COMBINED MULTIPLE SEQUENCE REDUCED PROTEIN MODEL APPROACH TO PREDICT THE TERTIARY STRUCTURE OF SMALL PROTEINS

ANGEL R. ORTIZ[1], ANDRZEJ KOLINSKI[1,2], JEFFREY SKOLNICK[1]

[1]*Department of Molecular Biology, The Scripps Research Institute, La Jolla, CA 92037 (USA)*
[2]*Department of Chemistry, University of Warsaw, Pasteura-1, 02-093, Warsaw, Poland*

By incorporating predicted secondary and tertiary restraints into *ab initio* folding simulations, low resolution tertiary structures of a test set of 20 nonhomologous proteins have been predicted. These proteins, which represent all secondary structural classes, contain from 37 to 100 residues. Secondary structural restraints are provided by the PHD secondary structure prediction algorithm that incorporates multiple sequence information. Predicted tertiary restraints are obtained from multiple sequence alignments via a two-step process: First, "seed" side chain contacts are identified from a correlated mutation analysis, and then, the seed contacts are "expanded" by an inverse folding algorithm. These predicted restraints are then incorporated into a lattice based, reduced protein model. Depending upon fold complexity, the resulting nativelike topologies exhibit a coordinate root-mean-square deviation, cRMSD, from native between 3.1 and 6.7 Å. Overall, this study suggests that the use of restraints derived from multiple sequence alignments combined with a fold assembly algorithm is a promising approach to the prediction of the global topology of small proteins.

1. Introduction

The question of how to relate a protein sequence to its native structure is commonly referred to as the protein folding problem[1]. It is widely believed that proteins obey the "thermodynamic hypothesis". This says that the protein's native conformation corresponds to a global free energy minimum[2]. However, due to the complexity of the interactions, the task of finding this free energy minimum in the myriad of multiple minima on the free energy landscape[3] is extremely difficult.

One way of partly surmounting the conformational search problem is to employ restraint information in the folding simulations. Such restraints might include known or predicted secondary structure and/or tertiary contacts. Assuming known secondary structure and using a genetic algorithm to search conformational space, Dandekar and

Argos[4] reported encouraging results for simple helical and β proteins. Furthermore, Mumenthaler and Braun[5] have developed a self-correcting distance geometry method that assumes known secondary structure and that successfully identified the native topology for 6 of 8 helical proteins. There have also been a number of studies that incorporate the known, correct secondary structure and a limited number of known, correct tertiary restraints to predict the global fold of a globular protein[6-8]. For example, the approach of Aszodi and Taylor[7] is in the spirit of Mumenthaler and Braun and is based on distance geometry, where a set of experimental tertiary distance restraints is supplemented by a set of predicted interresidue distances. These distances are obtained from patterns of conserved hydrophobic amino acids that have been extracted from multiple sequence alignments. They find that to assemble structures below 5 Å cRMSD, on average, typically more than N/4 restraints are required, where N is the number of residues. Even then, this method has problems selecting out the correct fold from competing alternatives. However, the Aszodi et al. approach is very rapid. More recently, Skolnick and coworkers have reported very encouraging results when N/4 exact tertiary restraints are employed[8] in their MONSSTER (MOdeling of New Structures from Secondary and TErtiary Restraints) algorithm, but the approach is computationally rather intensive.

In what follows, we explore whether use of predicted secondary structure and tertiary restraints are adequate to predict tertiary structure from sequence alone. If so, this would suggest a practical solution to the problem of tertiary structure prediction, at least for a subset of proteins that can be considered by this generation of models.

2. Methods

The tertiary structure prediction procedure can be logically divided into two parts: restraint derivation using information extracted from multiple sequence alignment and structure assembly/refinement using an improved version of the MONSSTER algorithm[8,9] modified to incorporate the expected accuracy and precision of the predicted tertiary restraints. A schematic overview of the entire approach is presented in Figure 1. In what follows, we discuss each aspect of the protocol in turn.

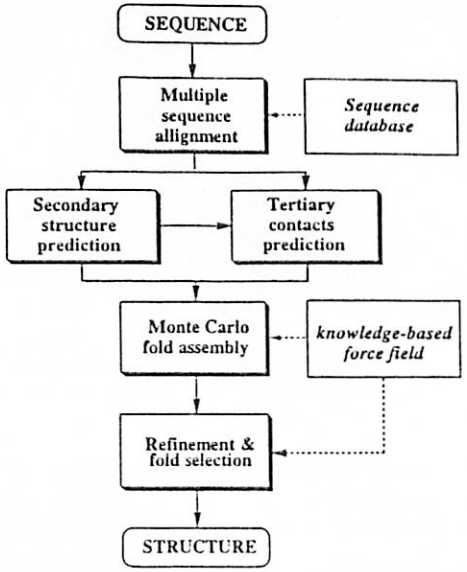

Figure 1. Schematic overview of the procedure for tertiary structure prediction.

