.

Hydrogen bond mixing rules. The hydrogen bond potential is modified for those residues assigned to a predicted type of secondary structure so that the resulting hydrogen bond pattern is compatible with the secondary structural prediction. The magnitude of this term is $E_{\text{H-bond}}$. More specifically: (1) continuous stretches of strands and extended states or their combinations cannot form intra-element hydrogen bonds. Strands can form hydrogen bonds only with other strands, extended states or non-assigned states. On the other hand, extended states can form hydrogen bonds with all states except helices. (2) For those residues assigned to be helical, hydrogen bonds beyond the fifth neighbor along the chain are not allowed.

β-Strand cooperativity term. In trial calculations, it was observed that predicted β-strands had considerable difficulty forming β-sheets. The same observation has been made by other authors (Dandekar & Argos, 1996; Friesner & Gunn, 1996; Simons *et al.*, 1997). In our case, this behavior appears to result from a combination of the excessive conformational entropy of the backbone and the highly permissive hydrogen bond scheme. To correct for these effects, a cooperativity term that stabilizes and propagates the formation of β-sheets ($E_{\beta\text{-prop}}$) has been included. For each predicted strand, the hydrogen bond state of each residue in the putative strand is scanned. If the residue of interest participates in two hydrogen bonds belonging to two different β-strands, then a stabilization energy equal to that of the hydrogen bond cooperativity term is added. Strand residues can both nucleate and participate in the cooperativity. In other words, blocks of secondary structure predicted to be strands can be located either in the core or at the edges of the β-sheet. Extended state residues can serve as cooperative hydrogen bond partners, but cannot nucleate cooperativity; therefore, their location in the β-sheet core is energetically penalized, but not forbidden. There is no directionality in this cooperativity term. Thus, it cannot distinguish parallel from antiparallel arrangements of the strands, rather the final arrangement is dictated by the connectivity of the chain and the predicted restraints. This hydrogen bond cooperativity term has the effect of propagating the β-sheet. It also helps to bury strands predicted by PHD into the core, and to locate extended states predicted by LINKER at the surface and at the edges of a β-sheet. Trial calculations indicated that a small bias is adequate to successfully build β-sheets. In the all -β and α/β proteins studied here, the total magnitude of this term has a value of roughly $-5\ kT$ in successfully folded structures.

Tertiary restraints

Restraint function

The restraint function used in this work consists of a simple flat-bottom harmonic potential. Let r_{ij} be the actual distance between two restrained residues. In practice, the restraint could operate between side-chain centers of mass or between the projection of the residue pair onto the principal axes of their respective secondary structural elements. This situation is discussed in greater detail in the next section, which describes restraint splinning. Thus, the restraint function is as follows:

$$\begin{aligned} &= 800 && \text{if } E_{\text{res}} > 800 \\ E_{\text{res}} &= k_{\text{rep}}(r_{ij} - r_{ij}^0)^2 && \text{if } r_{ij} > r_{ij}^0 \text{ and } E_{\text{res}} < 800 \\ &= 0 && \text{if } r_{ij} < r_{ij}^0 \end{aligned} \quad (2)$$

$k_{\text{rep}} = 4.0\ kT/(lu)^2$ with lu equal to one lattice unit (1.22 Å). In the case of side-chain centers of mass, $r\backslash rm\ ^0_{ij} = (\langle r_{\text{AB}}\rangle + \sigma_{\text{AB}})\ (1 + \omega)$. For a pair of residues, A and B, $\langle r_{\text{AB}}\rangle$ and σ_{AB} are the average separation distance and standard deviation of this distance observed in a structural database. The value of $\omega = 0.5$ is used by default. In the case of restraint splinning (see below), the value $(\langle r_{\text{AB}}\rangle + \sigma_{\text{AB}})$ is substituted by the average separation distance observed in a structural data base for the packing of secondary structure elements. The values used are: 10 Å for α-α, 8 Å for α-β and 6 Å for β-β super-secondary elements, respectively.

Restraint splinning. Most predicted seeds are shifted by at least one residue with respect to the experimentally observed contact. Moreover, after growth, the different patches of contacts can have different phases. For example, suppose that one helix is predicted to contact two other secondary structure elements, i.e. it has two seed contacts. Because each seed is obtained and grown independently, the overall predicted contact pattern of the helix with the two other elements could be impossible. These seed contacts can lie on opposite sides of the helix, but in fact in the folded structure their secondary structure partners can be on the same side of the helix face. This effect could either preclude successful assembly or distort the folded conformation such that distinction, using energy criteria from misfolded alternatives, is not possible. One way to eliminate these artifacts is to apply the restraints between the axes of the secondary structure elements. This is done by smoothing the local C^{α} chain using the method described in the Appendix of Skolnick *et al.* (1997b). This list of smoothed coordinates is continuously updated during the simulation.

Knowledge-based restraints. Knowledge-based information about the general features of protein topology is also used (Skolnick *et al.*, 1997b). This knowledge-based information acts to reduce the number of misfolded structures. Two types of knowledge-based rules are considered, namely the chirality of βαβ units and the angle formed in ββα supersecondary structure units (Chothia & Finkelstein, 1990). The implementation used here differs in some important aspects from that described by Skolnick *et al.* (1997b). First of all, because the secondary structure prediction scheme can miss an intervening element, the number of successive residues between secondary structure regions is counted. If the number of loop residues is greater than 15 residues, it is assumed that an intervening secondary structure element has been missed by the secondary structure prediction algorithm, and the knowledge-based rules are not applied at all. The knowledge-based rules themselves are also implemented in a different way than that described in our previous work. First of all, in the ββα rule, the chirality requirement is eliminated and only the angle between the elements is restricted. When predicted secondary structure is used, this rule is not sufficiently robust because strands, particularly at the edges of the fold in α/β proteins, can be missed. This results in the inappropriate application of the chirality requirement of the rule. In the case of the βαβ rule, the vector definitions and restraint potential are the same as in our previous work (Skolnick *et al.*, 1997b). However, the definition of the angles between elements is different from that previously employed because the previous implementation made implicit assumptions about the chain geometry that can be violated. In particular, in the method described in our previous work, the first strand of the βαβ element was not demanded to be in the plane formed by the vector describing the orientation of the second strand, and the vector connecting the beginning of the second strand with the end of the first strand. Therefore, unusual geometries were not penalized. These geometries did not appear in our previous work because the sparse set of restraints used in the folding simulations was exact and always involved some contacts between β-sheet forming strands. However, the use of incorrect, clustered restraints in the present work permits the appearance of such conformations. The new vector definitions of the βαβ element are shown in Figure 2. The energy penalty for the βαβ rule is then given by:

$$E_{\beta\alpha\beta} = \sum (V_{\pi-1}(k) + V_{\pi-2}(k))\eta(k) \qquad (3)$$

with $\eta(k) = 1$ if element $k-2$ is predicted to be β, $k-1$ is α and k is β, and $\eta(k) = 0$ otherwise. The sum is taken over the number of secondary structure elements N_{sec}. The potentials $V_{\pi-1}$ and $V_{\pi-2}$ are given by:

$$V_{\pi-1}(k) = \eta(k) K_{\beta\alpha\beta}(\mathbf{b} \cdot \hat{\mathbf{e}})^2 \qquad (4)$$

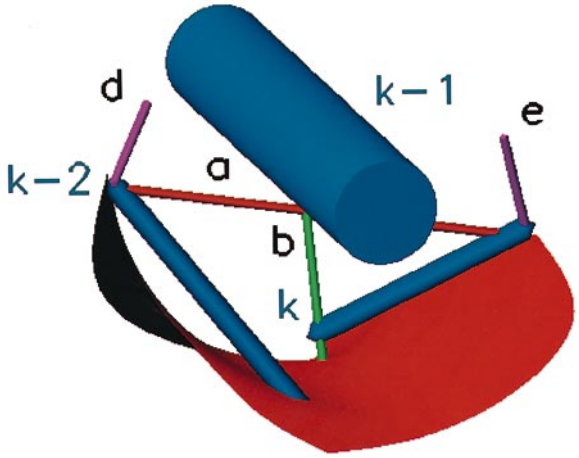


Figure 2. Scheme of the geometric definitions used to define the knowledge-based rules in βαβ supersecondary structures. The secondary structure elements are represented as blue cylinders; the strands are represented as two thin cylinders and the helix as a thick cylinder. The axes of the secondary structure elements are represented by the blue cylinders and are represented by the vectors $\mathbf{k}-2$, $\mathbf{k}-1$ and $\mathbf{k}$, respectively. The β-sheet to which the two strands belong is represented as a red membrane. The vector **a** connects the beginning of the last β-strand with the end of the first β-strand. The vector **b** connects the middle of the vector **a** with the middle of the $\mathbf{k}-1$ element; it is shown in green in the Figure. Note that it is shifted from its original position for display purposes. Vectors **d** and **e**, shown in magenta, are perpendicular to the plane defined by vectors $\mathbf{k}-2$ and **a**, and **k** and **e**, respectively. Figure generated with MOLMOL (Koradi *et al.*, 1996).

$$V_{\pi-2}(k) = \lambda(k) K_{\beta\alpha\beta}(0.7 - (\hat{\mathbf{e}} \cdot \hat{\mathbf{d}}))^2 \qquad (5)$$

where the vectors **b**, **d** and **e** are given in Figure 2. These vectors are obtained as follows: Let us denote $\mathbf{x}_\alpha^\beta$ as the splinned coordinates (Skolnick *et al.*, 1997b) of the starting ($\alpha = \text{s}$) and ending ($\alpha = \text{e}$) points of the secondary structure elements $\beta = k-2$, $k-1$ or k, repectively (see Figure 2). The unit vectors describing the direction of the secondary structures of the β elements can be found as: $v^\beta = ||\mathbf{x}_\mathbf{e}^\beta - \mathbf{x}_\mathbf{s}^\beta||$. We can also define the vector connecting the two strands as $\mathbf{a} = (\mathbf{x}_\mathbf{e}^{k-2} - \mathbf{x}_\mathbf{s}^k)$. The vectors **d** and **e** can then be obtained as the following cross-products: $\mathbf{d} = (\mathbf{v}^{k-2} \times \mathbf{a})$ and $\mathbf{e} = (\mathbf{v}^k \times \mathbf{a})$. For the derivation of **b**, we refer to the derivation of equation (A8a) in our previous work (Skolnick *et al.*, 1997b). The value of $\eta(k) = 1$ if $(\mathbf{b} \cdot \mathbf{e}) < 0$, as the connection is then left-handed, and it is $\eta(k) = 0$ otherwise. Also, if $||\mathbf{e} \cdot \mathbf{d}|| > 0.7$, then $\lambda(k) = 0$, otherwise $\lambda(k) = 1$. This allows an angle of up to 45° between the strands, therefore taking into account the possible β-sheet twist (see Figure 2). The typical value of $K_{\beta\alpha\beta} = 20\ kT$.

Relative weighting of the various contributions. The total energy of a given conformation is given

by:

$$E = 0.5E_{14} + 1.5E_1 + E_{\text{density}} + 2.75E_{\text{pair}} + E_{\text{target,sec}} + E_{\text{U-turn}} + E_{\text{H-bond}} + E_{\beta\text{-prop}} + E_{\text{res}} + E_{\text{know}} \quad (6)$$

Conformational sampling

Sampling of conformational space occurs *via* a standard asymmetric Monte Carlo Metropolis scheme (Metropolis *et al.*, 1953). Several types of local conformational micromodifications of the chain backbone and rare, small distance motions of larger chain fragments, together with side group equilibration cycles, are used. From 10 to 40 independent assembly simulations for each protein are carried out, each from a fully extended initial conformation. Each simulation starts at a reduced temperature of 5.0, and then the temperature is slowly lowered to 1.0. Low energy structures are then subject to isothermal refinement. The predicted fold is the one exhibiting the lowest average energy during the isothermal calculation. (In correctly folded structures, this energy is roughly $5kT$ per residue.)

Computational details

The typical computational time of assembling a 100-residue protein with MONSSTER, consisting of a run of about 5×10^6 Monte Carlo steps, is 5.0 hours on a single SGI MIPS R10000 processor running at 180 MHz clock speed and using a cache size of 32 kb. Each isothermal calculation needs an additional 5.0 hour run.

Structural analysis

The folded structures were compared with the experimental conformations using two sets of measurements, describing the global similarity of the structures and the local matching of the secondary structure elements. For the global similarity, the total coordinate root-mean-square deviation (cRMSD) of the two structures was calculated after computing the best superposition of the predicted structure with the experimental structure, using the McLachlan algorithm. In all the cRMSD calculations, all residues of both structures were included.

The comparison of the secondary structure prediction of the predicted models with that of the experimental structure and that coming from secondary structure prediction algorithms is complicated, and we feel that no satisfactory method can be used at the moment. The standard method for assigning secondary structure is due to Kabsch & Sander (1983) and is implemented in the DSSP program. The method relies heavily on the hydrogen bond pattern of the structure as computed from the peptide plates. The C^α models of the predicted structures lack these peptide plates. An all-atom reconstruction of these models is possible, and has actually been carried out (see below), but the DSSP method is very sensitive to small shifts in the coordinates in the secondary structure assignment; thus, the direct use of the DSSP assignments in the predicted models is prone to introduce considerable artifacts. For this reason, we adopted the Richards & Kundrot (1988) definition of secondary structure, based on the pseudodihedral angles of consecutive C^α atoms. Still, this approach is not entirely satisfactory, as the PHD method has been developed using the DSSP definition of secondary structure. However, trial calculations indicated that artifacts introduced by this approach are smaller than when the DSSP assignment is used. Computation of the Richards-Kundrot secondary structure was carried out on the basis of reconstructed all-atom models. All-atom reconstruction of each of the predicted structures was carried out using the MODELLER program (Sali & Blundell, 1993) using default parameter settings.

