PROTEINS: Structure, Function, and Genetics Suppl 3:177–185 (1999)

Ab Initio Folding of Proteins Using Restraints Derived From Evolutionary Information

Angel R. Ortiz,[1] Andrzej Kolinski,[1,2] Piotr Rotkiewicz,[1,2] Bartosz Ilkowski,[1,2] and Jeffrey Skolnick[1]*
[1]*Department of Molecular Biology, The Scripps Research Institute, La Jolla, California*
[2]*Department of Chemistry, University of Warsaw, Warsaw, Poland*

ABSTRACT **We present our predictions in the ab initio structure prediction category of CASP3. Eleven targets were folded, using a method based on a Monte Carlo search driven by secondary and tertiary restraints derived from multiple sequence alignments. Our results can be qualitatively summarized as follows: The global fold can be considered "correct" for targets 65 and 74, "almost correct" for targets 64, 75, and 77, "half-correct" for target 79, and "wrong" for targets 52, 56, 59, and 63. Target 72 has not yet been solved experimentally. On average, for small helical and alpha/beta proteins (on the order of 110 residues or smaller), the method predicted low resolution structures with a reasonably good prediction of the global topology. Most encouraging is that in some situations, such as with target 75 and, particularly, target 77, the method can predict a substantial portion of a rare or even a novel fold. However, the current method still fails on some beta proteins, proteins over the 110-residue threshold, and sequences in which only a poor multiple sequence alignment can be built. On the other hand, for small proteins, the method gives results of quality at least similar to that of threading, with the advantage of not being restricted to known folds in the protein database. Overall, these results indicate that some progress has been made on the ab initio protein folding problem. Detailed information about our results can be obtained by connecting to http://www.bioinformatics.danforthcenter.org/CASP3. Proteins Suppl 1999;3:177–185.** © 1999 Wiley-Liss, Inc.

Key words: ab initio protein folding; tertiary structure prediction; lattice models; contact prediction

INTRODUCTION

The successful prediction of the native structure of a protein requires both the efficient sampling of conformational space and an energy function that recognizes the native conformation as being the lowest in energy. To address the former, we employ a reduced protein model that facilitates sampling in a discretized lattice space,[1–3] together with predicted secondary structure and a small number of tertiary restraints that bias the search over putative nativelike folds. Although such restraints restrict the manifold of possible conformations, by themselves they are insufficient to select nativelike states.[4] Thus, the second issue, native state identification, depends on the use of a knowledge-based force field.[5] The synergism of the predicted secondary and tertiary restraints with the knowledge-based potential yields the native topology among a handful of possible contenders.

For the CASP3 competition, two different lattice-based protein models were used by our group. The most successful approach, used by the *skol-ort-kol* team, employed two interaction centers per residue with the original Cα plus side chain center of mass (CAPLUS) representation of the protein.[5] In the *kolinskol* team, we employed a protein model that uses only one interaction center per residue located at the side chain center of mass (the side chain only model, or SICHO).[6] The CAPLUS model has the advantage of a more accurate backbone description and a side chain description with larger inherent flexibility, but because it uses fewer lattice vectors, its conformational sampling is poorer. As expected, the CAPLUS model worked better for smaller proteins (up to about 100 residues) in which conformational sampling is not a major issue. However, as protein size increases, the better sampling of the SICHO model shows its advantages. In the interest of brevity, we present the results of the CAPLUS model, which in CASP3 produced better results on average.

METHOD

Because the MONSSTER method has been previously described elsewhere,[5,7] here we summarize only its main features. The methodology can be logically divided into secondary and tertiary restraint derivation, topology assembly using the CAPLUS model, and selection of the native structure. We address each issue in turn.

Restraint Derivation

Because our approach has not yet been fully described in the literature, we present the restraint prediction methodology in some detail. Restraints were derived from an analysis of multiple sequence alignments (MSAs). We generated these MSAs mainly by first selecting a pool of homologous sequences after a search in the National Center for Biotechnology Information (NCBI) "nr" database (see http://www.ncbi.nlm.nih.gov/BLAST/

Grant sponsor: National Institutes of Health; Grant number: P41 RR12255.
*Correspondence to: Jeffrey Skolnick, Laboratory of Computational Genomics, Danforth Plant Science Center, 4041 Forest Park Avenue, St. Louis, MO 63108. E-mail: skolnick@danforthcenter.org
Received 28 January 1999; Accepted 21 May 1999

blast_databases.html) with the target sequence using PSI-BLAST,[8] with default parameters and a maximum of three iterations. With the selected sequences, pairwise alignments were carried out. If the percentage of gaps was larger than 22% with respect to the target sequence or the pairwise sequence homology was larger than 95%, then the candidate sequence was discarded. The filtered set of sequences was used to prepare an MSA with the CLUSTAL[9,10] program using default parameters.

