PROTEINS: Structure, Function, and Genetics 53:469–479 (2003)

TOUCHSTONE: A Unified Approach to Protein Structure Prediction

Jeffrey Skolnick,[1]* Yang Zhang,[1] Adrian K. Arakaki,[1] Andrzej Kolinski,[1,2] Michal Boniecki,[2] András Szilágyi,[1] and Daisuke Kihara[1]

[1]*Center of Excellence in Bioinformatics, University at Buffalo, Buffalo, New York*
[2]*Faculty of Chemistry, University of Warsaw, Warsaw, Poland*

ABSTRACT **We have applied the *TOUCHSTONE* structure prediction algorithm that spans the range from homology modeling to *ab initio* folding to all protein targets in CASP5. Using our threading algorithm *PROSPECTOR* that does not utilize input from metaservers, one threads against a representative set of PDB templates. If a template is significantly hit, Generalized Comparative Modeling designed to span the range from closely to distantly related proteins from the template is done. This involves freezing the aligned regions and relaxing the remaining structure to accommodate insertions or deletions with respect to the template. For all targets, consensus predicted side chain contacts from at least weakly threading templates are pooled and incorporated into *ab initio* folding. Often, *TOUCHSTONE* performs well in the CM to FR categories, with *PROSPECTOR* showing significant ability to identify analogous templates. When *ab initio* folding is done, frequently the best models are closer to the native state than the initial template. Among the particularly good predictions are T0130 in the CM/FR category, T0138 in the FR(H) category, T0135 in the FR(A) category, T0170 in the FR/NF category and T0181 in the NF category. Improvements in the approach are needed in the FR/NF and NF categories. Nevertheless, *TOUCHSTONE* was one of the best performing algorithms over all categories in CASP5. Proteins 2003;53:469–479.**

Key words: comparative modeling; fold recognition; ab initio prediction

INTRODUCTION

As typified by the CASP5 target categories, there are three basic protein structure prediction approaches: comparative modeling, CM, threading or fold recognition, FR, and *ab initio* folding or New Folds, NF. In CM, the target sequence has a clear evolutionary relationship to another protein (template) whose structure has already been solved.[1] Here, the goal is to predict structures with a root mean square deviation, RMSD, of 1-2 Å from native. The next more difficult protein structure prediction method is fold recognition,[2-8] where one attempts to find the closest matching structure in a library of solved structures. The challenge is to recognize not only homologous but also analogous proteins where the target and template proteins are not necessarily evolutionarily related, but adopt very similar structures perhaps by convergent evolution.[9] Both threading and comparative modeling suffer from the limitation that structure of the homologous or analogous protein must have been solved. To address this, the most difficult and general approach is *ab initio* folding, where one attempts to fold a protein from a random conformation.[10-13] As expected, models in the FR/NF categories tend to be of lower resolution. To be successful, a structure prediction method must be able to span the range from the CM to NF categories. In this paper, we describe our automated ***TOUCHSTONE***[14] protein structure prediction approach that has partially achieved this goal. Furthermore, to establish its applicability, large scale benchmarking is required. CASP5 partly addresses this need as the predictions are blind, but it is also important to know beforehand how the algorithm should behave. If similar results as under the stress of truly blind predictions are found, the validation protocols can be used with greater confidence to improve the methods. Finally, a key goal of our participation in CASP5 was to identify the strengths and weaknesses of the current ***TOUCHSTONE*** algorithm.

METHOD

An overview of the ***TOUCHSTONE*** methodology is shown in Figure 1(A). First, one uses ***PSIPRED***,[15] to predict secondary structure that is fed into our threading algorithm ***PROSPECTOR***.[8] ***PROSPECTOR*** provides predicted side chain contacts, sets of local distances and when applicable, a predicted template [see Fig. 1(B), for a more detailed discussion see Section a. below]. If there is at least one predicted template, then Generalized Comparative Modeling (see Section b below) in the vicinity of the template is done. In all cases, *ab initio* folding is done using predicted contacts and secondary structure in the context of a new lattice protein model, CABS (Zhang et al, 2003 submitted). All structures are clustered using SCAR.[16] On the basis of cluster population and related parameters,

Grant sponsor: Division of General Medical Sciences of the National Institutes of Health; Grant numbers: GM-48835 and GM-37408.