Proteins tested

The above protocol has been applied to the set of 20 small proteins listed in Table 1 sorted by size and named according to the entry in the Brookhaven database (Bernstein *et al.*, 1977). The size of the proteins ranges from the 29 amino acids of 3cti to the 100 amino acids of 1ife. They were chosen to examine how well the methodology performs on the following different structural motifs: small disulfide-rich proteins; all-α proteins; all-β proteins and α/β proteins. The fold description of each of the proteins is presented in Table 1 according to the SCOP database (Murzin *et al.*, 1995). In all cases, the structures were chosen at random, with the only constraints being that no global misassignment of the secondary structure prediction took place, i.e. no long helical segment was assigned as extended or *vice versa*, and that a sufficient number of homologous sequences, at least ten, was available in the HSSP database (Sander & Schneider, 1991). For the two smallest disulfide-rich proteins, 3cti and 1ixa, which are substantially devoid of secondary structure, the same protocol was applied. However, here we assumed knowledge of the identity of the disulfide bridges, as the general protocol would presumably fail due to the small predicted content of secondary structure. Such disulfide bridges are used as seeds in the tertiary restraint derivation protocol for contact prediction. Here, the objective is to study whether the knowledge of disulfide bridges, together with the protocol we now present, can produce low resolution models of small disulfide-rich proteins, which, in general, are considerably less regular than larger proteins. For T0042, we also used the known disulfide pattern, as it was available to the prediction teams in the CASP2 contest. In the case of the other 17 proteins, the native disulfide bridges were not used as seeds in those proteins that had disulfide bridges in their native structure. Rather, cross-

links were chosen so as to be compatible with the predicted contact map. To this effect, the following algorithm was used: first, the set of all possible pairs of disulfide bridges is computed. From this set, the first disulfide bridge is assigned as that with the closest contact map distance to any of the predicted contacts. The selected pair of cysteine residues is removed from the list of free cysteine residues, and the list of possible disulfide bridges is recomputed. The algorithm continues until all pairs are assigned.

All additional information in this study, such as the derived set of predicted contacts for all proteins, and structural predictions, is available *via* the WWW on the URL http://www.scripps.edu/skolnick/ORTIZ/ortiz.html

Results

Derivation of restraints

Secondary structure restraints

The secondary structure bias for each residue used in these series of calculations has been obtained by mixing the PHD and LINKER predictions, as described in Methods. The accuracy of the PHD predictions is shown in Table 2, and the final states assigned to each residue in a representative subset of the proteins studied are shown in Figure 3. The average secondary structure prediction accuracy for the set of proteins used here is higher than the expected accuracy of the method, but it is still within one standard deviation of the current accuracy of PHD. Combination of the LINKER algorithm with the PHD method seems to improve, on average, the quality of the predictions. This improvement comes from two sources: (1) the ability of LINKER to break long PHD-predicted helices by inserting loops in unphysically long helices assessed on the basis of the expected protein size, and (2) the ability of LINKER to predict extended states of local hydrophilic stretches, usually missed by PHD as a result of the fitting of PHD to the DSSP assignment of secondary structure, which demands a hydrogen-bond network in the assignment (Kabsch & Sander, 1983). Thus, PHD sometimes misses stretches of sequence with poor β-sheet propensities, but which populate extended states. For example, the second strand in 1gpt was missed by PHD, but localized as an extended state by LINKER (observe in Figure 3 the difference between 1 and 4 states). A similar situation is observed in the first strand of 1tfi. For 1hmd, helices two and three are merged, but LINKER successfully corrects this overprediction (Figure 3). A similar situation was observed in 3icb (not shown). However, because of the limited accuracy in loop positioning of the algorithm at the residue level, the final outcome of the secondary structure assignment is usually more shifted with respect to the experimental secondary structure than in the original PHD prediction. Although the overall secondary structure assignment of the proteins tested here can be considered to be quite good by state-of-the-art standards, it is worth pointing out that there are still some proteins for which entire elements of secondary structure are missed. Thus, in the case of 1ftz, the third helix is missed, and an additional helix is predicted in the C terminus. Similarly, the third strand of 1shg is missed, as is the second and third helix and the fourth strand of 1ego (Figure 3), as well as the third strand of 1poh at the edge of the fold (not shown). There are also considerable discrepancies between the predicted and observed lengths and locations of the secondary structure elements (Figure 3).

Contact prediction

The results of side-chain contact prediction are compiled in Tables 2 and 3. Not considering cysteine-rich proteins where the disulfide contact pattern has been assumed to be known, the overall accuracy of the procedure of contact prediction is similar to that reported by other authors (Göebel *et al*., 1994; Olmea & Valencia, 1997), on the order of 25%. Turning to the issue of precision, allowing an error of one residue in the assignment of partners yields an accuracy on the order of 60%; within two residues is 77%; within three residues is 85%; and within four residues is 95%. The average prediction of the contact map coverage is on the order of 25%, which corresponds to about $N/4$ restraints per N protein residues and is consistent with our previous findings (Skolnick *et al*., 1997b). Thus, in general, the number of contacts predicted is a small portion of the whole protein contact map and usually contains a significant amount of noise. In some cases, wrong pairings of secondary structure elements are obtained, as is the case with 1poh, 1ife and 1hmd. In general, as expected, the bigger the protein, the higher the chance to assign a wrong pairing of secondary structure elements. The structures used for ''growing'' the contacts predicted by correlated mutations for each of the proteins studied are compiled in Table 3. The structures selected are entirely unrelated to the target sequence, i.e. no remote sequence homologues were used. With the contact derivation procedure used in this work, the final predicted contacts are highly clustered in structure; therefore, the effective number of contacts that can be considered to be independent is lower than the number of predicted contacts. Some other effects are also worth noting. Comparison of the predicted and observed contact maps (see Figure 4) shows that, as a result of the independent growth of each one of the restraints, frequent phase shifts for the different restraint subsets are observed. This problem is particularly important for restraints involving α-helices. This can produce helix unwrapping and other structural distortions. As explained in Methods, we have tried to avoid this problem by introducing the ''splinning'' procedure.

Table 1. Classification of the proteins folded in this study according to the SCOP (URL http://scop.mrc-lmb.cam.ac.uk/scop/) database

Protein	*N*res	Class	Fold description	Name
3cti	29	Small	Disulfide-bound fold, beta hairpin with adjacent disulfide	Trypsin inhibitor from squash (*Cucurbita maxima*)
1ixa	39	Small	EGF-like (disulfide-rich fold; nearly all-beta)	Factor IX from human (*Homo sapiens*)
protA	47	α	Three-helix bundle	Protein A
1gpt	47	small	Disulfide-bound fold beta hairpin with adjacent disulfide	Gamma-thionin from barley (*Hordeum vulgare*)
1tfi	50	Small	Rubredoxin-like (metal bound fold, with 2 CXXC motifs)	Transcriptional factor SII from human (*Homo sapiens*)
6pti	58	Small	BPTI-like (disulfide rich α + β fold)	Pancreatic trypsin inhibitor from bovine (*Bos taurus*)
1fas	61	Small	Snake toxin like (disulfide rich; nearly all beta)	Fasciculin from green mamba (*Dendroaspis angusticeps*)
1shg	62	β	SH3-like barrel (partly opened; $n^* = 4$, $S^* = 8$; meander)	alpha-Spectrin, SH3 domain from chicken (*Gallus gallus*)
1cis	66	α + β	CI-2 family (α + β sandwich; loop across free side of β)	Hybrid protein from barley (*Hordeum vulgare*) hiproly strain
1ftz	70	α	DNA-binding 3-helix bundle (right-handed twist; up-down)	Fushi Tarazu protein from fruit fly (*Drosophila melanogaster)*
1pou	71	α	DNA-binding domain (4 helices, folded leaf, closed)	Oct-1 POU-specific domain from human (*Homo sapiens*)
1c5a	73	α	Anaphylotoxins (4 helices; irreg. array, disulfide linked)	C5a anaphylotoxin from pig (*Sus scrofa domestica*)
3icb	75	α	EF-hand (2 EF-hand connected with Ca bind loop)	Calbindin D9K from bovine (*Bos taurus*)
1ubi	76	α + β	β-Grasp (single helix packs against β-sheet)	Ubiquitin from human (*Homo sapiens*)
T0042	78	α	Five-helix bundle	NK-lysin from pig (*Sus scrofa*)
1lea	84	α	DNA-binding 3-helix bundle (right-handed twist; up-down)	LexA repressor, DNA-binding domain (*Escherichia coli*)
1ego	85	α/β	Thioredoxine-like (3 α/β/α layers; β-sheet order 4312)	Glutaredoxin from bacteriophage t4
1hmd	85	α	Four helical up-and-down bundle (left-handed twist)	Hemerythrin from sipunculid worm (*Themiste dyscrita*)
1poh	85	α + β	α + β sandwich	Histidine-containing phosphocarrier proteins (*E. coli*)
1ife	100	α + β	IF3-like (β-α-β-α-β(2); 2 layers; mixed sheet 1243)	Translation initiation factor IF3 from *Escherichia coli*

The protein name is assigned according to the Brookhaven Protein Data Base entry, except for Protein A (not in the PDB dababase) and T0042, corresponding to the sequence target 42 of the recent CASP2 meeting (URL http://PredictionCenter.llnl.gov/); see the text for details.

Table 2. Statistics of the restraint derivation procedure

Protein	*N*seq[a]	*N*seed[b]	*N*pc[c]	δ = 0[d]	δ = 1[d]	δ = 2[d]	δ = 3[d]	δ = 4[d]	δ = 5[d]	*N*w[e]	*N*con[f]	%PC[g]	Q3[h]
3cti[i]	19	3	6	83.3	100.0	100.0	100.0	100.0	100.0	0	39	15.3	82.4
1ixa[j]	70	2	5	100.0	100.0	100.0	100.0	100.0	100.0	0	48	10.4	97.4
prota	25	3	17	0.0	35.2	70.5	82.3	94.1	100.0	0	91	18.6	83.0
1gpt	8	3	13	46.1	76.9	100.0	100.0	100.0	100.0	0	70	18.5	72.3
1tfi	10	8	37	21.6	54.0	88.8	93.3	97.7	100.0	0	84	44.0	78.0
6pti	45	6	19	68.4	94.7	100.0	100.0	100.0	100.0	0	92	20.6	80.4
1fas	44	1	19	26.3	57.8	78.9	89.4	100.0	100.0	0	98	19.3	90.2
1shg	20	6	39	28.2	89.7	100.0	100.0	100.0	100.0	0	109	35.7	64.9
1cis	17	5	23	8.6	65.2	78.2	95.6	100.0	100.0	0	144	15.9	86.4
1ftz	312	2	12	25.0	33.3	58.3	58.3	75.0	91.6	1	149	8.0	71.4
1pou	47	5	48	28.6	77.5	89.8	95.9	100.0	100.0	0	122	39.3	84.5
1c5a	20	9	45	24.4	62.2	73.3	82.2	95.5	95.5	2	105	42.8	93.8
3icb	67	3	25	28.0	68.0	68.0	76.0	100.0	100.0	0	154	16.2	89.3
1ubi	33	7	17	23.5	88.2	94.1	94.1	100.0	100.0	0	153	11.1	77.6
T0042	18	3	24	29.1	45.8	58.3	70.8	91.6	95.8	1	150	16.0	80.8
1lea	16	10	41	9.7	34.1	75.6	90.2	95.1	95.1	2	131	31.2	87.5
1ego	10	7	33	15.1	84.8	93.9	96.9	100.0	100.0	0	223	14.7	71.8
1hmd	11	5	20	10.0	45.0	65.0	70.0	90.0	90.0	2	157	12.7	85.0
1poh	19	12	36	8.3	33.3	55.5	80.5	91.6	91.6	3	162	29.6	74.1
1ife	13	6	21	14.2	23.8	38.0	61.9	80.9	85.7	3	148	14.1	70.0

[a] Number of sequences contained in the multiple sequence alignment.
[b] Number of predicted contacts obtained from the multiple sequence alignment (see the text for details).
[c] Number of predicted contacts used as restraints in the simulations.
[d] Percentage of predicted contacts within δ residues of a native contact.
[e] Number of contacts that are incorrect when δ = 5.
[f] Number of contacts found in the experimental structure. A contact between two amino acids occurs when any of their side-chain heavy atoms are within 5.0 Å from each other.
[g] Percentage of the contact map of the experimental structure predicted by the contact prediction method.
[h] Per residue percentage accuracy of secondary structure prediction, based on a three-state model, obtained from the PHD method.
[i] In the case of 1ixa, the restraints come from the known disulfide bridges (contacts: 3-20; 10-22; 16-28), plus three additional contacts predicted by the correlated mutations method (contacts: 8-17; 7-27; 14-21).
[j] Restraints come from the known disulfide pattern (contacts: 6-17; 11-26; 28-37) and predicted contacts from the correlated mutation analysis (contacts: 10-24; 13-32).

```
Protein: 1shg      #residues :  57
....,....1....,....2....,....3....,....4....,....5....,..
KELVLALYDYQEKSPREVTMKKGDILTLLNSTNKDWWKVEVNDRQGFVPAAYVKKLD
 EEEE         EEEEEE    EEEEEE  EEEEEEHHHEEEE
554444445555333354445533444445533335544555535444555544555

Protein: 1cis      #residues :  66
....,....1....,....2....,....3....,....4....,....5....,....6....,.
MKTEWPELVGKSVEEAKKVILQDKPEAQIIVLEKQAVDNAYAEYRIDRVRLAVDKLDNIAQVPRVG
             HHHHHHHHHHH     EEEEE                 EEEEE
554444445555222222222233344444444455555555333344444443344444444445

Protein: 1c5a      #residues :  65
....,....1....,....2....,....3....,....4....,....5....,....6....,.
MLQKKIEEEAAKYKYAMLKKCCYDGAYRNDDETCEERAARIKIGPKCVKAFKDCCYIANQVRAEQ
 HHHHHHHHHH     HHHHHHHHHHH         HHHHHHH     HHHHHHHHHHHHHHHHHH
222222222222333552222222233333335522222225533222222222222222222222

Protein: 1gpt      #residues :  47
....,....1....,....2....,....3....,....4....,..
RICRRRSAGFKGPCVSNKNCAQVCMQEGWGGGNCDGPLRRCKCMRRC
 EEEEE           HHHHHHHHHHH    EEEEE     EEEEEEE
444444335555113332222222225511111115333114444444

Protein: 1ftz      #residues :  56
....,....1....,....2....,....3....,....4....,....5....,.
TRQTYTRYQTLELEKEFHFNRYITRRRRIDIANALSLSERQIKIWFQNRRMKSKKD
      HHHHHHHHHHH         HHHHHHHHHH     HHHHHHHHHHH
111333322222222222223333322222222255553355555533322222225

Protein: 1fas      #residues :  61
....,....1....,....2....,....3....,....4....,....5....,....6.
TMCYSHTTTSRAILTNCGENSCYRKSRRHPPKMVLGRGCGCPPGDDYLEVKCCTSPDKCNY
 EEE          EEE      EEEEEE       EEEEEE         EEEEEE
544445555555544553355444445533354444453311111334444445555555555

Protein: 1lea      #residues :  72
....,....1....,....2....,....3....,....4....,....5....,....6....,....7..
MKALTARQQEVFDLIRDHISQTGMPPTRAEIAQRLGFRSPNAAEEHLKALARKGVIEIVSGASRGIRLLQEE
   HHHHHHHHHHHHHHHHHH         HHHHHHHH      HHHHHHHHHHHHHH EEE         EEE
555222222222222222233333332222222225555333222222222555334444555555444555555

Protein: 1tfi      #residues :  50
....,....1....,....2....,....3....,....4....,....5
KTGGTQTDLFTCGKCKKKNCTYTQVQTRSADEPMTTFVVCNECGNRWKFC
    EE               EEEEEE           EEEEEE      EEE
111115544441133333444444441133355114444433311444444

Protein: 1ego      #residues :  85
....,....1....,....2....,....3....,....4....,....5....,....6....,....7....,....8....,
MQTVIFGRSGCPYCVRAKDLAEKLSNERDDFQYQYVDIRAEGITKEDLQQKAGKPVETVPQIFVDQQHIGGYTDFAAWVKENLDA
 EEEEE       HHHHHHHHHHHHHHHHH    EEEEE HHHH     HHHHHH           EEEE  EEEE  HHHHHHHHHHHH
444444455333332222222222553333444444335533553335555555511144444443333553322222222255

Protein: 1hmd      #residues :  84
....,....1....,....2....,....3....,....4....,....5....,....6....,....7....,....8....
TIIDDEHKTLFNGILLLSQADNADHLNELRRCTGKHFLNEQQLMQSSQYAGYAEHKKAHDDFIHKLDTWDGDVTYAKNWLVNHI
HHHHHHHHHHHHHHHHHHHH    HHHHHHHHHHHHHHHHHHHHHHHHH       HHHHHHHHHHHHHHHH       HHHHHHHHHHHH
222222222222222225553333222222222222222222222333222222222222222223333332222222222222
```