Secondary and tertiary restraints were then derived from the MSA. First, the secondary structure was predicted using PHD[11] with the derived sequence alignment, although sometimes the consensus JPRED prediction (http://circinus.ebi.ac.uk:8081) was selected. These predictions were encoded as energetic biases in the folding simulation. Tertiary restraints were then derived by contact map prediction from the MSA, which was carried out in two steps. The first involves the extraction of *seed restraints* from the analysis of residue covariation in the MSA. Using a modified McLachlan mutation matrix, we first calculated a correlation matrix by computing the Pearson correlation coefficient between all pairs of positions in the aligned sequences.[12] Care is required because the dominant contribution to residue covariation comes from phylogenetic relationships. To disentangle this effect from the covariation putatively arising from structural relationships, we applied factor analysis.[13] Here, we exploit the fact that phylogenetic and structural relationships often explain different proportions of the total variance of the correlation matrix and tend to be orthogonal, i.e., they tend to load in different factors. After eliminating phylogenetic relationships, the effects of intervening variables were removed using the partial correlation multivariate technique. The structural character of these seed contacts seems to be sustained by some of our unpublished results that seem to indicate that, for proteins that fold via a nucleation mechanism, these contacts are spatially related to their experimentally known protein folding nucleus in a non-random manner (A.R.O. and J.S.,unpublished data). The resulting set of seed contacts, on the order of 5 to 10 for a 100-residue protein, is too small to generate a native conformation using MONSSTER; thus, they were *expanded* using a local threading method. This method uses the predicted secondary structures and contacts, and searches and scores, using a protein database, pairs of packed secondary structure elements consistent with these restraints. If the search finds patterns that satisfy certain thresholds in the score and conformational diversity of the fragments, the contacts from the packed elements are extracted and incorporated into the target contact map, using the predicted contacts as a reference.

Fold Generation, Topology Selection, and Addition of Atomic Details

The predicted contacts were used as restraints to drive the conformational search in the folding algorithm. For each target, 1,000 independent Monte Carlo folding simulations were undertaken using simulated annealing, starting from a totally extended conformation. For each simulation, the lowest energy conformation found during the trajectory was stored. The 1,000 energies were then sorted, and the lowest energy model was submitted as model 1. The reduced models were converted into all-atom models using MODELLER.[14]

RESULTS

We tried to derive an MSA for all available CASP3 ab initio targets smaller than 140 residues. If the resulting MSA had a sufficient number of sequences, we attempted to derive a model. This procedure resulted in a set of 11 prediction targets (T0052, T0056, T0059, T0063, T0064, T0065, T0072, T0074, T0075, T0077, T0079). A more complete description of the targets is available at http://PredictionCenter.llnv.gov/casp3/SUMMARY/cm1/, as well as at http://bioinformatics.danforthcenter.org/casp3. T0072 is not included in the subsequent analysis because its structure has not yet been solved.

Contact Predictions

Table I reports the percentage of correctly predicted contacts within $|\delta|$ residues, measured along the sequence, of the experimental contact. A contact is assigned if the minimal distance between heavy atoms belonging to the side chains of two residues i and j (separated along the chain so that $|j\text{-}i| \geq 2$) is < 4.5 Å. When $|\delta| > 0$, this measures the *precision* of the contact prediction. By precision, we mean the error in contact map space of a predicted contact compared with the observed contact, as evaluated by $|\delta|$. In our experience, the *precision* and not the *accuracy* (i.e., the average value at $|\delta|=0$) is the relevant quantity in evaluating the utility of a given contact prediction procedure for tertiary structure prediction. This is so because with only one or a small number of very dissimilar contacts (contacts situated far away in the contact map space) it is possible to favor an alternative fold over the correct answer, even if the majority of the contacts are given exactly and the *accuracy* of the contact prediction is high. On the other hand, a small shift in contact map space in all predicted contacts with respect to the real contacts only very slightly affects, on the order of 1 to 2 Å, the reconstructed three-dimensional model. Thus, it is possible to create a model with a root-mean-square deviation (RMSD) of 1 to 2 Å from the real structure by shifting all real contacts by one residue, even though the *accuracy* of this new set of contacts is zero (i.e., there are no predicted contacts that correspond *exactly* to any experimental contact).

Four representative values of $|\delta|$ (=0, 2, 4, and 5) are shown in Table I, which also presents the values expected by chance. On average, our tertiary restraint derivation protocol is more *precise* than a random prediction (a higher percentage of predicted contacts are predicted within a given precision level), but it is also more *accurate* (as can be seen by observing the improvement over a random prediction at $|\delta|=0$). The results for target 56 (Fig.1 A and B and Table I) are worth noting because they point out the importance of a good MSA. Using the alignment generated by the procedure described in Methods, the percentage of

TABLE I. Precision of Predicted Contacts Obtained From the Contact Prediction Method

Target	Enrich[a] ($\|\delta\| = 0$)	% Correct[b] ($\|\delta\| = 0$)	Enrich[a] ($\|\delta\| = 2$)	% Correct[b] ($\|\delta\| = 2$)	Enrich[a] ($\|\delta\| = 3$)	% Correct[b] ($\|\delta\| = 3$)	Enrich[a] ($\|\delta\| = 5$)	% Correct[b] ($\|\delta\| = 5$)
T0056	0.000	0.000	0.854	11.765	1.154	27.451	1.263	37.255
T0056[c]	11.643	13.636	3.200	43.939	2.744	65.152	2.521	74.242
T0059	8.621	14.286	3.344	57.143	2.077	71.429	2.373	100.000
T0064	4.687	3.093	4.491	44.330	3.840	65.979	3.486	75.258
T0065	9.434	5.882	2.900	52.941	2.102	70.588	1.570	70.588
T0074	12.136	13.514	4.652	72.973	3.535	97.297	2.854	97.297
T0075	0.000	0.000	0.961	13.333	1.931	50.000	1.858	60.000
T0077	2.538	2.326	2.744	39.535	1.862	51.163	1.786	60.465
T0079	15.355	15.686	4.387	58.824	3.150	72.549	2.721	76.471