*Correspondence to: Jeffrey Skolnick, Center of Excellence in Bioinformatics, University of Buffalo, 901 Washington St., Buffalo, NY 14203. E-mail: skolnick@buffalo.edu

Received 28 January 2003; Accepted 20 May 2003

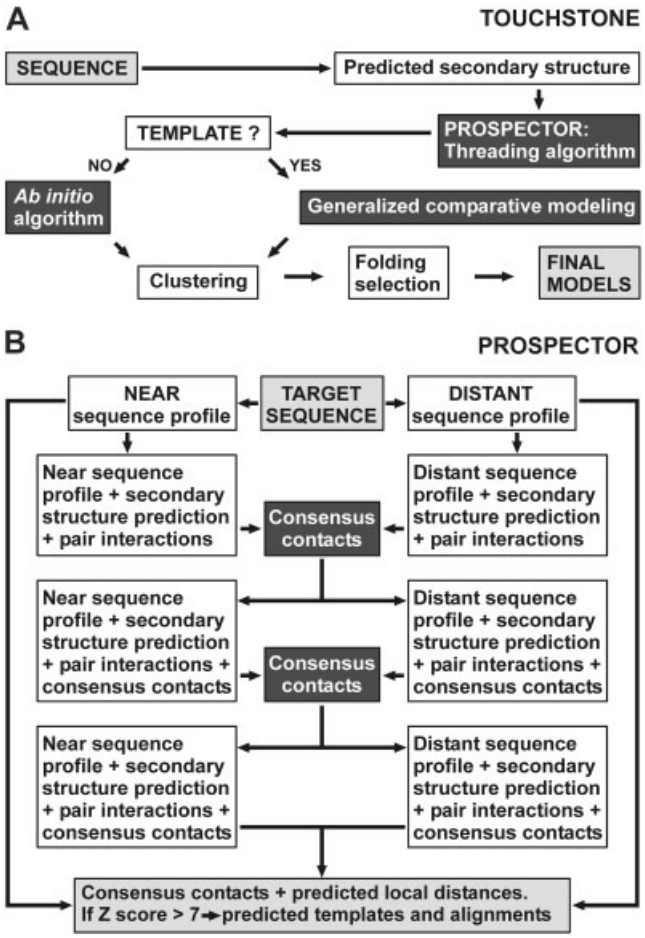

Fig. 1. The flowcharts show the ***TOUCHSTONE*** methodology (**A**) and the ***PROSPECTOR*** algorithm (**B**). For details, see METHODS.

fold selection is done and the top five models submitted to CASP (see Section d).

Overview of *PROSPECTOR*

As shown in Figure 1(B), ***PROSPECTOR*** is an iterative threading algorithm that combines near (between 35%-90% sequence identity) and distant sequence profiles (sequences from ***FASTA***[17] with an E-value <10) to generate alignments from which the partners for the evaluation of pair interactions are extracted and used in subsequent threading iterations. For each of the four scoring functions (near/distant sequence profiles, and the corresponding set of near/distant sequence profiles plus predicted secondary structure plus pair interactions), the top 5 scoring structures are selected. If a contact is present in at least 5/20 alignments, this constitutes a predicted contact that is converted into a pair potential for use in subsequent threading. Contacts are predicted for a total of three iterations and pooled for use in subsequent Generalized Comparative Modeling and *ab initio* folding. ***PROSPECTOR*** also provides a set of local distance predictions (between residues separated by no more that 5 residues). If the Z-score of a predicted template is >7, then the target sequence is assigned as having the structure of the template. No metaservers are used in ***PROSPECTOR;*** apart from predicted secondary structure and sequence profiles generated by ***FASTA***,[17] it is stand-alone.

Generalized Comparative Modeling

In Generalized Comparative Modeling, depending on the sequence identity between the target and template, different protein models were used. When the sequence identity of was $<40\%$, we used an algorithm that could improve the model while retaining the accuracy of the well aligned regions. Thus, using the CABS protein model [for additional details see Section c and (Zhang et al, 2003 submitted)], we kept the well-aligned regions fixed and built an initial model from the ***PROSPECTOR*** identified template. When a gap in the template is physically connectible, we perform a random walk that connects one extremity to the other. When the specified number of unaligned residues cannot close the gap, we release the neighboring aligned residues until the gap can be spanned; then, a random walk procedure builds the initial model. Residues having identical spatial coordinates in the initially built model as the template are frozen and the rest are moveable, but since the two residues at the end of the aligned regions are often unreliably predicted, these are also moveable. The conformational space of the movable regions is explored using our Parallel Hyperbolic Sampling (PHS) Monte Carlo approach,[18] with the energy calculated for the whole chain. Depending on the number of the movable residues, we use 20$\sim$30 replicas and submit the structures from the 8 lowest temperature replicas to SCAR clustering and choose the final clusters according to cluster density, the similarity to the template, and the combination of energy and free energy (see Section d below).