Figure 3. Secondary structure assignment for a representative subset of the proteins used in this study. The amino acid sequence is given for each protein. The observed secondary structure in the experimental conformation according to the DSSP assignment (Kabsch & Sander, 1983) of three states is also shown, as is the assigned secondary structure state in the folding simulations, according to the prediction results. 1 stands for coil assignment; 2 for helix assignment; 3 for a U-turn assignment; 4 for strand assignment, and 5 corresponds to no assignment of secondary structure.

Table 3. For each of the predicted proteins, the source structures used for contact map *growth* are shown

Protein	Source structures used for contact growth[a]					
3cti	————————————					
1ixa	————————————					
protA	1bbhA	4ts1A	256bA			
1gpt	1lhm_	2gbp_	1ald_			
1tfi	1pk4_	1fdlH	1hoe_	1vaaB	1sarA	1pk4_
6pti	2fb4_	1er8_	2gb1_			
1fas	1fxa_	1dtx_	1atx_			
1shg	4tms_	1atx_	1gp1_	6taa_	3ebx_	
1cis	1pcy_	2hlaA	1ppd			
1ftz	1s01_	1c5a_				
1pou	1prc_	1cdp_				
1c5a	1pbxA	3adk_	1avr_	8catA	3wrp_	1col_A
1ubi	1ovaA	3enl_	1paz_			
T0042	1mba_	2utgA				
1lea	2liv_	1akeA	2lhb_	2timA	2liv_	
1ego	2azaA	2fx2_	5rubA	1rnh_		
1hmd	6taa_	1lig_				
1poh	5rubA	3blm_	1abp_	1rbp_		
1ife	2fcr_	1lh1_	1ald_			

[a] In the case of 3cti and 1ixa, the contact map growth step was not used, as a result of the insufficient secondary structure content in these structures. Here, restraints are given by the predicted contacts plus the known disulfide bridges. See also Table 2.

Fold assembly and discrimination

Fold assembly is carried out starting from an extended chain; therefore, the initial restraint energy is very large in the first cycles of the algorithm where the first motions of the chain mainly decrease the restraint energy. Thus, compact states are generated very quickly. Finally, the secondary structure forms, and the adjustments of secondary structure elements take place. In some cases, during the first annealing run, the structures are trapped in misfolded states. Then, during subsequent annealing runs, the correct registration of elements takes place. This effect is particularly observed in α/β proteins. In the final folds, typically the restraint energy is close to zero as a result of the soft implementation of restraints (Table 4), although on average the number of satisfied predicted contacts is about half the number of predicted contacts. Based on restraint satisfaction, it is not possible to discriminate among alternative answers (Table 4). Because the energy landscape is rugged, individual structures obtained from the assembly runs are not able to provide a reliable energy for the particular fold they represent. As a result, to rank order the folds, it is necessary to carry out isothermal calculations for at least the lower part of the energy spectrum of the created folds (Table 4).

The superimposed predicted and experimental conformations of a representative subset of the proteins tested in this work can be seen in Figure 5. Details of the resolution achieved for each particular protein can be found in Table 4. The average cRMSD is about 5 Å. Thus, in comparison with the use of ''exact'' restraints, a price in resolution between 1.0 to 2.0 Å has to be paid. When the different protein classes are considered, the average cRMSD of the lowest energy set of structures ranges from about 4 Å for helical proteins to roughly 6 Å for β and mixed motif proteins. In all cases, the global topology is recovered either as the best energy (in 17 out of 20 cases) or as the next best energy alternative fold. Of the three that failed (1ixa, 1hmd and 1ife), the misfolded state of 1ixa results from the misplacement of a few residues in the C-terminal region; in the case of 1hmd, it is not possible to distinguish between the two topological mirror images, which are essentially isoenergetic. In the last case, 1ife, the selected fold is actually correct in spite of the unacceptably high cRMSD. Here, a coil region shifts from the edge of the fold to the back of the protein. These numbers could be compared with the expected cRMSD obtained by random for protein chains of the length of the sequences used in these folding studies (Table 4), using the expression given by Cohen & Sternberg (1980). The average value that could have been obtained by random is around 12 Å.

Turning again to the assembly process, it is interesting to note that the different secondary structure elements do not simply pack as rigid bodies; that is, as shown in Figure 6, changes in secondary structure status produced by long-range interactions are common in order to assemble the fold. This can be quantified by the comparison of the secondary structure predictions and the secondary structure assignment of the predicted models (Table 5). In some cases, the secondary structure elements extend in length from the original predictions, as in the case of the α-helices of 6pti or the β-strands in 1gpt. In some other cases, they need to shorten to accommodate themselves into the protein fold, as in the case of 1lea. And in a number of cases, additional secondary structure elements form. The most striking case is that of 1ubi, as an example of α-helix formation, and 1shg for β-strands. Overall, our results suggest that fixing the length of secondary structure elements and treating them as rigid bodies can have a deleterious effect in fold assembly. For example, as a result of ''fraying ends'', some hydrophobic residues can be exposed and some wrong contacts between elements can be formed, reducing or even eliminating the energy gap with alternative folds. Some flexibility is required in order to correctly pack the secondary structure elements. Usually their predicted length is incorrect, and there are shifts in registration with respect to the tertiary restraints. Therefore, if a rigid model is used to define the secondary structure, the correct fold can be missed. However, on average, the correctness of the secondary structure in the predicted models does not improve when compared with the original secondary structure predictions used as restraints during the simulations (see Table 5 and Figure 6).

The case of T0042 as a blind prediction

The second meeting on the Critical Assessment of Techniques for Protein Structure Prediction (CASP2) was held in Asilomar, California, recently. Several protein targets were available to researchers as blind predictions covering different aspects of protein structure prediction: docking, homology modeling, threading and *ab initio* folding (URL http://iris4.carb.nist.gov/casp2/). At the time of the meeting, the present work was being carried out, and we felt that the method was too immature for us to participate. However, once we obtained enough experience with the approach presented here, and in order to compare our results with those of other groups using different methods under similar conditions, we pursued the blind prediction of target 42 (T0042) of the meeting. It must be stressed that the prediction was made without knowledge of the target conformation, as this structure has been released only recently. (All of the numerical assessments for this and the rest of the targets, as well as the experimental structure of some of the targets, are available through the World Wide Web URL http://PredictionCenter.llnl.gov/.) T0042 was chosen because it was the most popular target sequence for most groups doing *ab*

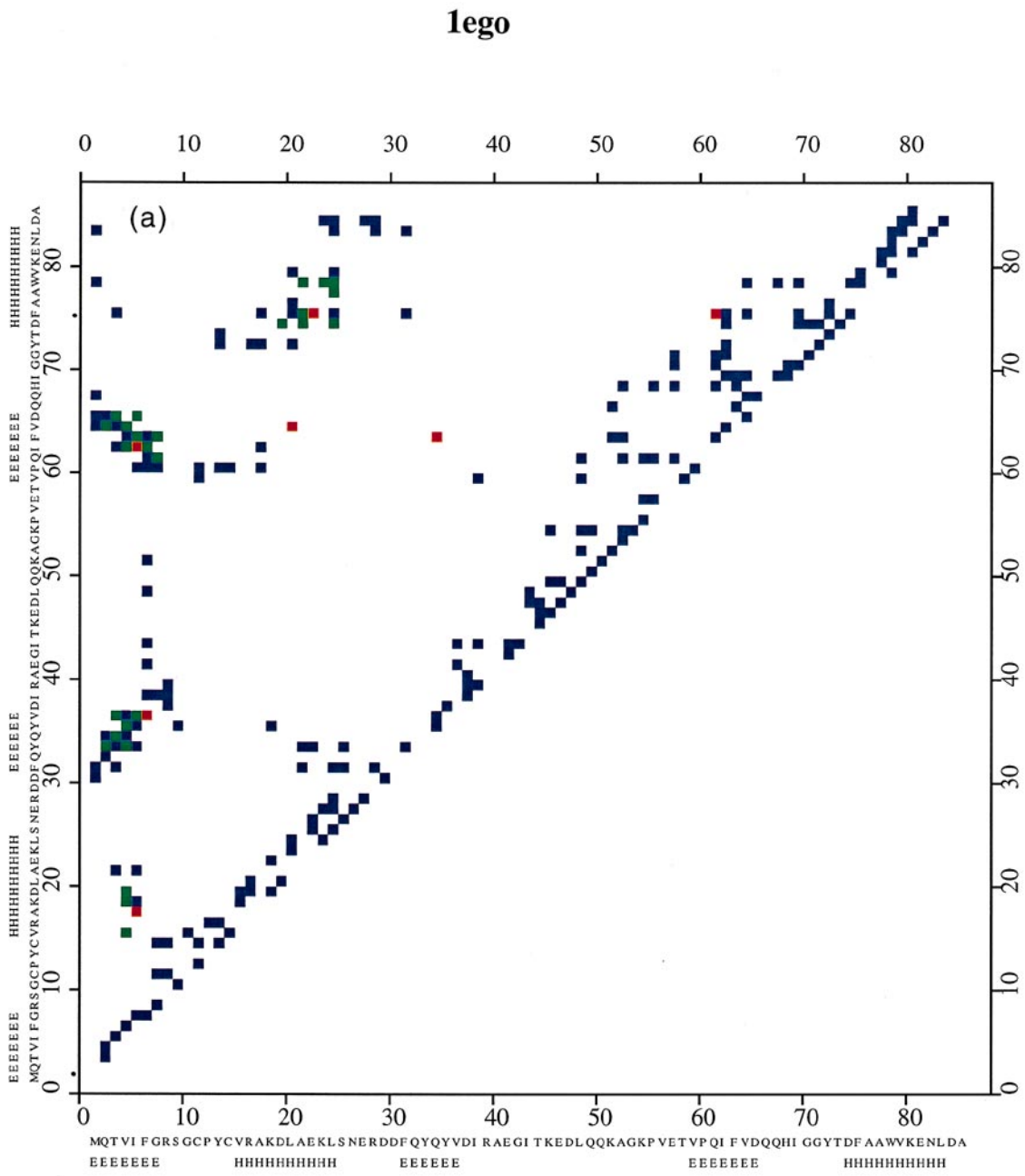


Figure 4(a) (*legend on page* 434)

initio folding. Here, we will describe in detail the results of the prediction of this protein. It will be used as an example to illustrate the prediction process following the present approach.

T0042 is a protein of 78 amino acids. The correct pairing of the three disulfide bridges of the protein was made available to the prediction teams, and it was used by us as well. A multiple sequence alignment was obtained for this sequence scanning the EMBL/SWISSPROT database with FASTA (Pearson & Lipman, 1988) and filtering the sequences found using MAXHOM (Sander & Schneider, 1991). However, it was necessary to manually edit this alignment in order to remove short sequences because the initial alignment did not provide any predicted contacts using our method. After filtering the alignment by hand, the final multiple sequence alignment contained 15 homologous sequences plus the target sequence (Table 6). Secondary structure predictions were carried out combining PHD and LINKER, as described in Methods (Table 7). The experimental structure contains five α-helices, but the PHD prediction merges helices III and IV and partially misses helix V. When the PHD predictions are combined with the LINKER predictions, the resulting secondary struc-