2.1 Secondary Structure Prediction

Existing secondary structure prediction schemes provide a logical starting point for the prediction of secondary structure. Multiple sequence alignments are obtained from the HSSP database[10] and serve as input to the PHD secondary structure prediction algorithm[11,12]. Elements predicted as U-turns (regions where the chain reverses global direction) by our *Linker algorithm*[13] override PHD predictions because of their high prediction accuracy. At the end of this stage, residues are assigned to one of five states: strand, helix, U-turn, extended state/loop and non predicted. The set of predicted secondary elements (helix or strand) between U-turns comprises the putative core region of the molecule.

2.2 *Prediction of Tertiary Contacts*

Troughout this work a *contact* between two aminoacids is defined when any two heavy atoms of the corresponding side chains have a distance between them smaller than or equal to 5 Å in the native structure of the protein. It has been proposed by different authors that multiple sequence information can be used to predict contacts based on residue conservation[5,7] or covariation[14-16]. Both approaches might be combined for increased sensitivity[16]. Obviously, residues that are highly conserved are important, but it is quite difficult to differentiate functional from structural information. Residue covariation might be sensitive to variations arising from contacting pairs of topological elements. Here, for the sake of simplicity, we slightly modify the approach of Göbel and coworkers[14] and calculate the covariation between all residues predicted to be in the putative core of the molecule. The rationale is that by restricting contact predictions to those occurring between rigid elements of the putative core, the assumption of spatial closeness might be more correct. In practice, a relatively high cutoff of 0.5 for the correlation coefficient for pairwise mutations is used for contact prediction. Unfortunately, there are too few of these "seed" contacts to assemble a protein from the unfolded state using MONSSTER.

To enrich the set of seed contacts, we extract additional contacts via a combined structural fragment search and inverse folding procedure[9,17]. First, all pairs of secondary structural elements compatible with the predicted pair of contacting elements (±1 residue) are identified in a structural database. Then, the top 10 scoring fragment pairs are extracted based on their secondary structural propensities and burial energy. The cRMSD of all fragment pairs is calculated. If there is no clear clustering (with an upper limit of 5.5 Å for the most divergent pair), then additional side chain contact restraints are not derived. If the fragments cluster, then the average member of the cluster is selected and additional tertiary restraints are extracted. This procedure produces predicted contacts, 25% of which are exactly correct and about 60% of which are correct within ±1 residue. About 25% of the total number of contacts seen in the native state are obtained by this combined approach to contact prediction. The final outcome of the prediction protocol is a set of noisy secondary and tertiary restraints.

2.3 *Fold Assembly/Refinement/Selection*

The predicted secondary and tertiary restraints are entered into an improved version of MONSSTER[9] as flat-bottom harmonic functions. Implementation of the tertiary restraints takes into account their accuracy and precision by using both a residue pair dependent flat bottom with an increase of 50% over the average contact distance in a representative protein database, and use of the *restraint splinning* technique[9]. The protein model employs a Cα based lattice protein model and incorporates potentials reflecting statistical preferences for secondary structure, side chain burial, pair interactions, and hydrogen bond contributions. In addition, predicted U-turn regions experience an energetic bias to lie at the protein surface. In order to improve the packing of putative β strands, an interstrand hydrogen bond cooperativity term is introduced where β type residues having hydrogen bonds to residues in two different strands are energetically favored.

For each protein sequence, 10-40 independent simulated annealing simulations from a fully extended initial conformation are carried out. If, from repeated simulations, the structures do not cluster into a handful of distinct topologies, no structural prediction is made. If the structures cluster, then all low energy structures are subject to low temperature, isothermal refinement. The predicted structure is the one having the lowest average (roughly 5 kT per residue) and minimum energies.