[a]Contact prediction enrichment with respect to a random prediction. For non-zero values, it is computed as the ratio of the percentage of predicted contacts at each precision level with respect to the average value of random predictions (after 1000 randomizations), given the same number of predicted contacts for that structure. Results are presented at the different precision levels. Note that, because of the binning of the precision levels, this ratio converges to 1 (i.e., saturates) for $|\delta| > 5$; thus, the value carries more information content at low $|\delta|$ values.
[b]Percent of contacts correctly predicted within $|\delta|$ residues of a contact in the native structure.
[c]Predicted contacts generated from an alternative multiple sequence alignment kindly provided by H. Scheraga and coworkers, Cornell University, Ithaca, NY (personal communication).

correct contacts at a precision of $|\delta|=4$ is too small for correct global fold prediction. In fact, the prediction is not very different from a random prediction. However, an alternative MSA provided by Dr. Scheraga et al. (entry labeled as T0056* in Table I) yields a set of predicted contacts of sufficient quality to generate a nativelike fold based on previous benchmarks. The reason for these differences at the alignment level remains unknown to us.

Structure Predictions

The quality of the predicted models was assessed with the DALI program[15] to generate the best sequence-independent structural superimposition between the experimental conformation and the predicted conformation. As shown in Table II, for 50% of the predicted targets, the best matches with the experimental structure constitute a substantial fraction (more than 50% of the structure) of the entire fold and have coordinate RMSDs from experiment below 4 Å. In addition, columns 6 and 7 report the RMSD of the superimposed secondary structural elements and the percent of the sequence composed of such regular secondary structure elements. Overall, our results are as follows: the global fold can be considered "correct" for targets 65 and 74; "almost correct" for targets 64, 75, and 77; "half-correct" for target 79, and "wrong" for targets 52, 56, 59, and 63. Next, we briefly dissect the results of the 10 sequences for which experimental structures are available.

1. T0052 is a beta protein with a novel fold. Although our prediction is the best of all submissions according to the SIPPL score (Table III), the predicted global fold is incorrect. This is due partly to the poor MSA quality (the selected sequences are at best weakly related to the target) and partly to the incorrect prediction of a helix in an alanine-rich region of the molecule.
2. T0056 is an alpha protein also with a novel fold. Here, the prediction also fails. Based on Table I, it is apparent that the quality of the predicted contacts is poor. However, the use of an alternative MSA provided to us by Scheraga and coworkers (T0056* in Table I) produced a good set of predicted contacts, illustrating the effect of the alignment on the contact prediction quality (Fig. 1A and B).
3. T0059 is beta protein. While the quality of the predicted tertiary contacts is acceptable, there are too few for the native topology to be successfully assembled. Yet, according to the SIPPL score (Table III), it is one of the best models submitted.
4. T0063 is a large, 138-residue beta protein. Given its size and secondary structure content, our benchmarks indicated that the likelihood of a good prediction was low. However, we wanted to explore the limits of our method. This target has two domains in the real structure, but it was modeled as a single domain.
5. T0064 is a two-domain, alpha helical protein. Both domains are correctly predicted, but their relative orientation is incorrect. This is partly due to a lattice artifact. Lattice models have great difficulty translating preassembled elements of structure. The incorrect orientation, however, also has to do to with the fact that the number of interdomain contacts in real structures is small, and the contact prediction method is not efficient in predicting them, perhaps because of the physical origin of the covariation in an MSA.
6. T0065 is a small helical hairpin protein. Good results are obtained with a best DALI structural superposition of 2.05 Å over 81% of the sequence (Table II). According to DALI, this prediction ranks fourth of models submitted by all groups (Table III).
7. T0074 is another helical protein. Because the MSA exhibited a large number of deletions in the C-terminal region, it was not possible to derive restraints for that part of the sequence. Thus, we attempted to predict only the tertiary structure of the four-helix bundle, calcium-binding domain. For model 1, 75% of the structure has an RMSD of 3.2 Å, with an RMSD of the secondary structure elements of 3.8 Å.