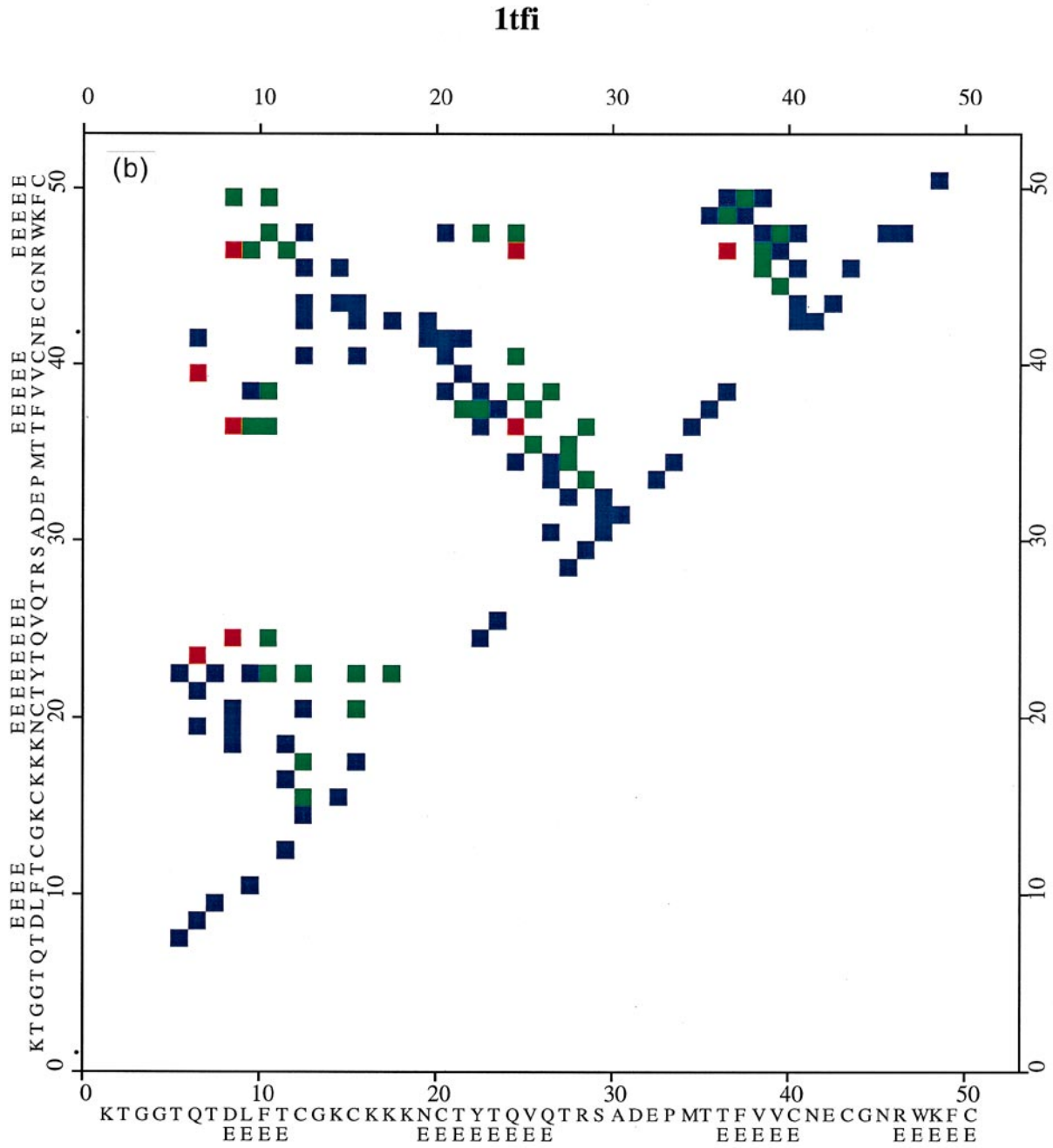


Figure 4(b) (*legend on page* 434)

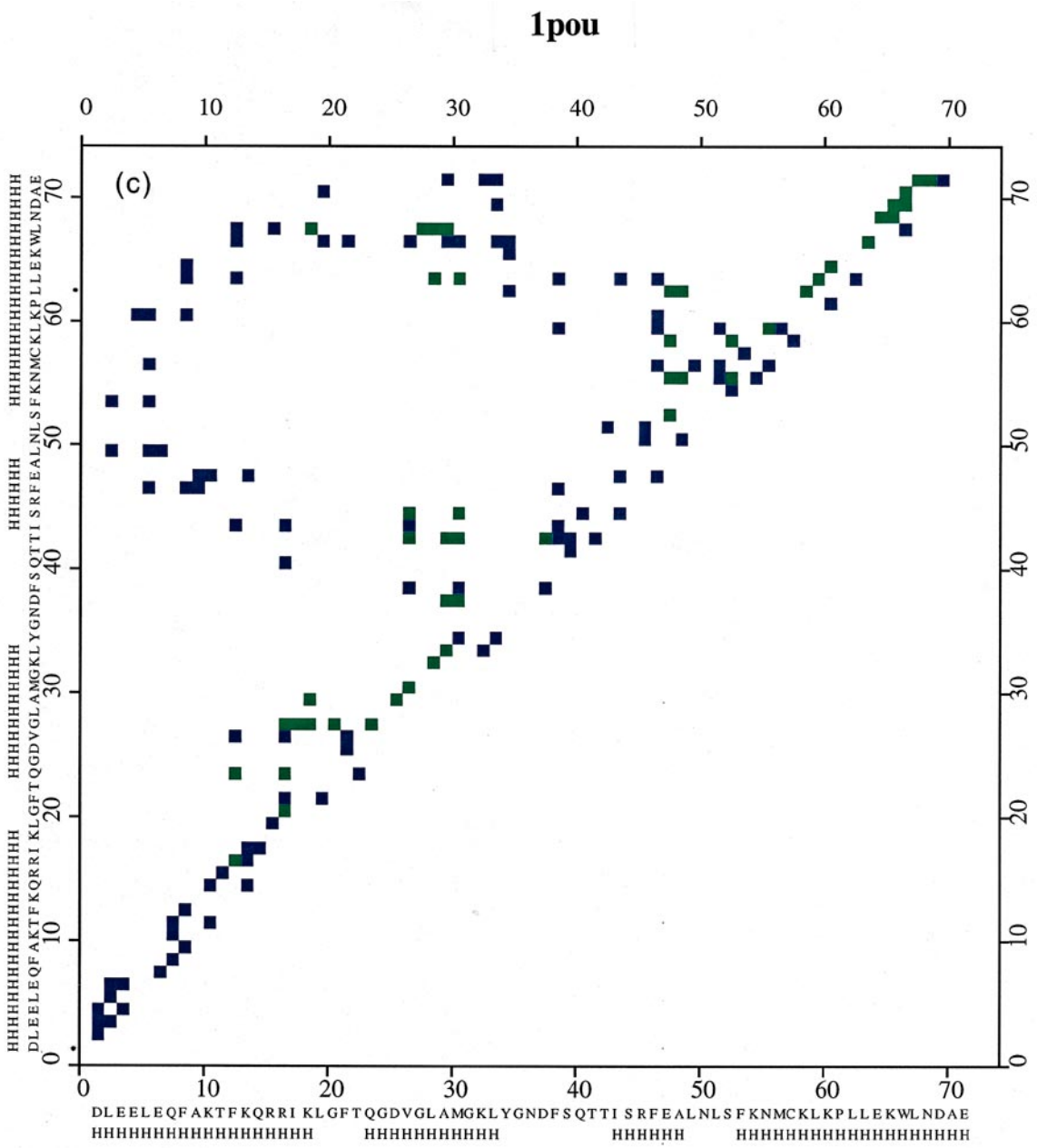


Figure 4. Contact maps of a representative set of proteins used in this work. The experimental contact map is shown in blue, the predicted seeds are shown in red, and the expanded contacts obtained by inverse folding are shown in green (see the text for details). A contact distance cut-off of 4.5 Å between side-chain heavy atoms is used. Contact maps for the following proteins are shown: (a) 1ego; (b) 1tfi; (c) 1pou.

ture assignment is actually worse than the PHD prediction alone: helices III and IV remain merged, helix V is totally missed and helix II is considerably shortened (Table 7). Using the correlated mutation analysis, three contact seeds could be predicted from the multiple sequence alignment (see Table 8). One of them involves a disulfide bridge observed in this protein, which made us feel more confident about the quality of the predicted seeds. These seeds, together with the known disulfide bridges, were used in the inverse folding calculation with the objective of ''expanding'' the predicted contacts. Only two of the three seeds could ''grow''. Thus, the final number of restraints was 23 (Table 9). Interestingly, one of the fragments selected for the enrichment process involved 2utg, which has been found by other researchers during the CASP2 contest to be a popular template when global threading of the sequence is done. The obtained secondary and tertiary restraints were used as input for MONSSTER. Ten simulations were performed. After the isothermal calculations, the lowest average energy fold was selected as the predicted structure.

After these calculations were completed, the experimental structure of T0042 was then available

Table 4. Results of the folding simulations

Protein	RMSD[a]	⟨*E*⟩[b]	σ[c]	*rs*[d]	*E*pen[e]	RMSD[f]	⟨*E*⟩[g]	σ[h]	*rs*[i]	*E*pen[j]	RMSD(*r*)[k]
3cti	**3.8**	−106.9	7.4	6	0.0	6.7	−103.1	7.8	6	0.0	10.60
1ixa	7.7	−131.2	7.0	2	13.2	**5.6**	−130.2	8.0	5	10.0	11.07
prota	**3.1**	−246.2	6.6	2	0.2	9.4	−240.0	5.5	1	0.3	11.45
1gpt	**5.9**	−276.1	12.4	9	4.3	6.6	−142.3	7.0	10	2.8	11.45
1tfi	**5.9**	−201.6	7.3	28	8.2	7.0	−191.2	9.4	31	2.2	11.59
6pti	**4.7**	−410.0	10.0	19	0.0	9.7	−397.0	10.0	18	0.00	11.96
1fas	**6.2**	−330.0	6.3	19	1.5	9.3	−284.0	7.6	20	10.7	12.10
1shg	**4.5**	−420.0	4.5	11	14.2	6.7	−397.0	5.5	17	21.1	12.15
1cis	**6.4**	−240.0	8.2	7	2.7	7.6	−232.0	6.6	7	0.1	12.34
1ftz	**5.1**	−276.9	8.0	11	0.7	10.13	−270.0	7.9	15	0.5	12.53
1pou	**3.5**	−418.0	3.4	18	31.8	11.9	−364.0	4.0	22	23.5	12.57
1c5a	**4.2**	−194.0	4.0	20	9.4	9.8	−182.0	5.2	26	5.3	12.66
3icb	**4.5**	−406.0	7.0	21	17.6	12.6	−342.0	3.9	11	15.0	12.76
1ubi	**6.1**	−238.0	6.3	9	0.0	11.5	−203.0	5.7	8	2.8	12.80
T0042	**5.6**	−362.2	8.2	15	10.4	11.7	−359.8	9.6	8	11.9	12.90
1lea	**6.1**	−136.0	7.7	26	8.8	9.4	−115.0	7.5	27	7.3	13.18
1ego	**5.7**	417.2	8.9	20	1.3	9.0	−396.4	15.0	14	1.19	13.22
1hmd	9.3	−459.7	5.4	13	0.3	**4.6**	−458.0	7.2	3	0.15	13.22
1poh	**6.5**	−336.0	9.5	42	24.6	11.7	−299.0	8.6	23	16.1	13.22
1ife	8.2	−481.8	7.3	16	5.8	**6.7**	−419.0	10.8	15	11.8	13.93

After the column corresponding to the protein name, the next five columns correspond to parameters describing the lowest energy fold obtained during the simulations. The following five columns correspond to these same parameters describing the alternative fold of lowest energy found during the simulations. The numbers in bold correspond to the lowest cRMSD among the competing folds. Note: Both folds of 1ife correspond to the same topology; however, the selected conformation has a strongly distorted strand at the edge of the fold. The final column describes the expected cRMSD obtained by random for a protein chain of the length of the corresponding sequence, according to the Cohen & Sternberg (1980) model.

[a] Average coordinate RMSD of the lowest energy fold found in the folding simulations.
[b] Average energy (in *kT* units) of the fold obtained from the isothermal calculation ($T = 1.0$).
[c] Standard deviation of the energy during the isothermal calculation ($T = 1.0$) for the lowest energy fold.
[d] Number of predicted contacts satisfied in the final predicted fold.
[e] Residual restraint energy (in *kT* units) in the predicted fold.
[f] Average coordinate RMSD of the lowest energy alternative fold found in the folding simulations.
[g] Average energy (in *kT* units) of the lowest energy alternative fold obtained from the isothermal calculation ($T = 1.0$).
[h] Standard deviation of the energy during the isothermal calculation ($T = 1.0$) for the lowest energy alternative fold.
[i] Number of predicted contacts satisfied in the alternative fold.
[j] Residual restraint energy (in *kT* units) in the alternative fold.
[k] Expected coordinate RMSD for a random chain of the length of the sequence under consideration according to the Cohen & Sternberg (1980) model.

to us. The RMSD between all C^{α} atoms of the experimental and computed structures is 5.6 Å (Table 4). A superimposition of the predicted and the experimental structure can be seen in Figure 5. Two striking features of the predicted fold are worth noting. First, helix III of the predicted secondary structure needs to break around residues 55 and 56 to assemble the fold, forming two independent helices in the predicted fold, as observed in the experimental structure. Thus, helix IV in the predicted structure extends from residues 59 to 62, as compared to residues 57 to 61 observed in the experimental conformation. The second point to note is the partial formation of the last C-terminal helix, helix V in the experimental structure, missed by the secondary structure predictions, even though a soft bias was used towards extended states. Both observations highlight the fact that proteins are frustrated systems from the energetic point of view, and that any prediction scheme must consider this frustration. The local secondary structure biases provided by the secondary restraints were in both cases overridden by tertiary interactions. It must be emphasized that many predictions submitted to CASP2 failed to provide the correct answer because the secondary structure was assumed to be completely correct. In our case, helix IV of the real structure is shorter and slightly shifted when compared to the experimental structure.