3. Results

3.1 *Summary*

The above protocol has been applied to the set of 20 proteins listed in Table 1. Such a large test set is necessary to demonstrate that the current approach can handle a wide variety of folds and different secondary structure types. All are extrinsic to the set of proteins employed in the derivation of the potentials. It is very important to emphasize that all predictions use the identical parameter set and folding protocol. Table 1 also shows the accuracy of the predicted secondary structure and tertiary contacts as well as the results from the folding simulations. As is clearly indicated by the folding simulation results, in spite of the moderate accuracy of the predicted

secondary and tertiary restraints, the native topologies are recovered either as the best energy (in 18 of 20 cases) or the next best energy structure for all classes of protein structure. The average cRMSD of the structures having the native topology ranges from about 3 Å for some helical proteins to roughly 6 Å for β and α/β proteins. Since a structure with a cRMSD deviation from native of 6 Å might have one or more incorrect topological elements, we present representative predicted structures (the structure at the end of the run whose average energy is the lowest) alongside the experimentally determined conformation in Figure 2. The remaining 17 structures will be made available on our Web site[18].

Of the three cases that did not select the native conformation as being lowest in energy, these are related to the native fold as follows: The misfolded state of 1ixa results simply from the wrong placement of the C-terminal β strand; the topology of the remainder of the molecule is correct. 1hmd is a four-helix bundle whose topological mirror image is essentially isoenergetic with the native fold. The discrimination of helical bundles from topological mirror images is an outstanding problem that taxes the potential used in these models. The final misidentified protein, 1ife, actually has the same global topology as native, but a strand is shifted from the edge of one β-sheet to the back of the protein. Finally, by way of example, we focus on the case of a protein whose structure was unknown to us at the time the prediction was done, but whose results are typical of the method.

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

Prot[a]	Type	N[b]	Q_3[c]	N_p[d]	N_w[e]	δ=0[f]	δ=2[f]	rms_n[g]	E_n[h]	rms_w[i]	E_w[j]
3cti	small	29	82.4	6	0	83.3	100.	**3.8**	-107	6.7	-103
1ixa	small	39	97.4	5	0	100.	100.	**5.6**	-130	7.7	-131
1gpt	small	47	72.3	13	0	46.1	100.	**5.9**	-276	6.6	-142
1tfi	small	50	78.0	37	0	21.6	88.8	**5.9**	-202	7.0	-191
prota[k]	α	47	83.0	17	0	0.0	70.5	**3.1**	-246	9.4	-240
1ftz	α	56	71.4	12	1	25.0	58.3	**5.1**	-277	10.1	-270
1c5a	α	66	93.8	43	1	24.4	73.3	**4.2**	-194	9.8	-182
1pou	α	71	84.5	49	0	28.6	89.8	**3.5**	-418	11.9	-364
3icb	α	75	89.3	25	0	28.0	68.0	**4.5**	-406	12.6	-342
t0042	α	*78*	*80.8*	*24*	*1*	*29.1*	*58.3*	*5.6*	*-362*	*11.7*	*-360*
1hmd	α	85	85.0	20	2	10.0	65.0	**4.6**	-458	9.3	460
1shg	β	57	64.9	39	0	28.2	100.	**4.5**	-420	6.7	-397
1fas	β	61	90.2	25	1	26.3	78.9	**6.2**	-330	9.37	-284
6pti	αβ	56	80.4	19	0	68.4	100.	**4.7**	-410	9.7	-397
1cis	αβ	66	86.4	23	0	8.6	78.2	**6.4**	-240	7.6	-232
1lea	αβ	73	87.5	41	2	9.7	75.6	**6.1**	-136	9.4	-115
1ubi	αβ	76	77.6	17	0	23.5	94.1	**6.1**	-238	11.5	-203
1poh	αβ	85	74.1	36	3	8.3	55.5	**6.5**	-336	11.7	-299
1ego	αβ	85	71.8	33	0	15.1	93.9	**5.7**	-417	9.0	-396
1ife	αβ	100	70.0	21	3	14.2	38.0	**6.7**	-419	8.2	-482

[a] Prot refers to the PDB access number.
[b] N is the number of residues in the protein in the PDB file.
[c] Q_3 is the percent of correctly predicted secondary structure. All proteins have a Q_3 within one standard deviation of the average.
[d] N_p is the number of predicted contacts.
[e] N_w is the number of contacts that are incorrect when δ=5.
[f] % of predicted contacts within d residues of a native contact.
[g] rms_n is the cRMSD deviation in Å from the native structure.
[h] E_n is the lowest average energy (in kT) after refinement for the native-like topology.
[i] rms_w is the cRMSD deviation from native in Å of the alternative topology of lowest energy.
[j] E_w is the lowest average energy (in kT) in the alternative topology after refinement runs.
[k] The B domain of protein A[18].
[l] t0042 is target 42 of the CASP2 meeting.