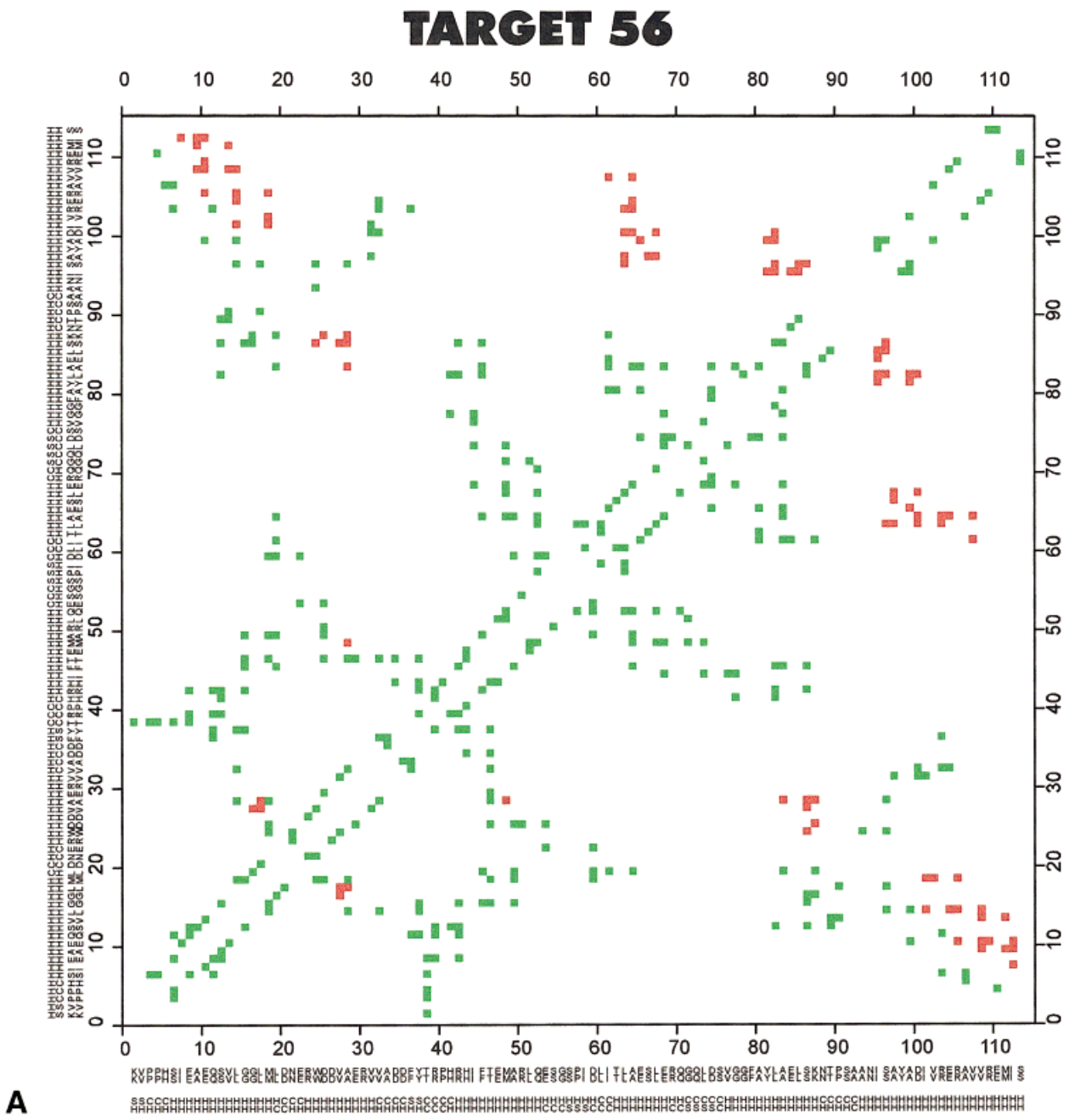

Fig. 1. An example of the comparison of experimental (green) and predicted (red) contacts from an MSA. Black dots correspond to the points for which experimental and predicted contacts exactly agree. Contact maps are shown for targets 56 (**A**), 56*(**B**) (see Table II), and 77 (**C**). Comparison of contact maps for all targets is available at http://bioinformatics.danforthcenter.org/casp3.

Although close structures of calcium-binding domains were available in the PDB, and these were picked by the threading programs, our ab initio prediction was the best of all submissions as ranked by DALI (Table III).

8. T0075 is an alpha helical protein. Its fold is very rarely observed in known proteins (Fig. 2A). The predicted model can be aligned quite closely to the experimental conformation over about half the structure (3.7 Å for 54.5% of the structure). The N- and C-terminal halves of the molecule are correctly predicted, and the assigned fold class is correct. However, one internal helix is incorrectly flipped, giving a poor global RMSD (Table II). Our model is the fourth best model produced by all groups according to DALI (Table III).
9. T0077 is a 108-residue mixed alpha/beta protein with a complex topology that, depending on the author, can be classified as a novel fold. The prediction was