When the sequence identity of the target to the template is $>40\%$, a continuous space C_α- model with side groups represented by 1-3 united atoms was used to avoid loss of resolution due to lattice discretization. Short and long-range interactions were controlled by a statistical potential, similar to the CABS model. Conformational space was searched by Replica Exchange Monte Carlo.[19,20] Small insertions or deletions were allowed in the alignment provided by ***PROSPECTOR***, with the template and target always very close. A similar gap-closing algorithm was used as above. Usually, after short simulations, conformations very close to the target protein's native structure are obtained.

Ab Initio Folding

The CABS model describes the protein as a set of Cαs, Cβs and where appropriate the center of mass of the remaining side chain heavy atoms (Zhang et al, 2003 submitted). The force field consists of a set of short-range interaction terms describing local, protein like conformational stiffness, hydrogen bonds, predicted secondary structure from ***PSIPRED***,[15] and threading based local distance restraints. The long-range potential is comprised of terms describing side chain burial and contact environment, pair

interactions, Debye-Hückel electrostatic interactions, a bias towards predicted contact order and number, and threading based side chain restraints whose accuracy strongly biases the likelihood of success. To allow for template refinement, *ab initio* folding is done over all non-CM targets, and the resulting best cluster is included as one of the choices.

Depending on protein length, from 30-80 replicas at temperatures that span the conformational transition region are used in PHS, with different replicas starting from different structures. Half of the initial structures are random and half start from structures selected from the PDB using gapless threading. At the simulation end, structures from the 12 lowest temperature replicas are subjected to SCAR and subsequent structure selection.

Best Cluster Selection

A number of quantities were previously used to select the best cluster obtained from the clustering program SCAR. In a benchmark test,[21] we tried to rank the clusters based on their energy, E, or cluster population, M. However, these were not optimal. One of the quantities, which turned out more sensitive to the correctness of structures, is a combination of the energy and free energy, Y,

$$Y = E - kT\log(M)\,. \tag{1}$$

A more sensitive quantity found in the benchmark test is the ratio of the structure population in a cluster to the average RMSD of the structures to the cluster centroid, i.e.,

$$D = {}^{M}/_{M_{\rm tot}}\langle RMSD\rangle_{cluster} \tag{2}$$

where M_{tot} is the total number of structures submitted to the cluster processes. The normalization by M_{tot} is necessary when we compare the clusters from multiple simulations using different sets of restraints. The idea of the quality D is that the cluster density reflects the coordination between the threading-based restraints and general force field of CABS model. The higher the cluster density is and the more convergent the trajectory is, the more likely the cluster is to be correct.

For target selection in CASP5, we first sort the simulations of different restraint sets according to D; and then select and submit the structures according to Y for a given set of clusters.

BENCHMARK TESTING

Prior to CASP5, we undertook extensive benchmark calculations. For ***PROSPECTOR,*** we considered a representative set of all solved protein structures (<35% identity) 200 residues or shorter in length. At the time, there were 2186 protein domains. These are a subset of the representative PDB library of 3990 proteins no pair of which has greater than 35% sequence identity and excludes coiled coils and chains from multimers that are very open. Of the 2186 proteins, 1317 (60%) are matched to templates with an average sequence identity of 25.5%. Thus, ***PROSPECTOR*** can identify weakly evolutionarily related proteins. There are 899 (68%) proteins with a coordinate RMSD <6.5 Å over the aligned region (this threshold was chosen because based on the relative RRMSD defined previously,[22] statistically significant structures are predicted). The average RMSD from native of the best template with average rank 1.2 (ranked on the basis of the template's Z-score) is 3.7 Å, and the average coverage is 89% of the target sequence. If we identify good local regions even though the global RMSD may be above 6.5 Å, then there are 1090 (83%) proteins with an average RMSD of 2.03 Å and 70% coverage. 96% of the 1317 proteins have a correct structural alignment to the template (RMSD <6.5 Å). Thus, there are problems with 177 proteins in generating better alignments with an additional 50 proteins whose templates were incorrectly selected. The sources of both errors are under examination.