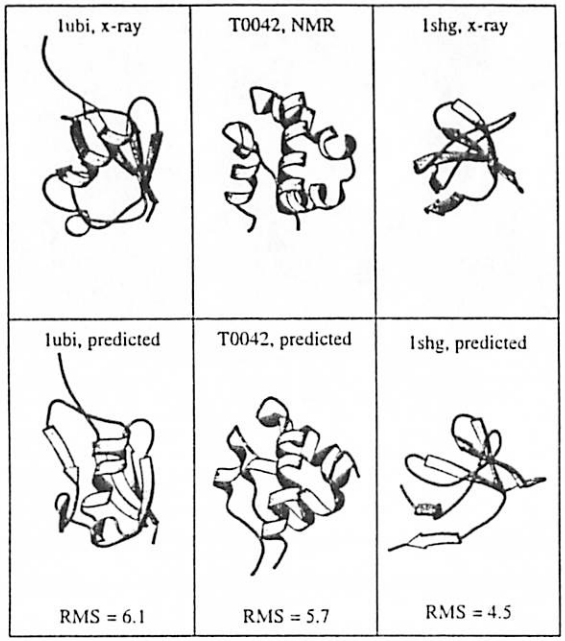

Figure 2. Comparison of the experimentally observed and predicted structure for the α/β protein, 1ubi, the α protein, target 42 of the CASP2 prediction competition, and the β protein 1shg, respectively. The top panel shows the experimental structures, while the lower panel shows the predicted structure. Also shown is the cRMSD in Å of all Cα atoms for the predicted conformation with respect to the native one.

3.2 The case of target 42 from the CASP2 meeting

Recently, the second meeting on the Critical Assessment of Techniques for Protein Structure Prediction (CASP2) was held (see URL: http://iris4.carb.nist.gov/casp2/). A number of protein targets were made available for which researchers could make different types of blind predictions, one category of which was *ab initio* folding. At the time of the CASP2 meeting, the work described here was being carried out and it was premature to attempt blind predictions prior to

gaining adequate experience on how the algorithm behaved. However, once we obtained the requisite experience, we undertook the blind prediction of target 42 (T0042), one of the prediction targets available at CASP2. We chose target 42 because it was the most popular target sequence for the groups participating in *ab initio* folding, and thus, we could compare our method to other approaches. *It must be stressed that the prediction shown in Table 1 and in Figure 2 was made entirely without knowledge of the target structure*, which has been only recently released. Here, we shall describe our prediction results in some detail precisely because they behave in a typical fashion and are illustrative of both the strengths and weaknesses of our approach.

T0042 is a 78-residue protein containing three disulfide bridges whose correct pairing was made available to the prediction teams. This information was incorporated by us into the prediction scheme as well. The experimental structure contains five helices; however, the PHD secondary structure prediction fuses helices III and IV and partially misses helix V. When the PHD predictions are combined with the U-turn predictions, the resulting secondary structure prediction is actually worse than when PHD alone is used. Helices III and IV are still fused, helix V is now totally missed, and helix II is considerably shorter. Nevertheless, the predicted core secondary elements were used to predict seed contacts. These seeds, along with the disulfide bridges, were used in the restraint enrichment procedure. The number of predicted tertiary restraints was 24.

Ten independent folding simulations were performed. Following the isothermal refinement calculations, the lowest average energy fold was selected as the predicted structure. Subsequent to these calculations, the experimental structure was made available. A superposition of the predicted and the experimental structure is shown in Figure 2. The cRMSD is 5.6 Å. A striking feature of the predicted structure is that helix III indeed breaks around residue 55 to give the correct topology. In the real structure, helix IV stretches from residues 57 to 61, as compared to residues 59 to 62 in the predicted conformation. Here, the local secondary prediction (which fuses helices III and IV) is overridden by the tertiary interactions. In contrast, many of the predictions submitted to CASP2 failed to predict the correct topology because they assumed that the secondary structure is absolutely correct[20]. The results presented here

compare favorably with those obtained by the other groups participating in CASP2 (see URL:http://predictioncenter.llnl.gov/). The cRMSD of the predictions ranged from 6.2 Å for Jones to 15.6 Å for Baker. It is interesting that Jones's relatively good results are due to the fact that his algorithm can introduce kinks in the secondary structural elements, which were not considered fixed. Nevertheless, he predicts that T0042 is a four-helix bundle and does not capture the topology of the global fold as the present method does.