On the other hand, the prediction of T0042 also illustrates some of the shortcomings of the method. The last helix observed in the experimental structure was predicted as an extended state, and an extended state partially persists in the C-terminal region of the final structure, being one of the main errors in the predicted conformation. Another problem is related to the energy discrimination of the fold. As seen in Figure 7, the energy difference between the lowest average energy structure and that of an alternative fold is about 4 *kT*. Figure 7 also demonstrates that most of the noise in the energy evaluation comes from the sequence-independent terms. Thus, when only the pair potential energy is considered, the energy differences increase to about 10 to 20 *kT*. It is of interest to note that the restraint energy, by itself, favors an alternative topology by 10 *kT*. This again illustrates the need to incorporate the restraint function as a soft bias to provide a manifold of topologies,

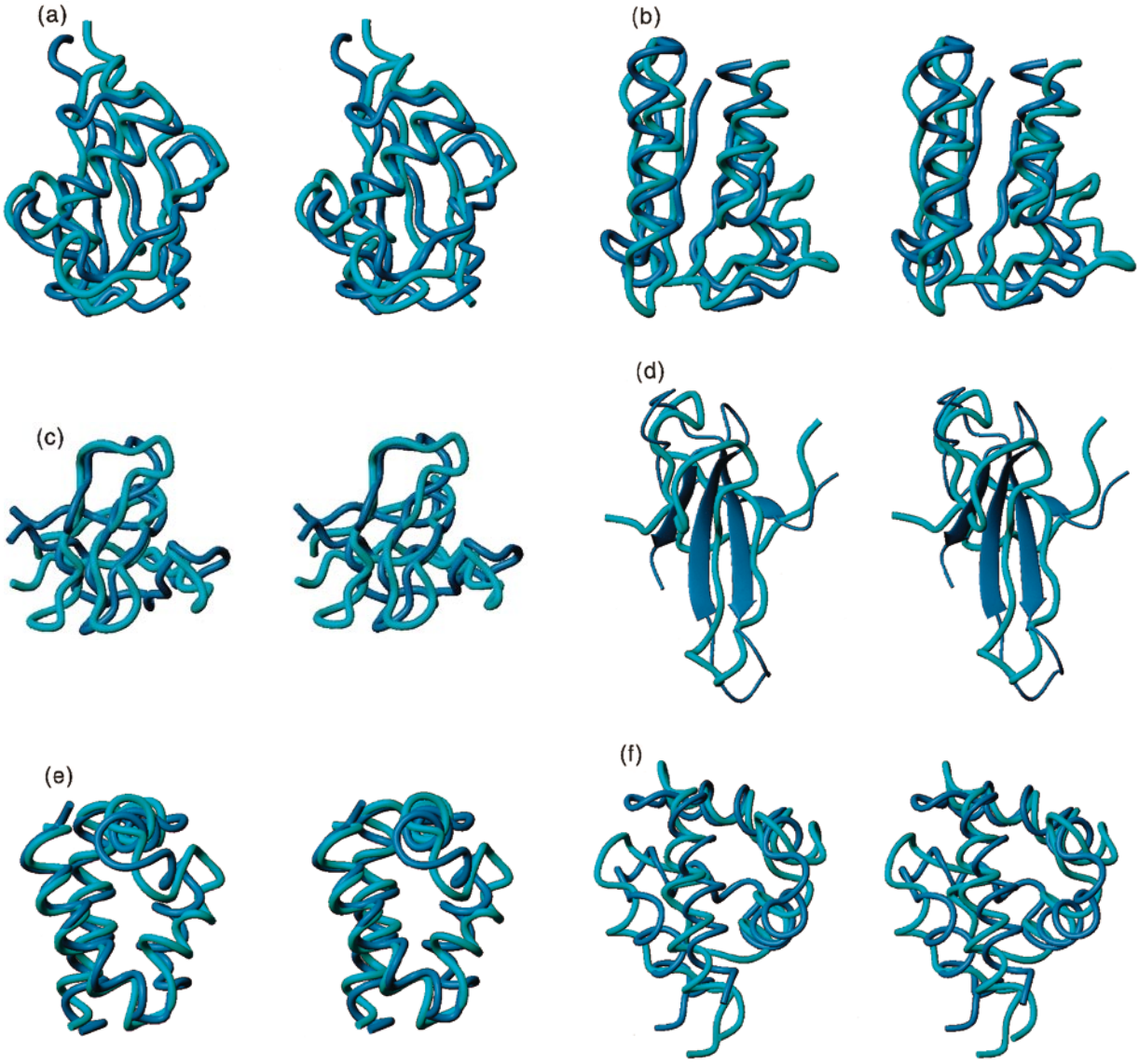


Figure 5. Predicted and experimental structures of: (a) 1ubi; (b) 1ego; (c) 1shg; (d) 1tfi; (e) 1pou; (f) T0042. The experimental structures are shown in blue, while the predicted structures are shown in cyan. Figure generated with MOLMOL (Koradi *et al.*, 1996).

among which the energy function must select the correct one. Thus, the restraint energy bias cannot be too large.

Our results compare favorably with those obtained by other groups in the CASP2 contest. In *ab initio* folding, the best result was obtained by Jones's group, who were able to obtain predictions of 6.2 Å RMSD with respect to the target protein, although their predicted four-helix bundle topology was incorrect. It is interesting that Jones's relatively good results (see results of *ab initio* folding at URL: http://PredictionCenter.llnl.gov/) were due to the possibility, in his simulation algorithm, of introducing kinks in the secondary structure elements; that is, the secondary structure elements were not considered fixed during the simulations, as was assumed by the other researchers (Dunbrack *et al.*, 1997).

Discussion

Factors affecting the performance of the approach

Given the myriad of problems that any *ab initio* folding algorithm must face, and the difficulties encountered so far, it is important to ascertain why the approach described here is reasonably successful. In our view, the ability to assemble low to moderate resolution structures of small proteins is related to the following features. First and foremost, MONSSTER does not require a precise description of secondary structure nor a large number of tertiary restraints to assemble the global topology. This is made possible by including generic protein-like features into the model, using sequence-specific terms and adjusting the restraint

implementation to the expected accuracy and precision of the predicted restraints. As a result, a substantial number of prediction errors are tolerated. Thus, one can employ existing secondary structure prediction algorithms and can focus the tertiary restraint derivation process on the generation of a relatively small number of tertiary contacts of enhanced reliability.

Turning explicitly to the latter, another important aspect of this approach is the increased signal-to-noise ratio in the predicted contacts obtained by restricting the analysis to residue covariation in the predicted core secondary structural elements. Such contacts are extracted through the use of a two-step procedure that minimizes the appearance of wrong pairings of secondary structure elements and generates a locally self-consistent set of restraints. We stress that prediction of contacts based only on the local threading of all possible pairs of secondary

```
1fas
|EEEEEEEEEE EEEEE    EEEEEEE   EEEE EEEE EEEE    EEEEEE    E E|RK
| EEEE        EE      EEEEE      EEEEE          EEEEEE       |PHD
|EEEEEEE      EEE    EEEEEEE    EEEE    EEEEEEEEEEEEEE    EEEE|PRD

1cis
|EEEE       HHHHHHHHHHHHH   EEEEEE EEEEEEEEE     EEEEE       EEEE E|RK
|       E    HHHHHHHHHH    EEEEEEEE             EEEEEEE     EE     |PHD
|EEEE EEEEEHHHHHHHH HHHHHEEEEEEEE HHHHHHHHHHH  EEEEEEEE EE EEEEEEEE|PRD

1pou
|HHHHHHHHHHHHHHHHHHHH   HHHHHHHHHHHEEEEE HHHHHHHHHH   HHHHHHH HHHHHHHHH E|RK
| HHHHHHHHHHHHHHHHH    HHHHHHHHHH       HHHHHHHHHHHHHHHHHHHHHHHHHHHHHH  |PHD
|HHHHHHHHHHHHH           HHHHHHHH  EEEE HHHH      HHHHHHHHHH  HHHHHHHHH|PRD

prota
|HHHHHHHHH   EEEEHHHHHHHHHHHH   HHHHHHHHHHHHHE E|RK
| HHHHHHHHH    HHHHHHHHHHHH  HHHHHHHHHHHHHH   |PHD
|  HHHHHHHH      HHHHHHHHHH  HHHH   HHHHHHHH EEE|PRD

1shg
|EEEEEE EEEEE   EEEEEEEEEE EEE      EEE    EEEE     EEEEE|RK
|  EEEEEEE        EEE    EEEEE     HHHEE      EEE    EE   |PHD
| EEEEEE       EEEEEE    EEEEEEE     EEEE    EEEEE EEEEEEE|PRD

1ife
|EEEEEEEEEE       HHHHHHHHHHHHHHHHEEEEEEEEE         HHHHHHHHHHHHH    EEEE         EEEEEEEEEE|RK
| EEEEEEEE        HHHHHHHHHHHH    EEEEEEE   HHHHHHHHHHHHHHHHHHHHHHHHH          EEEEEEE     |PHD
|EEEEEEEEE EEEE    HHHHHHHHHHH EEEEEEEEEE  HHHHHHHHHHHHH    EEEEEE EEEEE    EEEEEE EE EE|PRD

1ego
|EEEEEEE   HHHHHHHHHHHHHH    EEEEEEE HHHHH HHHHHHHHHEEE EE EEE    EEEE HHHHHHHHH EEE|RK
| EEEEEE    HHHHHHHHHHHH      EEEEEEE       HHH          E EEEEE      HHHHHHHHH   |PHD
|EEEEEEE EE HH    HHHHHHH   EEEEEEEEE       EEEE     EEEEEEEEE     EEEE HHHHHHHHE E|PRD

3icb
| HHHHHHHHHHHHH         HHHHHHHHHHHHHHH     HHHHHHHH     EEE HHHHH HHHHHHH|RK
|  HHHHHHHHHH           HHHHHHHHHHHHHHH    HHHHHHHHH         HHHHHHHHHH  |PHD
|HHHHHHH  HHHHHHHH  HHHHHHHHH HHHHHH      HHHH  HHHHHHH       HHHHHHHHHHHHH|PRD

1lea
|EEEE HHHHHHHHHHHHHHHHHHEEEE HHHHHHHH EE HHH  HHHHHHHHH EEEE EEE EEEEE EE|RK
| HHHHHHHHHHHHHHHHHHHH     HHHHHHHH     HHHHHHHHHHH   EEEE      EEE    |PHD
|                          HHHHHHHHH EEEEHHHHHHHHH   EEEEE      EEE EE E|PRD

1hmd
|HHHHHHHHHHHHHHHHHHH HHHHHHHHHHHHHHHHHHHHHHHHH  HHHHHHHHHHHHHHHHH  HHHHHHHHHHHH|RK
|HHHHHHHHHHHHHHHHH    HHHHHHHHHHHH HHHHHHHHHHHHHHHHHHHHHHHHHHHHHH   HHHHHHHHHHH|PHD
| HHHHHHHHHHHHH     EEEEHHHHHHHHHHHHHHHHHHHHHH    HHHHHHHHHHHHHHHH  EEEE HHHHHHHH|PRD

1gpt
|EEEEE EEE      HHHHHHHHHH EE EEE    EEEEEEEEE|RK
|  EEEE            HHHHH               EEEEE |PHD
|EEEEEEEEEE     HHHHH HHHHEEEEEE    EEE EE EE|PRD

T0042
|  HHHHHHHHHHHHHHHHHEEE HHHHHHHHH HHHHH  HHHHHHHHH      HHHHH HHHHHHHHHEEEE|RK
|    HHHHHHHHHHHHH     HHHHHHHHHH        HHHHHHHHHHHHHHHHHH      HHEHE      |PHD
```

Figure 6 (*continued on page* 438 *with legend*)

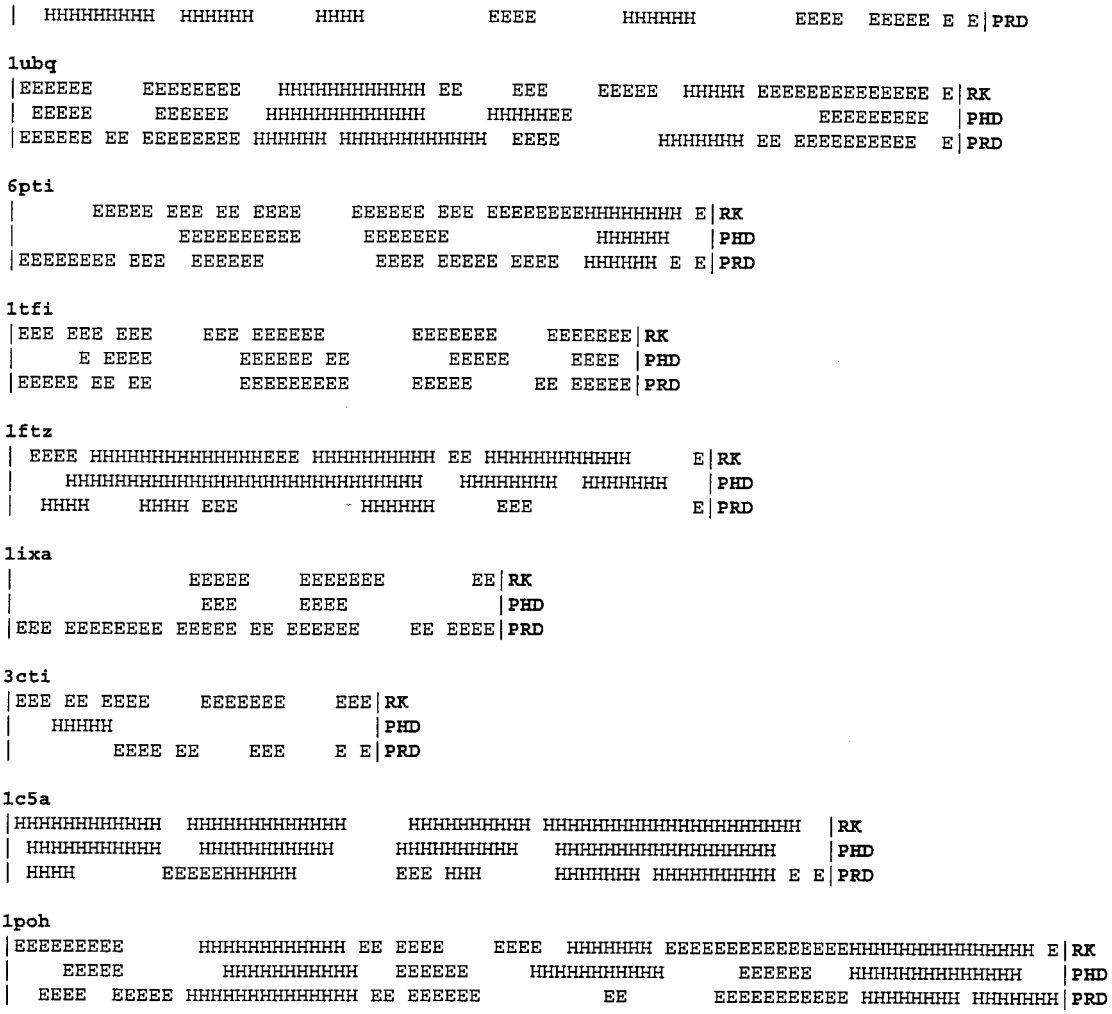


Figure 6. Observed and predicted secondary structures for all proteins studied in this work. Calculation of the secondary structure using the three-dimensional coordinates of the protein models was done using the Richards & Kundrot (1988) method. RK refers to the secondary structure assignment of the experimental structure; PHD to the predictions of secondary structure using the PHD method; and PRD is the secondary structure assignment based on the coordinates of the predicted structure.

structure elements is unreliable (Hu *et al.*, 1997; A.R.O. & J.S., unpublished results). For example, by using only local threading, almost all β-strands would like each other and it would be, in general, impossible to discriminate the β-sheet pattern in α/β and β proteins. There is not enough specificity in the potential (local secondary structure plus burial) to discriminate reliably among the preferred pairings of β-strands. These limitations of local threading are well known and have been reported by us (Hu *et al.*, 1997) and others (Hubbard & Park, 1995). It is the use of a few contacts derived from correlated mutations obtained with high reliability together with the fragment threading/clustering protocol of restraint growing that yields restraints of sufficient quality for successful structure assembly.