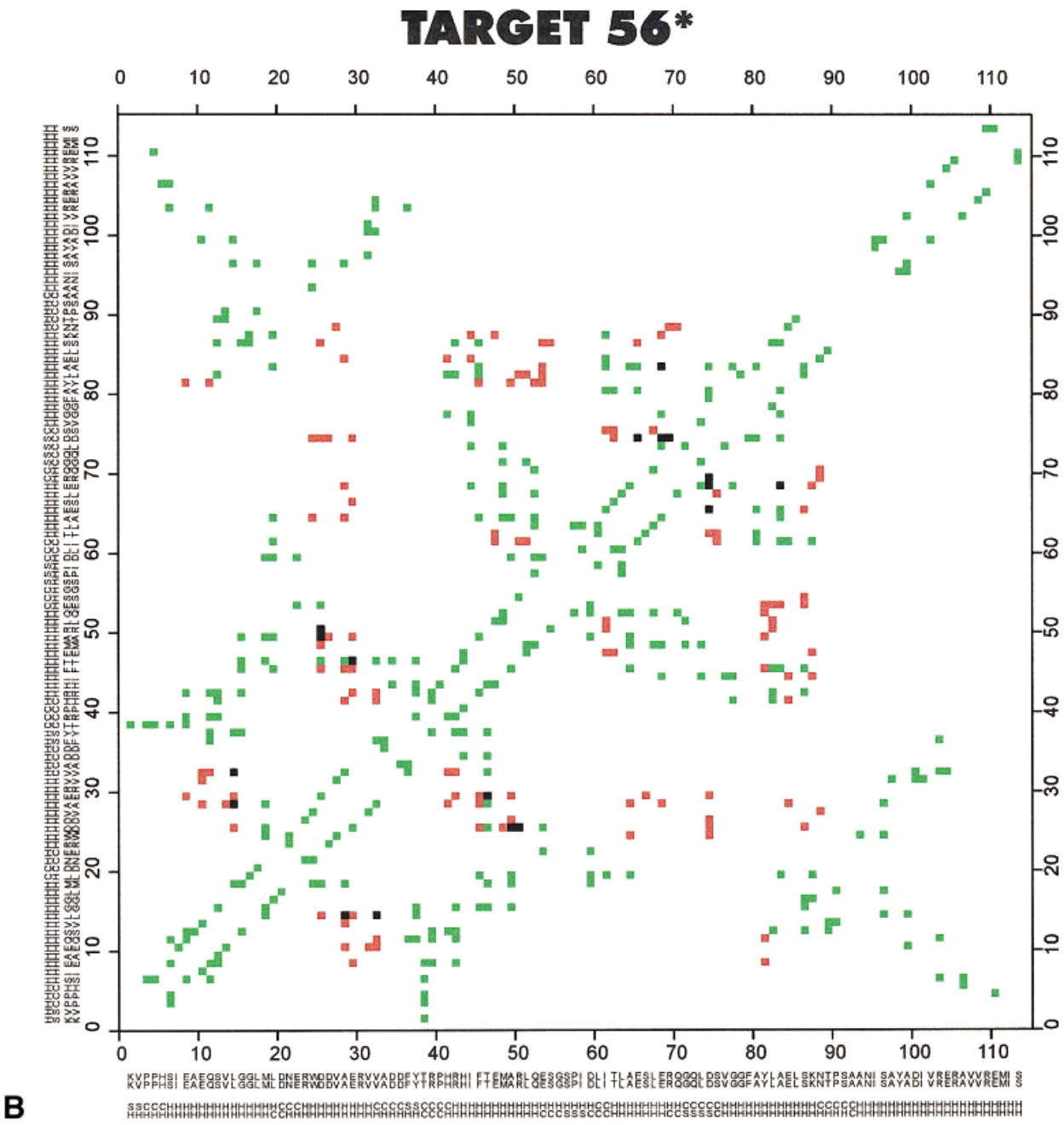

Figure 1B. (Continued.)

essentially correct; however, as shown in Figure 2B, the two central beta strands are swapped, an effect that may be partly the result of kinetic trapping, although the contact map prediction (Fig. 1C) misplaces the packing of these two strands. Otherwise, the fold is quite well reproduced, with an RMSD of 3.1 Å over 52% of the structure and an RMSD of 7.3 Å for all elements of secondary structure (Table II). According to the DALI score (Table III), ours was the best global prediction for this structure.

10. T0079 is an alpha helical DNA-binding protein of 116 residues with a novel arrangement of the relative position of two helix-turn-helix motifs. The two motifs are separated into two domains joined by a helical stalk, giving the molecule a very elongated shape. About 40% of the structure can be aligned with an RMSD of ≈4 Å. The structure of the first lobe and the helical stalk are reasonably well predicted. However, the second domain is docked onto the first, thereby incorrectly creating a spherical model. This may be in