For the 1317 threaded proteins, the average side chain contact prediction accuracy is 41%, with 64% correctly predicted within ±1 residue; their average contact order is 31.6 and an average number of contacts/residue, $f_{c,}$ of 2.3. For nonthreaded proteins, the average contact prediction accuracy is 17%, with 39% correctly predicted within ±1 residue; their average contact order is 16.2, and f_c, is 0.8. In Figure 2, in the solid histogram, we plot the cumulative fraction of structures having an RMSD <6.5 Å as a function of f_c for the 1317 threaded proteins. If f_c >1.5, then more than half of the predicted structures are likely to have an RMSD <6.5 Å; if f_c <1 then only about 1/3 of the templates are of this quality.

TOUCHSTONE was applied to the *ab initio* folding of 125 test proteins (36-174 residues in length) representing all secondary structural classes. 83/125 (66%) had a predicted structure with a RMSD below 6.5 Å in the top five clusters when predicted tertiary restraints are used. In the absence (with) of predicted tertiary restraints, 41/100 (70/100) small proteins (36~120 residues) have one of the top five lowest energy clusters with a RMSD below 6.5 Å. On average, inclusion of predicted contacts enhances the yield. If the goal is to select the single best cluster, then min E, max M, and min Y identifies 58 (61), 60 (67), and 65 (73), structures with an RMSD <6.5 Å (the best cluster independent of any RMSD threshold) respectively. Clearly, Y is the most sensitive.

RESULTS

Overview of Results

Structure predictions were made on all CASP5 targets with a total of 327 models submitted. With the exception of errors in submitting T0130 and T0150, five models were submitted for each target. A more detailed analysis of all our CASP5 predictions may be found at http://bioinformatics.buffalo.edu/casp5/. Included on our website are all models and fleshed out Tables.

In Table I, we summarize our results for all categories in terms of the length of the protein, the global RMSD, coverage and rank, the best structural alignment, coverage and rank, the best contiguous fragment, coverage and rank, the GDT_TS (Global Distance Test Total Score)[23] and rank, and the selected template, length and sequence identity to the target. For each category, as ranked by the

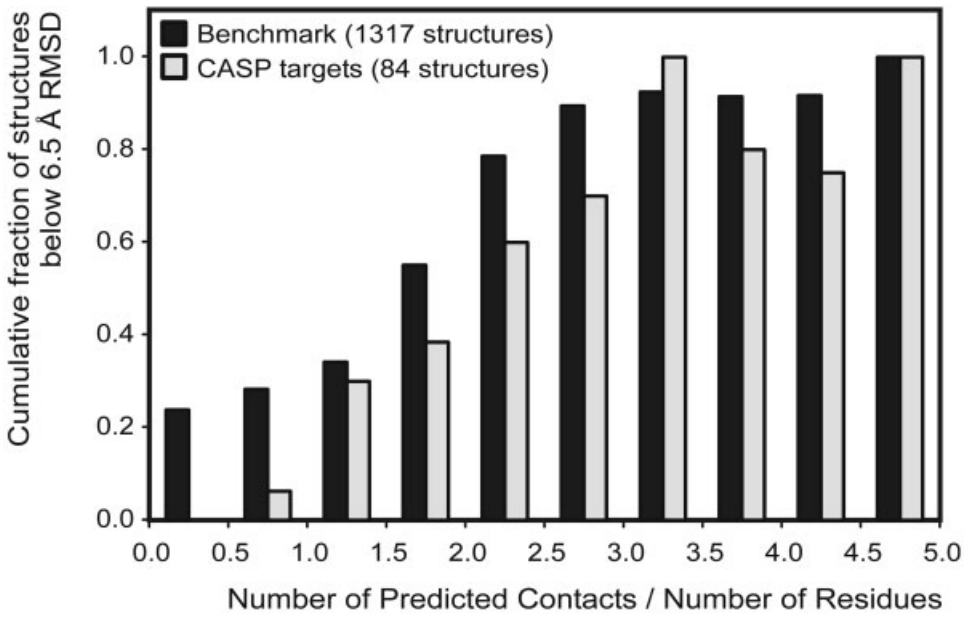

Fig. 2. Comparison of the cumulative distribution for the fraction of structures with an RMSD below 6.5 Å *versus* f_c (number of predicted contacts per residue) for 1317 benchmark structures (solid histogram) and 84 CASP5 targets/domains (open histogram).