An interesting observation made here is that it is better not to implement a restraint than to include grossly wrong information. The appearance of false positives is one of the main factors affecting the performance of the approach. Another interesting outcome is that the average overall precision of the predicted restraints is more important than the average overall accuracy in low resolution folding simulations. This effect can be appreciated in Figure 8(A), where a correlation can be observed between precision at $\delta = 2$ and cRMSD for α-helical proteins on the one hand and β-containing proteins on the other. For two of the proteins, 1ixa and 6pti, that do not lie in any of the correlation lines, the discrepancy can be explained by the good accuracy of the contact predictions in both cases (Figure 8(B)). However, no clear correlation with the cRMSD was found in the case of the accuracy (Figure 8(B)). Moreover, it is interesting to note that the dependency of the quality of the predicted fold on the precision of the restraints is higher for α-helices than for β-containing proteins, something

Table 5. Accuracy of secondary structure prediction

Protein	Q3_PRD	Q3_PHD
3cti	61.176	50.588
1ixa	57.647	70.588
protA	69.412	77.647
1gpt	80.000	70.588
1tfi	77.647	60.000
6pti	69.412	58.824
1fas	77.647	67.059
1shg	70.588	67.059
1cis	64.706	64.706
1ftz	54.118	63.529
1pou	71.765	78.824
1c5a	62.353	85.882
3icb	68.235	82.353
1ubi	75.294	62.353
T0042	42.353	61.176
1lea	44.706	63.529
1ego	57.647	60.000
1hmd	77.647	90.588
1poh	58.824	65.882
1ife	67.059	75.294
AVER	65.412	68.824
SDEV	10.366	9.875

Q3_PRD refers to the Q3 value obtained from the predicted tertiary structure model, while Q3_PHD refers to the Q3 value obtained using the PHD method. Secondary structure assignment of the three-dimensional structures is made according to the Richards & Kundrot (1988) definition.

probably related to the difficulties of modeling β-strands.

Comparison with previous studies

Different authors have now started to investigate the use of multiple sequence information and/or predicted secondary structure using different fold prediction techniques. Perhaps the field that has most strongly pursued by this approach is threading. There, as recently reported by different authors, the use of variability patterns in multiple sequence alignments (Defay & Cohen, 1996; Taylor, 1997) or predicted secondary structure (Rice & Eisenberg, 1997; Rost *et al.*, 1997; Russell *et al.*, 1996), provide considerable improvement over the single sequence approach. But, and at least up to the best of our knowledge, the only reported application of predicted restraints in threading is due to Russell *et al.* (1996). In their study, putative contacts derived from biochemical arguments were used as filters in a fold recognition technique that makes use of predicted secondary structure. However, no distance restraint information coming from multiple sequence alignments has been reported in threading. The reasons for this might be that the predicted contact information has been regarded as not being reliable, and certainly it is not without post-processing, and that introduction of distance restraints requires incorporation of double dynamic programming techniques (Taylor, 1997) or Monte Carlo approaches, making the threading procedure rather computationally expensive.

On the other hand, the combined effort of the Cohen and Benner groups has produced approaches recently to fold prediction based on hierarchical building procedures that are similar in spirit to the one presented here (Gerloff *et al.*, 1997a,b). In their case, multiple sequence alignments are built and secondary structure is predicted. Later, an analysis of compensatory variations found in the multiple sequence alignment between pairs of positions is carried out. This allows the authors to derive relationships of closeness in space between the secondary structure elements. In the final stage of their prediction method, folds are searched in the database that meet at least a fraction of the predicted structural features. *Bona fide* predictions of the C-terminal domain of the β and γ chains of fibrinogen have been published making use of this approach (Gerloff *et al.*, 1997a).

Perhaps the most similar approach to fold prediction to that presented here has been suggested by Aszodi & Taylor (1995). Their model studies using simulated restraints were very similar to our previous studies on the determination of the requisite number of exact restraints necessary for successful fold assembly (Skolnick *et al.*, 1997b). However, later studies by Aszodi & Taylor (1996) have addressed the problem of remote homology modeling rather than fold prediction (Aszodi & Taylor, 1996), and also follow an approach similar in spirit to the one presented here. In their case, multiple sequence alignments are used to define conserved regions that are assumed to form part of the protein core. A fitting function derived from the protein database is used to map restraint distances from these conservation patterns. Aszodi & Taylor (1996) have shown convincingly that this is a promising technique for remote homology modeling.

Limitations of the current methodology

While the results described above are encouraging, there are problems with this approach that must be addressed. First and foremost, the yield of native topologies is only about 10 to 20%, and extraction of correct from incorrect topologies requires a long series of isothermal simulations. The fact that different assembly runs produce a wide dispersion in the final energies of the protein model (even for the same overall global fold) is a signature of sampling problems. While sampling is probably adequate for 50 to 60-residue proteins, sampling problems become acute as the size of the protein (or more precisely, the number of topological elements) increases. Thus, the development of better sampling approaches should permit us to extend the treatment to larger proteins having more complicated topologies. Another difficulty is related to fold selection. Typically, one of two situations arises. Either one must differentiate the native topology from its topological mirror image (a fold where the chirality of the secondary struc-

Table 6. Multiple sequence alignment used in the contact prediction of target T0042

NR	ID	%IDE	ACCNUM	PROTEIN_SEQUENCE
0	: NK-lys	1.00	——	GYFCESCRKIIQKLEDMVGPQPNEDTVTQAASQVCDKLKILRGLCKKIMRSFLRRISWDILTGKKPQAICVDIKICKE
1	: nkg5_human	0.33	P22749	GRDYRTCLTIVQKLKKMVD.KPTQRSVSNAATRVCRTGRswRDVCRNFMRRYQSRVIQGLVAGETAQQICEDLRLC
2	: yog2_caeel	0.33	P34611	GQFTEPSGVAVNGQGDIVVADTNNHRI.....QVFDkfKFQFGECGKRDgqFLRKFGANILQ..HPRGVCVDSK
3	: pp1b_drome	0.33	P48462	GDFDLNVDSLIQRLLEMRSCRTGK.........QVQMTEAEVRGLCLKSREIFLQQPI..LLELEAPLIICGDIH
4	: pspb_rat	0.28	P22355	LCQECEDIVHLLTKMTKEDAFQDTIRKFLEQECdpLKLLVPRCRQVLDVYLPLVIDYFQGQIKPKAICSHVGLC
5	: xyla_thene	0.26	P45687	FDAKVRRASYKVEDLFIGHIAGMDTFALGFKVAYKlgVLDKFIEEKYRSFREGIGRDIVEGKVDFEKLEEYIIDKE
6	: fbn1_bovin	0.26	P98133	GTPCELCPPVNTSEYKILCprPNPITVILEDIDECQELPGlgGKCINTFGSFQCRCPTGYYL.NEDTRVCDDVNECET
7	: fbn1_human	0.26	P35555	GTPCEMCPAVNTSEYKILCprPNPITVILEDIDECQELPGlgGKCINTFGSFQCRCPTGYYL.NEDTRVCDDVNECET
8	: vnsm_insv	0.26	Q01268	FCDSPRADLDKSCMIIPINRAIRAKSQAFIEAC.KLIIPKGNSEKQIRRQLAELSANLEKSVEEEENVTDNKI
9	: rpb1_eupoc	0.25	P28364	CSTCQGDSKECPGHFGHIELAQPVFHIgdLVKKILKCVCFNCNKLlySALKRVKDPKLKLNKVYKVCKDIKVCGK
10	: pu92_scico	0.26	P22312	KECQKNTENLKETIEQLKKELAEAQKALEKCKKEL...ADCKKENAKLLNKIecQLDECKKKLNICNNELI
11	: glne_ecoli	0.24	P30870	GYFEEDDRKQVLTLIADFRKELDKRTIGPRGRQVLDHlhLLSDVCAREDAAVLSRItyLELLSEFPAALKHLISLCAA
12	: ynh4_caeel	0.24	P32742	RYVCSSHDVTIHGLAAMLRDRYPEYDVPQRFPGIQDDLQPVRFSSKK.....LQDLGFTFRYKTLEDMFDAAIRTCQE
13	: imdh_bacsu	0.24	P21879	RYFQEENKKFVP..EGIEGRTPYKGPVEETVYQLVGGLRSGMGYCGSKDLRALrrMTGAGLRESHPHDVQITVHRN
14	: pspb_pig	0.24	P15782	FCWLCRTLIKRIQAVVP....KGVLLKAVAQVCHVVPlvGGICQCLAERYIVICLNMLLDRTLPQLVCGLVLR
15	: dfra_horvu	0.23	P51106	RYICSSHDATIHGLARMLQDRFPEYDIPQKFAGVDDNLQPIHFSSKKLlhGFSFRYTTEDMFD.AAIHTCRDKGL

NR is the sequence number in the alignment. The number 0 stands for the target sequence.
ID is the identification number according to the EMBL/SWISSPROT data base (except the target sequence).
%IDE is the percentage of identity between the corresponding sequence and the target sequence.
ACCNUM corresponds to the entry number in the EMBL/SWISSPROT data base (except the target sequence).
PROTEIN_SEQUENCE is the sequence alignment used for contact prediction. Lower case letters indicate that in that position an insertion (not shown) is found in the corresponding sequence. The alignment was created by scanning the SWISSPROT data base with FASTA, filtering the sequences with MAXHOM, and then finally selecting the sequences by hand.

Table 7. Assignment of secondary structure for T0042 from the CASP2 meeting

```
          10        20        30        40        50        60        70
           |         |         |         |         |         |         |
|GYFCESCRKIIQKLEDMVGPQPNEDTVTQAASQVCDKLKILRGLCKKIMRSFLRRISWDILTGKKPQAICVDIKICKE|   SEQUENCE
|    HHHHHHHHHHHHHHHH      HHHHHHHHHHHHHH     HHHHHHHHHH      HHHHH     HHHHHHH      |   OBSEC
|     HHHHHHHHHHHHHHH      HHHHHHHHHHH          HHHHHHHHHHHHHHHHHHHH       HHEHE      |   PDSEC (PHD)
|      HHHHHHHHHHHHH         HHHHHHHH           HHHHHHHHHHHHHHHHHHHH                   |   PDSEC (PHD + LINKER)
|555552222222222221113333322222225555533332222222222222222225553335511111155 5|   Assigned local structure
```

The sequence of the predicted protein is shown (SEQUENCE), together with the observed secondary structure in the experimental structure (OBSEC); the predicted secondary structure using the PHD method (PDSEC (PHD)); the predicted secondary structure when the PHD and LINKER predictions are combined (PDSEC (PHD + LINKER)); and the local secondary structure assigned in the simulations (see Methods for bin description).

Table 8. Predicted seeds for target T0042 using correlated mutation analysis

Residue number A	Residue number B	Residue name A	Residue name B	Correlation coefficient
34	61	V	L	0.662
6	56	S	I	0.588
7	70	C	C	0.535

Three contacts were predicted, each one of them being an entry in the Table. Residues A and B refer to the first and second partner of the predicted contact, respectively. The correlation coefficient of the mutational behavior of the corresponding positions in the multiple sequence alignment is also shown (see Methods for details).

tural elements is the same, but the chirality of the turns is reversed; Pastore *et al*., 1991), or there are a handful of distinct folds, some having a subset of their structures in common. Thus, the resulting energy differences between the different low energy topologies are small and on the order of the standard deviation of the mean energy per fold obtained from independent runs. We hope that the specificity for the native topology could be accentuated by the development of better energy functions.

A major effort is required to devise better methods of tertiary restraint derivation so that false positives in contact map prediction are minimized. If restraints could be predicted with higher reliability, then a tighter restraint function could be used that would eliminate some of the misfolded states currently encountered. For example, to account for the possibility of wrong restraints between non-interacting β-strands in β and α/β proteins, the restraint potential has to be rather permissive, resulting in a higher population of competing misfolded states. This work suggests that the accuracy of the current methods of contact prediction is already more or less adequate, but improvement in precision is still required.

Finally, secondary restraint derivation also needs improvement. While the PHD and LINKER algorithms have been combined in an *ad-hoc* manner, a more consistent protocol would be desirable. Overall, this study indicates that to predict low to moderate resolution structures, the prediction of the exact secondary structure element boundaries is not an important factor. However, missing a secondary structure element can have a strong negative impact on the results, particularly if the element belongs to the protein core. In this regard, our studies are in agreement with the recent findings of Dandekar & Argos (1994, 1996) and Simons *et al*. (1997).