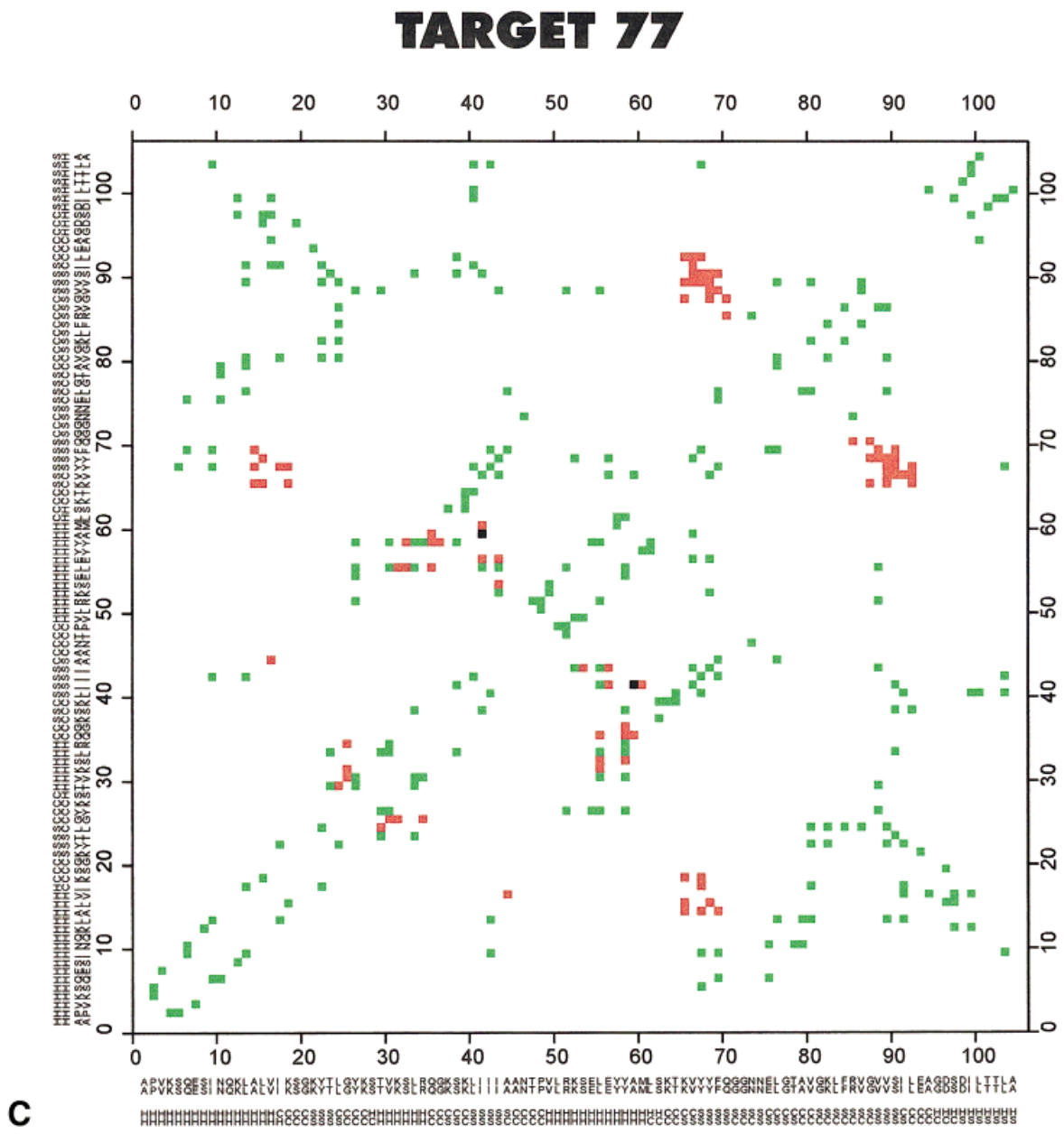

Figure 1. (Continued.)

part because the algorithm assumes that the protein is a single domain.

A word of caution, however, is required when analyzing these results in terms of structural superimpositions because different metrics can provide different answers, and no ultimate "standard of truth" exists. In our analysis we have used two of the most accepted methods in CASP3, namely, the DALI score and the SIPPL score. However, as can be inferred from inspection of Table III, little correlation exists between the rankings provided by each of the methods. This is not surprising because each method emphasizes a different aspect of structural similarity. For this reason, we have chosen to present our comparative results showing the performance of both methods in our set of predictions (Table III), noting in the text the metric we use in judging a particular prediction.

CONCLUSIONS

What Went Right?

For small helical and alpha/beta proteins on the order of 110 residues or smaller, the MONSSTER algorithm has predicted low-resolution structures with a reasonably good prediction of the global topology. For five models (targets 64, 65, 74, 75, and 77) out of the 10 predictions, it is possible to find a structural alignment with an RMSD