GDT_TS, we present our best prediction relative to all other groups, the best among our predictions and the average of all our predictions. For those cases where the two criteria coincide, we also present the second best target among our predictions.

In Table II, for each category as ranked by RMSD, we present a summary of our contact predictions along with the results on the target corresponding to our best prediction relative to all other groups, the best of our predictions among our all predictions and the average of all our predictions. The average contact prediction accuracy for all 84 domains considered (for which structures are available), is 35% with 59% correct within ± 1 residue, an average contact order of 33.9 and an average of 2.0 contacts/residue predicted. This is comparable to our average benchmark performance, where the corresponding values over the entire set of 2186 test proteins is 32%, 55%, 25.5 and 1.7 respectively. Figure 2 (open histogram) shows the cumulative fraction of all structures with an RMSD <6.5 Å as a function of f_c. Very similar trends as in the benchmark (solid histogram) are evident.

Table III summarizes the improvement of all available targets from the initial template in the CM and FR categories. Often on building a physically realistic model, the alignment to the template improves, sometimes with a significant decrease in the alignment length (e.g. see T0134_2). The average RMSD of the initial template alignment over 84 targets is 9.9 Å; after physically realistic molecules are built, it improves to 6.7 Å. Hence, core alignment regions can be identified by building a physically realistic (but shorter) molecule. In only four cases, T0141, T0147, T0165, and T0190 are the RMSDs of the continuous models worse, going from 7.4 to 7.8 Å, 11.3 to 11.7 Å, 17.5 to 19.6 Å and 1.8 to 1.9 Å, respectively. A particularly interesting and encouraging case is T0191_2, whose initial RMSD is 28 Å over 128 residues and whose rebuilt model has an RMSD of 2.9 Å over 75 residues, with a final global RMSD of 2.7 Å.

Another important conclusion is that in only 7/84 cases does our methodology do harm (by comparing the global RMSD of the originally aligned residues to the initial aligned region template RMSD, and even then only by no more than 0.25 Å), and in a significant fraction of the cases, the models are considerably improved, a trend seen for all ranges of RMSD of the initial models. This represents a significant improvement in our methodology over CASP4.[24]

Particularly interesting are those cases in Table III, where *ab initio* folding yielded the best model (labeled by A in the method column). Using *ab initio* folding, in the CM category, the average RMSD for the threading template aligned residues improves from 6.1 Å to 3.6 Å. In the CM/FR(H) category, the average RMSD improvement is 13.0 Å to 8.7 Å. In the FR(H), category, the average RMSD improves from 10.6 Å to 8.4 Å, with an improvement in the FR(A) category from 14.4 Å to 6.4 Å. Finally in the FR(A)/NF and NF categories, when a template has been identified, the average improvement is from 15.2 Å to 11.3 Å and 14.0 Å to 9.7 Å respectively. We now focus on some specific cases that show significant improvements. In T0133, the RMSD in the aligned regions improves from 15.7 Å to 6.9 Å with a global RMSD of 6.7 Å. T0135 has an alignment that improves from 6.3 Å to 4.0 Å over the residues given by the threading template, with a global RMSD of 4.8Å, T0170 alignment improves from 10.8 Å to 5.0 Å, with a global RMSD of 5.3 Å, and T0188 alignment improves from 3.2 Å to 2.2 Å with a global RMSD of 2.2 Å.