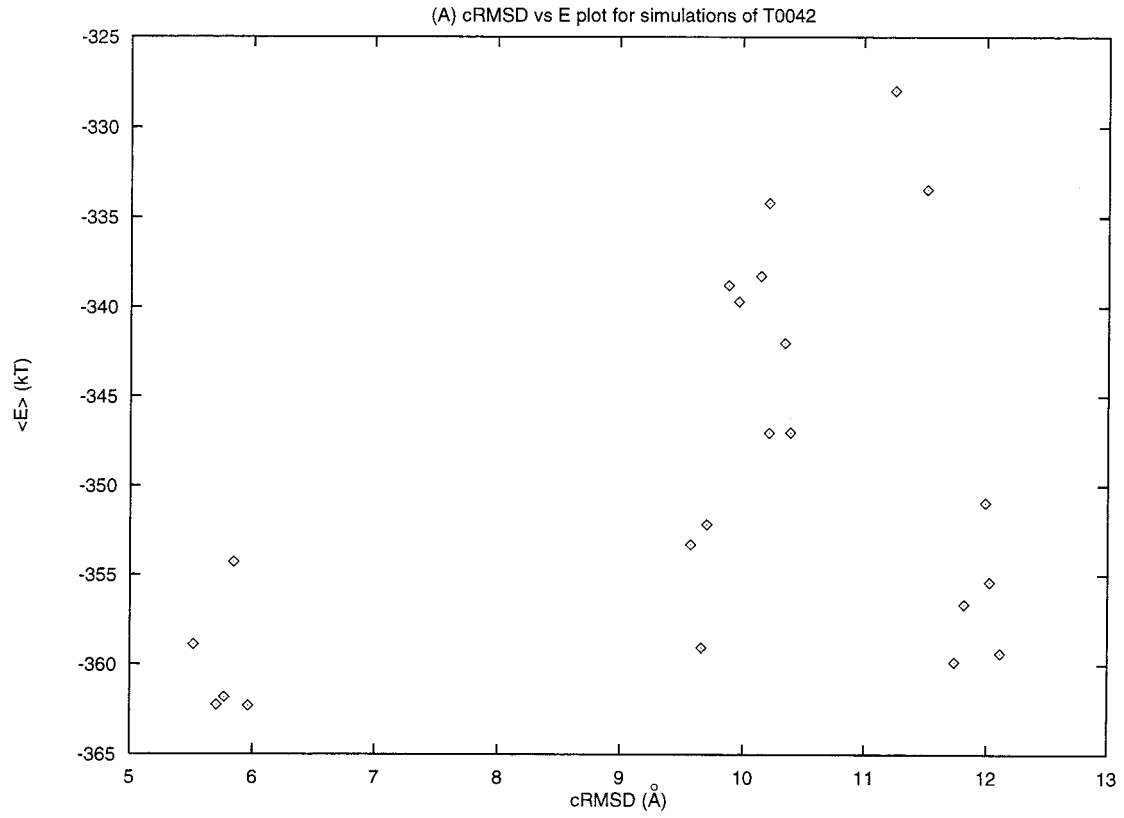
Conclusions

This paper has addressed the feasibility of deriving restraints from multiple sequence alignments for use in restraint assisted simulations designed to predict the native conformation of small proteins. Initial application has been made to a set of 20 different proteins, representing all secondary structural classes and having a wide variety of topolo-

Table 9. List of contacts used in the structure prediction of T0042

Res. A (T0042)[a]	Res. B (T0042)[b]	Template structure[c]	Res. A (template)[d]	Res. B (template)[e]
7	55	1mba	89	137
7	56	1mba	89	138
7	59	1mba	89	141
10	56	1mba	92	138
10	60	1mba	92	142
11	59	1mba	93	141
11	60	1mba	93	142
14	60	1mba	96	142
15	60	1mba	97	142
30	34	2utg	22	26
30	43	2utg	22	35
30	47	2utg	22	39
33	38	2utg	25	30
33	43	2utg	25	35
33	47	2utg	25	39
34	30	2utg	26	22
34	43	2utg	26	35
35	31	2utg	27	23
38	43	2utg	30	35
34	61	SEED	–	–
6	56	SEED	–	–
4	76	Disulfide bridge	–	–
35	45	Disulfide bridge	–	–

[a] First residue in the predicted contact in the target sequence of unknown structure.
[b] Second residue in the predicted contact in the target sequence of unknown structure.
[c] Template structure from which the predicted contact is extracted. The protein name corresponds to the PDB entry. The last four contacts are either from the correlated mutation analysis or from the known disulfide bridge partner of the structure.
[d] Residue number in the template structure of the first partner of the predicted contact.
[e] Residue number in the template structure of the second partner of the predicted contact.



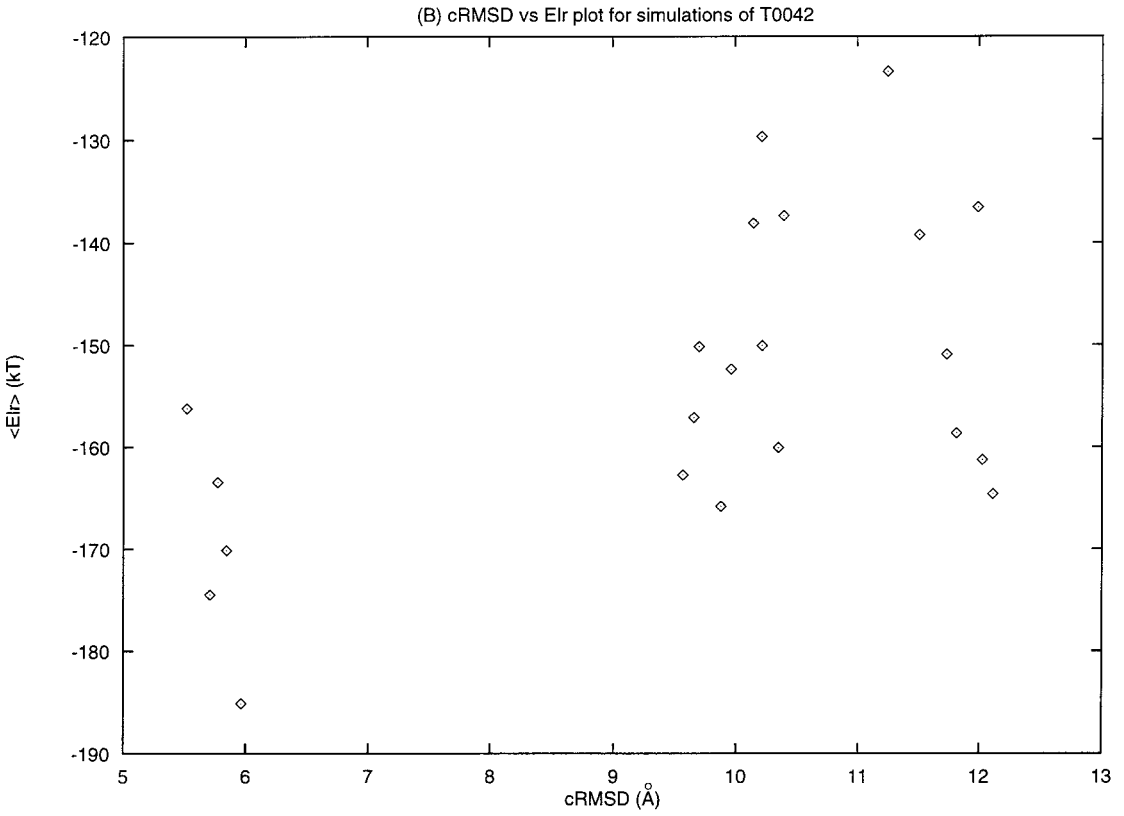
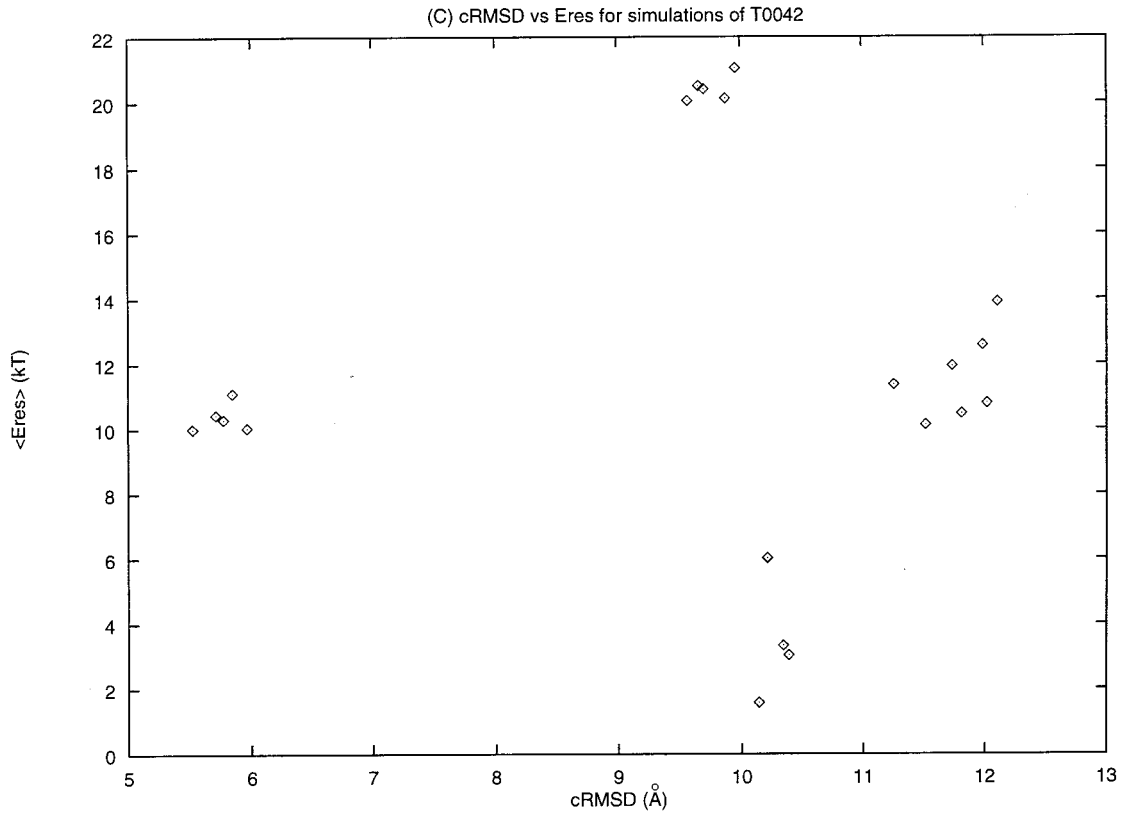


Figure 7(A–B) (*legend on page* 444)



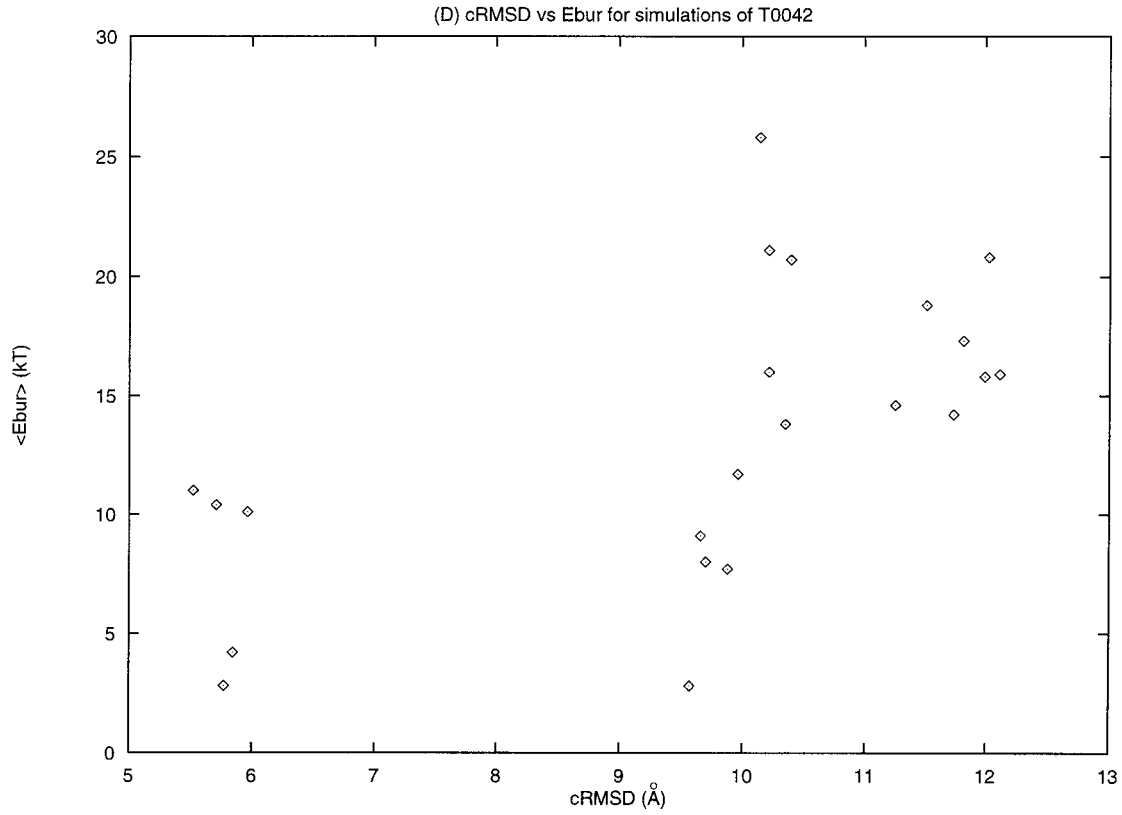


Figure 7. Plots of energy *versus* coordinate RMSD for the different folds obtained for T0042. (A) Total average energy of the isothermal run *versus* final coordinate RMSD of the fold. (B) Pairwise energy (E_{lr}) *versus* cRMSD. (C) Restraint energy (E_{res}) run *versus* final coordinate RMSD of the fold. (D) Burial energy (E_{bur}) *versus* cRMSD.

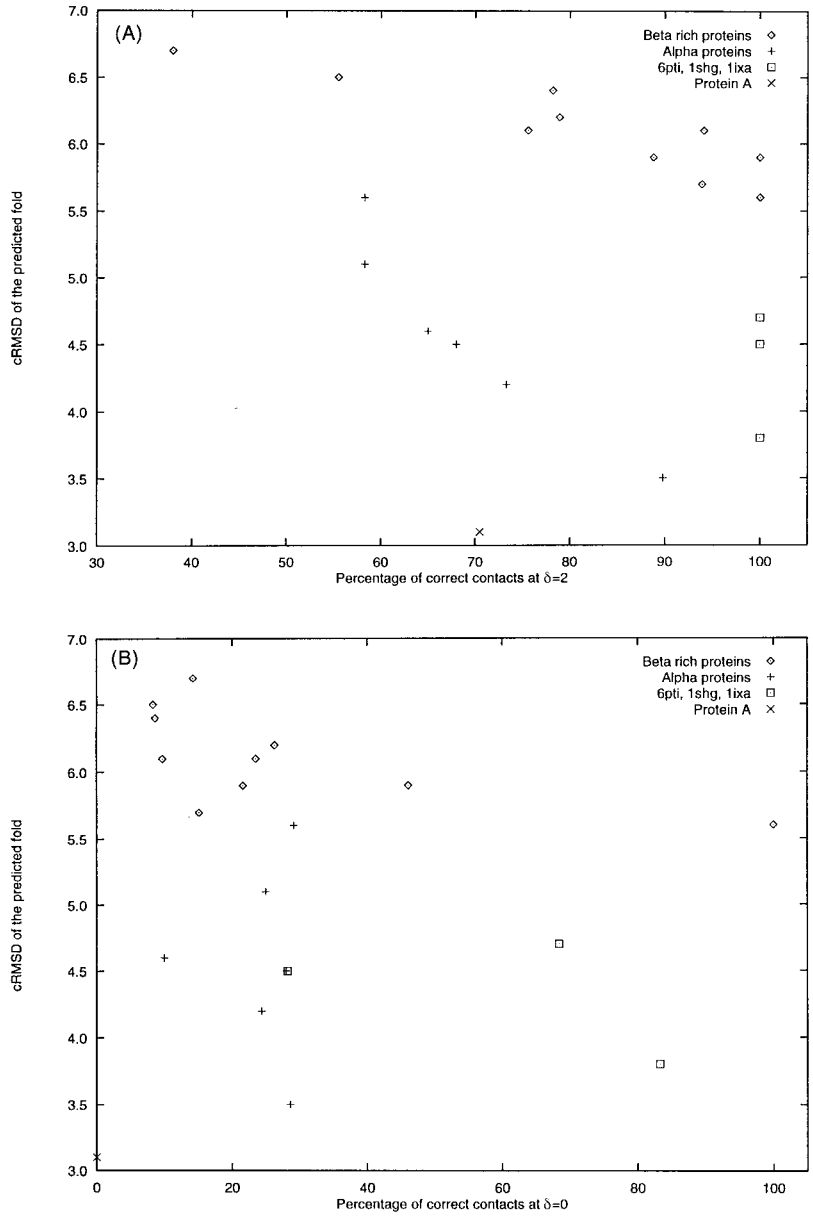


Figure 8. (A) Precision *versus* cRMSD for the set of proteins used in this work. The cRMSD values shown in bold in Table 3 are used in this plot. The $\delta = 2$ level was used (i.e. percentage of correctly predicted contacts allowing for a ± 2 residue error in prediction with respect to a correct contact), although similar results are obtained for $\delta = 1$ and $\delta = 3$. (B) The same plot, but using accuracy criteria (i.e. percentage of correctly predicted contacts without any residue error with respect to the correct contact, so that $\delta = 0$).

gies. From the experience gained so far, these conclusions can be drawn.