The prediction of T0042 also clearly delineates a number of deficiencies in the current approach. For example, using the *PHD/Linker* secondary structure prediction protocol, the C-terminal helix in the experimental structure is incorrectly predicted by us to be an extended state. This helix is only partly recovered in the predicted tertiary structure. This result strengthens the observations made for other folds that if an element of secondary structure is incorrectly predicted, it is very difficult but not impossible, as shown here, for the correct element to form in the final structure. Another crucial problem is related to the discriminative ability of the energy. The difference between the lowest average energy of the native topology and the best alternative fold is only about 2 kT. Analysis of the various contributions to the total energy reveals that when the pair potential alone is considered, the energy difference in favor of native increases to 10-20 kT. In contrast, the restraint energy favors an incorrect topology by about 10 kT. This highlights the necessity of implementing restraints as a relatively soft bias to a manifold of topologies. The restraint energy in and of itself cannot be used to select the native topology.

4. Conclusions

Based on our studies thus far on small proteins, the following conclusions can be drawn. First, the level of accuracy of existing secondary structure prediction schemes, at the secondary structure element level, is adequate for the present approach to tertiary structure prediction to work. Although the test set of proteins used in this work have an average Q_3 value of about 82%, about 10% higher than the average performance of the PHD method, this is because in most of cases all secondary structure elements of the native protein were predicted. However, if an element of secondary structure is entirely missed, depending on its location in the native

conformation, its absence might not necessarily prohibit successful tertiary structure prediction, as heas been demonstrated in the cases of 1gpt and 1ife, for example. Second, low resolution models of small proteins can be assembled from rather inaccurate predictions (about 77% at δ=2, see Table 1) of a subset (25%) of the total number of tertiary contacts in the native protein, as long as the presence of totally wrong contacts (Nw in Table 1) is minimized. Third, helical proteins are predicted with higher accuracy than α/β proteins and β proteins when sequences with the same number of residues are considered. Overall, a promising methodology for the prediction of low resolution tertiary structures of small proteins has been presented, although more studies are required to asses its generality.

Acknowledgments

This work is supported by NIH Grant No. GM-37408. ARO also acknowledges support from the Spanish Ministry of Education. AK also acknowledges support from University of Warsaw Grant BST-34/97.

References

[1]O. B. Ptitsyn, *J. Prot. Chem.* **6**, 273 (1987).

[2]C. B. Anfinsen, *Science* **181**, 223 (1973).

[3]L. Piela, J. Kostrowicki, and H. A. Scheraga, *J. Phys. Chem.* **93**, 3339 (1989).

[4]T. Dandekar and P. Argos, *J. Mol. Biol.* **256**, 645 (1996).

[5]C. Mumenthaler and W. Braun, *Prot. Sci.* **4**, 863 (1995).

[6]M. J. Smith-Brown, D. Kominos, and R. M. Levy, *Prot. Engn.* **6**, 605 (1993).

[7]A. Aszodi, M. J. Gradwell, and W. R. Taylor, *J. Mol. Biol.* **251**, 308 (1995).

[8]J. Skolnick, A. Kolinski, and A. R. Ortiz, *J. Mol. Biol.* **265**, 217 (1997).

[9]A. R. Ortiz, A. Kolinski, and J. Skolnick, *J. Mol. Biol.*, submitted (1997).

[10]C. Sander and R. Schneider, *Proteins* **9**, 56 (1991).

[11]B. Rost and C. Sander, *J. Mol. Biol.* **232**, 584 (1993).

[12]B. Rost, R. Schneider, and C. Sander, *TIBS* **18**, 120 (1993).

[13]A. Kolinski, J. Skolnick, A. Godzik and W-P. Hu, *Proteins* **27**,290 (1997).

[14]U. Gobel, C. Sander, R. Schneider, and A. Valencia, *Proteins* **18**, 309 (1994).

[15]D. J. Thomas, G. Casari, and C. Sander, *Prot. Engn.* **11**, 941 (1996).

[16]O. Olmea and A. Valencia, *Folding & Design*, in press. (1997).

[17]A. Godzik, J. Skolnick, and A. Kolinski, *J. Mol. Biol.* **227**, 227 (1992).

[18]J. Skolnick, http://www.scripps.edu/skolnick (1997).

[19]H. Gouda, H. Torigoe, A. Saito, M. Sato, Y. Arata, and I. Schimada, *Biochemistry* **40**, 9665 (1992).

[20]http://iris4.carb.nist.gov/casp2/, (1997).