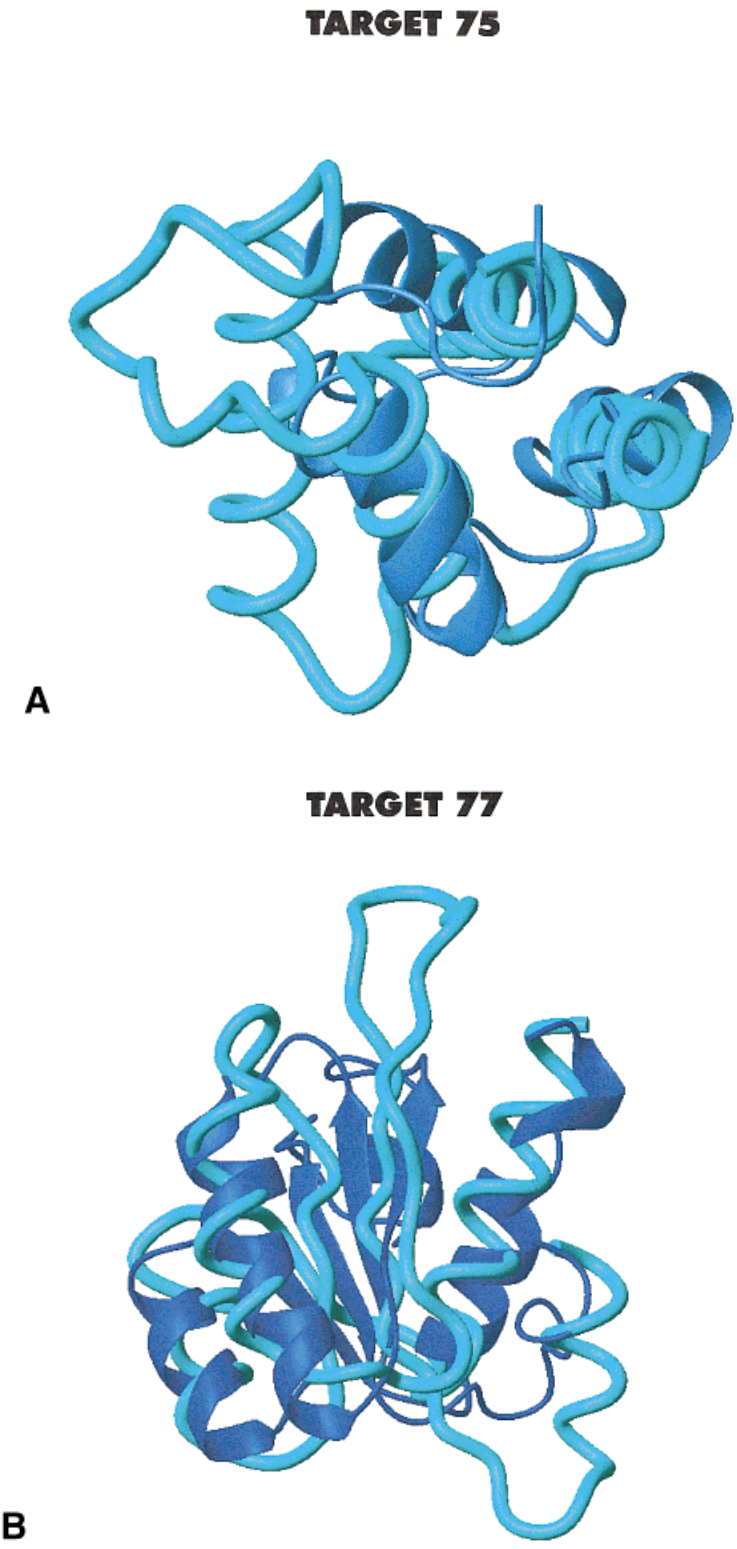

Fig. 2. Best structural superposition of the native (blue) and predicted (cyan) structures generated by DALI. Only the portion corresponding to the DALI structural alignment is shown for each of the structures. Targets 75(**A**) and 77(**B**) are displayed. Comparisons for the rest of the targets are available at http://bioinformatics.danforthcenter.org/casp3. Ab initio folding of proteins using evolution based restraints.

TABLE II. Assessment of the Predicted Structures by Their rmsd from Native Using the DALI Program, Together with the rmsd based on the Best Coordinate Structural Superposition of the Secondary Structure Elements

Target	Model[a]	ALI[b]	rmsd (DALI)[c]	% ALI[d]	rmsd (SSE)[e]	% ALI[d]
T0056	1	114	—	—	12.947	85.088
T0059	1	71	—	—	9.282	45.070
T0059	2	71	—	—	11.705	54.930
T0059	3	71	—	—	10.384	54.930
T0064	1	103	2.978	58.252	11.602	77.670
T0064	2	103	4.636	62.136	11.936	77.670
T0065	1	31	2.047	80.645	5.435	90.323
T0065	2	31	2.583	90.323	5.050	90.323
T0065	3	31	1.799	80.645	5.182	90.323
T0065	4	31	1.795	74.194	5.209	90.323
T0065	5	31	—	—	4.795	90.323
T0074	1	95	3.256	72.632	3.878	65.263
T0074	2	95	3.380	61.053	4.455	65.263
T0074	3	95	4.199	58.947	5.279	65.263
T0074	4	95	9.764	49.474	10.705	65.263
T0074	5	95	—	—	10.178	65.263
T0075	1	88	3.770	54.545	12.597	79.545
T0075	2	88	3.329	56.818	12.514	79.545
T0075	3	88	3.770	54.545	12.597	79.545
T0075	4	88	—	—	12.380	79.545
T0077	1	104	3.096	51.923	7.312	64.423
T0077	2	104	3.475	60.577	6.729	64.423
T0077	3	104	—	—	10.099	64.423
T0077	4	104	4.592	66.346	7.940	57.692
T0077	5	104	2.964	56.731	6.988	57.692
T0079	1	116	4.129	39.655	13.608	80.172

[a]Refers to the description of the model number, labeling each one of our CASP3 submissions.
[b]Number of aligned residues between the model and the experimental structure.
[c]rmsd of the best structural alignment produced by DALI between the prediction and experiment. Failures of DALI to find a significant alignment are indicated by a dash.
[d]Percentage of the total structure that is considered in the rmsd measure.
[e]rmsd between the secondary structure elements of the experimental and predicted structure. The secondary structure was measured in the experimental structure.

TABLE III. Ranking of the Predicted Structures Submitted as "Model 1" Compared with all "Model 1" Structures Submitted by Different Prediction Teams[†]

Target ID	# Models in CASP3	DALI rank*	% DALI rank	SIPPL rank*	% SIPPL rank
T0052	32	9	28.12	1	3.12
T0056	33	18	54.54	13	39.39
T0059	47	32	68.08	5	10.63
T0063	31	21	67.74	9	29.03
T0064	61	34	55.73	36	59.01
T0065	32	4	12.50	18	56.25
T0074	49	1	2.04	27	55.10
T0075	35	4	11.42	11	31.42
T0077	38	1	2.63	4	10.52
T0079	50	15	30.00	8	16.00

[†]Two different scoring schemes were used: DALI and the SIPPL scoring method, based on the percentage of residues correctly aligned within four residues of the experimental structure (see *http://PredictionCenter.llnv.gov/casp3/SUMMARY/cm1/* for a listing of all predictions according to each one of the scoring systems. The table was elaborated by analyzing these listings).

below ≈ 4 Å over more than 50% of the structure (Table II). For the other five cases, for which the method clearly fails, a portion of the native topology is recovered in some cases. In fact, MONSSTER is one of the methods, among all methods presented in CASP3, that allows a larger recovery of topology in these cases. For three (targets 52, 59, and 79) of these five targets, the SIPPL score ranks the MONSSTER structures in the upper 15% when considering all "model 1" structures. Note that these cases correspond to novel folds. Perhaps the most encouraging result is that in some situations, such as for target 75 and particularly target 77 (Fig. 2), the method is able to generate a rare or novel fold. This clearly points out the power of a method that does not rely on the availability of fold examples in a structural database.

Three features are thought to be responsible for the measure of success achieved. First, the approach works at the current level of accuracy of contemporary secondary structure prediction schemes. Only the identity and approximate location of most of the secondary structural elements are needed, but their exact location is not. Second, we employ a tertiary contact prediction algorithm that provides a set of contacts of sufficient precision and number so that MONSSTER can successfully assemble a fold. Moreover, these restraints are implemented in a manner that accommodates the expected contact precision; thus, the predicted contacts need to be precise enough only to spatially localize the contacting secondary structural elements. Third, the knowledge-based components of the potential can in many cases pick a nativelike solution from the handful of topologies allowed by the restraints.

What Went Wrong?