TABLE I. Summary of CASP5 Results for Skolnick–Kolinski Group

	Target		Global coordinate superposition			Best structural alignment			Best coordinate superposition			Global distance test		Best template		
Category	Name	Length	RMSD [Å]	Coverage	Rank	RMSD [Å]	Coverage	Rank	RMSD [Å]	Coverage	Rank	Total score	Rank	PDB id	Length	S.I.[a]
CM	T0177_1[b]	57	—	—	—	—	—	—	—	—	—	94.30	5	1lfpA	56	0.43
CM	T0137[c,d]	133	1.0	1.00	3	1.0	1.00	3	1.0	0.99	3	94.36	1	1pmpA	130	0.45
CM	average (n[e])	152.9 (37)	4.1 (31)	1.00 (31)	2.9 (31)	2.4 (26)	0.93 (26)	2.8 (26)	1.5 (26)	0.68 (26)	2.4 (26)	74.10 (37)	2.5 (37)	(32)	160.1	0.33
CM/FR(H)	T0130[d]	100	4.5	1.00	1	3.6	0.88	1	2.5	0.70	2	54.00	1	1fa0A	71	0.14
CM/FR(H)	T0186_1[c]	77	3.9	1.00	3	2.7	0.94	3	1.7	0.77	4	76.30	4	1k6wA	316	0.17
CM/FR(H)	average (n)	218.3 (24)	8.4 (22)	0.95 (22)	2.5 (22)	3.8 (20)	0.77 (20)	2.6 (20)	5.0 (20)	0.76 (20)	3.0 (20)	44.57 (24)	3.0 (24)	(22)	184.2	0.14
FR(H)	T0138[d]	135	5.2	1.00	1	2.6	0.81	3	1.9	0.61	1	58.70	1	1b00A	106	0.17
FR(H)	T0193_1[c]	74	3.9	1.00	5	3.5	0.96	1	0.5	0.24	3	62.84	1	1fseA	65	0.22
FR(H)	average (n)	165.5 (10)	10.4 (10)	0.95 (10)	2.0 (10)	4.4 (9)	0.76 (9)	1.6 (9)	7.6 (9)	0.72 (9)	1.8 (9)	35.91 (10)	1.6 (10)	(10)	105.9	0.21
FR(A)	T0135[d]	106	4.8	1.00	1	3.8	0.93	1	4.2	0.89	1	50.71	1	1h6kX	75	0.13
FR(A)	T0148_1[c]	71	2.6	0.89	2	2.0	0.86	2	0.4	0.21	4	63.73	2	1b9kA	44	0.21
FR(A)	average (n)	126.3 (9)	8.7 (7)	0.95 (7)	2.4 (7)	4.4 (7)	0.74 (7)	1.9 (7)	5.1 (7)	0.66 (7)	3.3 (7)	42.19 (9)	2.8 (9)	(7)	130.7	0.18
FR(A)/NF	T0186_3[b]	36	9.4	1.00	3	3.5	0.53	2	1.0	0.25	5	36.81	1	1k6wA	28	0.11
FR(A)/NF	T0170[c,d]	69	5.3	1.00	4	4.4	0.86	4	0.7	0.25	3	61.60	3	1gln_	64	0.25
FR(A)/NF	average (n)	127.8 (10)	12.0 (10)	1.00 (10)	2.4 (10)	6.1 (10)	0.64 (10)	2.7 (10)	7.4 (10)	0.60 (10)	2.6 (10)	28.40 (10)	1.8 (10)	(10)	119.2	0.18
NF	T0181_1[b]	95	—	—	—	—	—	—	—	—	—	43.16	2	1a3yA	67	0.12
NF	T0181_2[c,d]	16	—	—	—	—	—	—	—	—	—	62.50	2	1a3yA	0	0.00
NF	average (n)	107.5 (10)	10.9 (7)	1.00 (7)	2.6 (7)	4.8 (5)	0.54 (5)	3.2 (5)	4.0 (5)	0.45 (5)	1.0 (5)	34.51 (10)	2.1 (10)	(9)	82.4	0.12

[a]Sequence identity between target and best template over the aligned region.
[b]Our second best prediction (based on GDT_TS) after comparison of all our models for all targets in the specified category.
[c]Our best prediction (based on GDT_TS) after comparison of all our models for all targets in the specified category.
[d]Our best ranked prediction after comparison of all models for all targets submitted by all predictors. This model selection for each category is made as follows. 1) For each target in the specified category, we first collect the best model submitted by each predictor (based on GDT_TS), and then we rank the models, assigning as rank 1 the one with the highest GDT_TS. 2) The best ranked prediction made by our group is selected; if more than one of our models for different targets satisfies this criterion, the decision is based on the higher GDT_TS score.
[e]Number of targets in the specified category.