(1) At the level of secondary structure elements, the accuracy of existing secondary structure prediction algorithms appears to be acceptable for successful fold assembly, provided that some tertiary restraints are included in the assembly algorithm. Problems encountered with current secondary structure prediction algorithms arise either from entirely missing a secondary structure element or mistaking the secondary structural class (helix or β) of some elements. Depending on the position of the missed element in the global fold, this might or might not prohibit fold assembly. In some cases, when the missed elements lie at the edge of the fold, the ability to assemble the native topology is not affected. In fact, sometimes such missing elements are induced by tertiary interactions. If, however, the element missed corresponds to a crucial core element (e.g. a helix between two β-strands), it can exert a strong influence on the quality of the final predicted structure and in the worst case, could result in a grossly incorrect predicted structure.

(2) Low resolution models of small proteins can be assembled from rather inaccurate predictions of a subset of the total number of tertiary side-chain contacts. These predicted side-chain contacts need not span the entire structure, but can be strongly clustered. On average, the required level of accuracy in contact map prediction is on the order of 25%, provided that the predictions are reasonably precise, i.e. 95% lie within ± 4 residues of correct contacts. About 25% of the total number of side-chain contacts need to be identified. This study strongly suggests that precision is a more important factor than accuracy in determining the likeli-

hood of successful fold assembly. Although additional investigation is necessary to assess its generality, our results suggest that the required restraints can be reliably derived from multiple sequence alignments using a combination of correlated mutations followed by structural filtration/inverse folding (additional details will be given in a forthcoming publication).

(3) The strategy presented here yields predictions for all fold types whose RMSD from native ranges from 3.0 to 6.5 Å, depending upon fold complexity. Such structures are at the level of accuracy that can be obtained when threading techniques are applied to match a sequence to a structure whose sequence homology lies in or below the twilight zone of sequence identity. In the folding simulations, the accuracy and yield of correctly assembled structures is different for the different protein classes. In general, all-α proteins are predicted with higher accuracy than α/β proteins, and these, in turn, have better accuracy than all-β proteins.

(4) Because of errors in the tertiary restraints, the resulting structures exhibit shifts in registration and distorted mutual orientations of pairs of secondary structural elements. Such structural distortions also reduce the energy gap between the putative native conformation and alternative folds, as compared to our previous studies using ''exact'' restraints. Thus, the present protocol allows for the prediction of a small number of possible native conformations. However, selection of the specific fold based on the force field energy is more uncertain as a result of problems with the current energy function.

This study successfully demonstrates that, for a set of small proteins, the use of restraints derived from multiple sequence alignments incorporated into a tertiary structure prediction algorithm allows for assembly of native-like structures. The approach has been shown to be capable of assembling low resolution tertiary structures of greater complexity than was previously possible. However, the difficulties encountered in the assembly of these topologies, even when low resolution restraints are employed, paints a cautionary picture for the likelihood of *ab initio* assembly of complex folds without restraints. Given contemporary computer resources, sampling algorithms and existing force fields, the folding of more complex topologies is likely to be problematic. Probably, over the short term, the only way to progress is to combine insights gained from restraint-free folding studies on simple folds with strategies that reduce the conformational search space.

Acknowledgments

This work is supported by grant GM-37408 from the National Institutes of Health. A.K. also acknowledges support from the University of Warsaw (grant BST-34/97) and is an International Scholar of the Howard Hughes Medical Institute. A.R.O. also acknowledges support from the Spanish Ministry of Education, as well as access to the computational facilities of EMBL during this work. A.R.O. also thanks Dr Wei-Ping Hu for sharing some of his source codes at the beginning of this project, as well as Drs Li Zhang, Leszek Rychlewski and Adam Godzik for useful discussions. Finally, we thank both reviewers for their careful reading of the manuscript and useful suggestions that have significantly improved the quality of the text.

References

Aszodi, A. & Taylor, W. R. (1996). Homology modelling by distance geometry. *Folding Design,* **1**, 325–334.

Aszodi, A., Gradwell, M. J. & Taylor, W. R. (1995). Global fold determination from a small number of distance restraints. *J. Mol. Biol.* **251**, 308–326.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Simanouchi, T. & Tasumi, M. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542.

Chothia, C. & Finkelstein, A. (1990). The classification and origins of protein folding patterns. *Annu. Rev. Biochem.* **59**, 1007–1039.

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

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

Dandekar, T. & Argos, P. (1996). Identifying the tertiary fold of small proteins with different topologies from sequence and secondary structure using the genetic algorithm and extended criteria specific for strand regions. *J. Mol. Biol.* **256**, 645–660.

Defay, T. R. & Cohen, F. E. (1996). Multiple sequence information for threading algorithms. *J. Mol. Biol.* **262**, 314–323.

Dunbrack, R. L., Gerloff, D. L., Bower, M., Chen, X., Lichtarge, O. & Cohen, F. E. (1997). Meeting review: the second meeting on the critical assessment of techniques for protein structure prediction (CASP2), Asilomar, California, December 13–16, 1996. *Folding Design,* **2**, R27–R42.

Friesner, R. A. & Gunn, J. R. (1996). Computational studies of protein folding. *Annu. Rev. Biophys. Biomol. Struct.* **25**, 315–342.

Gerloff, D. L., Cohen, F. E. & Benner, S. A. (1997a). A predicted consensus structure for the C-terminus of the beta and gamma chains of fibrinogen. *Proteins: Struct. Funct. Genet.* **27**, 279–289.

Gerloff, D. L., Cohen, F. E., Korostensky, C., Turcotte, M., Gonnet, G. H. & Benner, S. A. (1997b). A predicted consensus structure for the N-terminal fragment of the Heat Shock Protein HSP90 family. *Proteins: Struct. Funct. Genet.* **27**, 450–458.

Godzik, A., Skolnick, J. & Kolinski, A. (1992). A topology fingerprint approach to the inverse protein folding problem. *J. Mol. Biol.* **227**, 227–238.

Godzik, A., Kolinski, A. & Skolnick, J. (1994). Lattice representation of globular proteins: how good are they? *J. Comput. Chem.* **14**, 1194–1202.

Göebel, U., Sander, C., Schneider, R. & Valencia, A. (1994). Correlated mutations and residue contacts in proteins. *Proteins: Struct. Funct. Genet.* **18**, 309–317.

Gouda, H., Torigoe, H., Saito, A., Sato, M., Arata, Y. & Schimada, I. (1992). Three-dimensional solution structure of the B-domain of staphylococcal Protein A: comparisons of the solution and crystal structures. *Biochemistry,* **40**, 9665–9672.

Gunn, G. J. R., Monge, A. & Friesner, R. A. (1994). Hierarchical algorithm for computer modeling of protein tertiary structure: folding of myoglobin to 6. 2 Å resolution. *J. Phys. Chem.* **98**, 702–711.

Hu, W.-P., Godzik, A. & Skolnick, J. (1997). Sequence-structure specificity: how does an inverse folding approach work? *Protein Eng.* **10**, 317–331.

Hubbard, T. J. & Park, J. (1995). Fold recognition and ab initio structure predictions using hidden Markov models and beta-strand pair potentials. *Proteins: Struct. Funct. Genet.* **23**, 398–402.

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

Kolinski, A. & Skolnick, J. (1994a). Monte Carlo simulations of protein folding: I. Lattice model and interaction scheme. *Proteins: Struct. Funct. Genet.* **18**, 338–352.

Kolinski, A. & Skolnick, J. (1994b). Monte Carlo simulations of protein folding: II. Application to protein A, ROP, and crambin. *Proteins: Struct. Funct. Genet.* **18**, 353–366.

Kolinski, A. & Skolnick, J. (1997). Determinants of secondary structure of polypeptide chains: interplay between short range and burial interactions. *J. Chem. Phys.* **107**, 953–964.

Kolinski, A., Galazka, W. & Skolnick, J. (1995a). Computer design of idealized β-motifs. *J. Chem. Phys.* **103**, 10286–10297.

Kolinski, A., Milik, M., Rycombel, J. & Skolnick, J. (1995b). A reduced model of short range interactions in polypeptide chains. *J. Chem. Phys.* **103**, 4312–4323.

Kolinski, A., Skolnick, J., Godzik, A. & Hu, W.-P. (1997). A method for the prediction of surface ''U''-turns and transglobular connections in small proteins. *Proteins: Struct. Funct. Genet.* **27**, 290–308.

Koradi, R., Billeter, M. & Wuethrich, K. (1996). MOLMOL: a program for display and analysis of macromolecular structures. *J. Mol. Graph.* **14**, 51–55.

Levitt, M. & Greer, J. (1977). Automatic identification of secondary structure in globular proteins. *J. Mol. Biol.* **114**, 181–293.

McLachlan, A. D. (1971). Test for comparing related amino acid sequences. *J. Mol. Biol.* **61**, 409–424.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.* **51**, 1087–1092.

Mumenthaler, C. & Braun, W. (1995). Predicting the helix packing of globular proteins by self-correcting distance geometry. *Protein Sci.* **4**, 863–871.

Murzin, A., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of protein database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540.

Olmea, O. & Valencia, A. (1997). Improving contact predictions by the combination of correlated mutations and other sources of sequence information. *Folding Design,* **2**, S25–S32.

O'Shea, E. K., Klemm, J. D., Kim, P. S. & Alber, T. (1991). X-ray structure of the GCN4 leucine zipper, a two-stranded, parallel coiled coil. *Science,* **254**, 539–544.

Park, B. H. & Levitt, M. (1995). The complexity and accuracy of discrete state models of protein structure. *J. Mol. Biol.* **249**, 493–507.

Pastore, A., Atkinson, R. A., Saudek, V. & Williams, R. J. P. (1991). Topological mirror images in protein structure computation: an underestimated problem. *Proteins: Struct. Funct. Genet.* **10**, 22–32.

Pearson, W. R. & Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA,* **85**, 2444–2448.

Rice, D. W. & Eisenberg, D. (1997). A 3D-1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence. *J. Mol. Biol.* **267**, 1026–1038.

Richards, F. M. & Kundrot, C. E. (1988). Identification of structural motifs from protein coordinate data: secondary structure and first level super-secondary structure. *Proteins: Struct. Funct. Genet.* **3**, 71–84.

Ripoll, D. R. & Scheraga, H. A. (1990). On the multiple minima problem in the conformational analysis of polypeptides. IV. Application of electrostatically driven Monte Carlo method to the 20–residue membrane bond portion of melittin. *Biopolymers,* **30**, 165–176.

Rost, B. & Sander, C. (1993). Prediction of secondary structure at better than 70% accuracy. *J. Mol. Biol.* **232**, 584–599.

Rost, B. & Sander, C. (1996a). Bridging the protein sequence-structure gap by structure predictions. *Annu. Rev. Biophys. Biomol. Struct.* **25**, 113–136.

Rost, B. & Sander, C. (1996b). Progress of 1D protein structure prediction at last. *Proteins: Struct. Funct. Genet.* **23**, 295–300.

Rost, B., Schneider, R. & Sander, C. (1997). Protein fold recognition by prediction-based threading. *J. Mol. Biol.* **270**, 471–480.

Russell, R. B., Copley, R. C. & Barton, G. J. (1996). Protein fold recognition by mapping predicted secondary structures. *J. Mol. Biol.* **259**, 349–365.

Sali, A. & Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779–815.

Sander, C. & Schneider, R. (1991). Database of homology derived protein structures and the structural meaning of sequence alignment. *Proteins: Struct. Funct. Genet.* **9**, 56–68.

Simons, K. T., Klooperberg, C., Huang, E. & Baker, D. (1997). Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.* **268**, 209–225.

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

Skolnick, J., Jaroszewski, L., Kolinski, A. & Godzik, A. (1997a). Derivation and testing of pair potentials for protein folding. When is the quasichemical approximation correct? *Protein Sci.* **6**, 676–688.

Skolnick, J., Kolinski, A. & Ortiz, A. R. (1997b). MONSSTER: A method for folding globular proteins with a small number of distance restraints. *J. Mol. Biol.* **265**, 217–241.

Smith-Brown, M. J., Kominos, D. & Levy, R. M. (1993). Global folding of proteins using a limited number of distance restraints. *Protein Eng.* **6**, 605–614.

Sun, S. (1993). Reduced representation model of protein structure prediction: statistical potential and genetic algorithms. *Protein Sci.* **2**, 762–785.

Taylor, W. R. (1997). Multiple sequence threading: an analysis of alignment quality and stability. *J. Mol. Biol.* **269**, 902–943.

Vieth, M., Kolinski, A., Brooks, C. L., III & Skolnick, J. (1994). Prediction of the folding pathways and structure of the GCN4 ''leucine zipper''. *J. Mol. Biol.* **237**, 361–367.

Wallqvist, A. & Ullner, M. (1994). A simplified amino acid potential for use in structure prediction of proteins. *Proteins: Struct. Funct. Genet.* **18**, 267–289.

Edited by F. Cohen

(Received 19 August 1997; received in revised form 5 December 1997; accepted 5 December 1997)