From the CASP3 results, it is apparent that the method fails in the treatment of beta proteins, proteins over the 110-residue threshold, and sequences for which only a poor MSA can be built. Part of the problem with beta proteins resides in our definition of hydrogen bonds, which allows for too much freedom in the mutual orientation of the strands. Another problem apparent in all beta proteins and in all larger proteins is conformational sampling. For target 64, for example, examination of the folding trajectories indicates kinetic trapping.

A third problem that limits the current approach to proteins below the threshold of 110 or so residues is the assumption that the protein contains just a single domain. Different treatments of side chain burial could perhaps partly avoid this problem. However, another reason for the poor performance in multidomain proteins is the general difficulty the contact prediction method has when applied to domain-domain interactions. This is partly because of the small fraction of contacts involving domain-domain

interactions as well as the observation that such interactions seem to impose too weak a constraint in sequence space to be reliably detected from the MSA.

Another source of problems is the MSA itself. The impact of the MSA on the quality of the predicted contacts can be great, as illustrated for target 56. However, this is a totally unexplored subject and, at present, we cannot easily distinguish a "good" from a "bad" alignment from the perspective of contact prediction. More work is required to clarify this issue and to improve the quality of the contact prediction.

Final Observations

The results presented here indicate that some progress has been made on the ab initio protein folding problem. Overall, our CASP3 results are about what we qualitatively expected based on our benchmarks[4,16] and previous blind predictions.[17,18] For small proteins, as indicated by the analysis given in Table III, the method gives results of quality at least similar to that of threading. On average, and independently of the evaluation method (DALI scoring or SIPPL scoring), our "model 1" predictions for the 10 sequences rank within the upper ≈30% of all submitted models marked as "model 1". Given that most models in CASP3 were created using a threading method, one would expect an average rank of ≈50% if the method performs as well as the average threading approach. However, the relative performance is ≈40% better. Even more encouraging is the fact that 40% of the models rank in the upper 20% by each scoring system, and when both scoring systems are considered together, 70% of the models rank in the upper 20%. Moreover, in 40% of the cases (for targets 65, 74, 75, and 77), the models produced are among the best four of all models produced by all predictor teams, as scored by DALI. Thus, while improvements in this method are clearly required, the results to date are encouraging.

ACKNOWLEDGMENTS

A.R.O. receives partial support from the Spanish Ministry of Education. A.K. is an International Scholar of the Howard Hughes Medical Institute. This research was supported in part by NIH grant P41-RR1225.

REFERENCES

1. Kolinski A, Skolnick J. Lattice models of protein folding, dynamics and thermodynamics. Austin, TX: R.G. Landes; 1996. 200 p.
2. Skolnick J, Kolinski A. Monte Carlo lattice dynamics and the prediction of protein folds. In: van Gunsteren WF, Weiner PK, Wilkinson AJ, editors. Computer simulations of biomolecular systems. Theoretical and experimental studies. Leiden, The Netherlands: ESCOM Science Publishing; 1997. p 395–429.
3. Godzik A, Kolinski A, Skolnick J. Lattice representation of globular proteins: How good are they? J Comp Chem 1993;14:1194–1202.
4. Ortiz AR, Kolinski A, Skolnick J. Nativelike topology assembly of small proteins using predicted restraints in Monte Carlo folding simulations. Proc Natl Acad Sci USA 1998;95:1020–1025.
5. Skolnick J, Kolinski A, Ortiz AR. MONSSTER: a method for folding globular proteins with a small number of distance restraints. J Mol Biol 1997;265:217–241.
6. Kolinski A, Skolnick J. Assembly of protein structure from sparse experimental data: an efficient Monte Carlo model. Proteins 1998;32:475–494.
7. Ortiz AR, Kolinski A, Skolnick J. Fold assembly of small proteins using Monte Carlo simulations driven by restraints derived from multiple sequence alignments. J Mol Biol 1998;277:419–448.
8. Altschul SF, Madden TL, Schaffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
9. Higgins DG, Sharp PM. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene 1988;73: 237– 244.
10. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994;22:4673– 4680.
11. Rost B, Sander C. Prediction of secondary structure at better than 70% accuracy. J Mol Biol 1993;232:584–599.
12. Goebel U, Sander C, Schneider R, Valencia A. Correlated mutations and residue contacts in proteins. Proteins 1994;18:309–317.
13. Johnson RA, Wichern DW. Applied multivariate statistical analysis, 3rd ed. Upper Saddler River, NJ: Prentice Hall; 1992. 592 p.
14. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 1993;234:779–815.
15. Holm L, Sander C. Protein structure comparison by alignment of distance matrices. J Mol Biol 1993;233:123–138.
16. Simmerling C, Lee M, Ortiz AR, Kolinski A, Skolnick J, Kollman PA. Combining MONSSTER and LES/PME to predict protein structure from amino acid sequence: application to the small protein CMTI-1. J Am Chem Soc 1999; in press.
17. Ortiz AR, Kolinski A, Skolnick J. Tertiary structure prediction of the KIX domain of CBP using Monte Carlo simulations driven by restraints derived from multiple sequence alignments. Proteins 1998;30:287– 294.
18. Skolnick J, Kolinski A, Ortiz AR. Reduced protein models and their application to the protein folding problem. J Biomol Struct Dyn 1998;16:381– 396.
19. Bernstein FC, Koetzle TF, Williams GJ, et al. The Protein Data Bank: a computer-based archival file for macromolecular structures. J Mol Biol 1977;112:535– 542.