TABLE II. Evaluation of Contact Prediction Accuracy

Category	N[a]	Target	Length	Npc[b]	CO[c]	f_c[d]	Accuracy (Allowed shift) 0	1	2	R[e] (Å)
CM		167_0[f]	180	434	51.4	2.41	0.71	0.86	0.93	2.9
CM		137_0[g]	133	556	38.5	4.18	0.57	0.77	0.87	1.0
CM	26	average	165.2	529.3	40.0	3.10	0.56	0.75	0.85	3.8
CM/FR(H)		185_3[f]	130	466	36.4	3.58	0.41	0.77	0.87	3.3
CM/FR(H)		185_1[g]	101	272	26.4	2.69	0.41	0.80	0.92	2.6
CM/FR(H)	20	average	200.1	404.0	38.9	1.96	0.36	0.64	0.77	8.3
FR(H)		157_0[f]	120	182	40.0	1.52	0.42	0.64	0.88	4.4
FR(H)		193_1[g]	74	142	24.1	1.92	0.35	0.58	0.72	3.9
FR(H)	9	average	166.6	176.2	22.9	1.23	0.25	0.44	0.60	10.3
FR(A)		135_0[f]	106	138	38.6	1.30	0.35	0.68	0.74	4.8
FR(A)		148_2[g]	91	218	21.3	2.40	0.24	0.50	0.73	2.6
FR(A)	7	average	147.1	311.4	29.8	2.30	0.23	0.48	0.64	8.7
FR(A)/NF		146_0[f]	299	380	31.9	1.27	0.10	0.25	0.43	18.1
FR(A)/NF		170_0[g]	69	84	19.5	1.22	0.33	0.62	0.71	5.3
FR(A)/NF	10	average	127.8	123.2	21.2	0.82	0.15	0.39	0.57	12.0
NF		129_0[f]	170	154	21.5	0.91	0.14	0.38	0.56	11.9
NF		139_0[g]	62	88	30.3	1.42	0.16	0.21	0.36	4.7
NF	5	average	106.2	89.6	21.3	0.88	0.15	0.38	0.56	9.8

[a]Number of targets in the specified category.
[b]Number of predicted contacts.
[c]Contact order.
[d]Number of predicted contacts/number of residues.
[e]RMSD of the best model.
[f]Our best ranked prediction after comparison of all models for all targets submitted by all predictors. This model selection for each category is made as follows. 1) For each target in the specified category, we first collect the best model submitted for each predictor (based on RMSD), and then we rank the models, assigning as rank 1 the one with the lowest RMSD. 2) The best ranked prediction made by our group is selected; if more than one of our models for different targets satisfies this criterion, the decision is based on the lower RMSD.
[g]Our best prediction (based on RMSD) after comparison of all our models for all targets in the specified category.

CM Results

As shown in Table I, the average RMSD, (GDT_TS) is 4.1 Å (74.1), with a best model rank of 2.9 (2.5) for a total of 31/37 (37/37) available targets. The relatively high average RMSD is due to three targets: T0140 has about 52% of its residues with a RMSD under 4 Å, but the global RMSD is poor, 13.6 Å, with the predicted C-terminus in error. In T0184, we get each domain correct, but their mutual orientation is wrong (the RMSD is 12.8 Å for the entire molecule, while the RMSD to native of T0184_1 and T0184_2 are 4.4 Å and 2.7 Å), a persistent problem. T0193_2 is also problematic; the alignment was driven by the N-terminal domain alignment, with the C- terminus weakly hit. Note that for the C-terminal domain, the fraction of predicted contacts/residue, f_c, = 0.7 where only about 28% of the threading templates have good alignments to the target.

T0137

In Figure 3(A), we display the coordinate superposition of the best predicted and native structures of T0137. The template is 1pmpA, 130/133 aligned residues with a sequence identity of 45%. The average contact prediction accuracy is 57%, with 4.2 contacts/residue predicted. A good model would be expected, and the resulting full-length structure has a RMSD from native of 1.0 Å for model 3.

CM/FR(H) Results

The 20 protein domains where structures are available were analyzed. The average RMSD is 8.4 Å, with the best structures having a mean rank of 2.5, and 10 (8) domains having a RMSD below 6.8 (6.5) Å. Examining structural alignments, the average RMSD is 3.8 Å with 77% average coverage. Even when the global alignment is in error, threading identifies significant fragments. Similarly, when we search for the longest continuous fragment, for 20 proteins, the average RMSD for the longest continuous piece is 5.0 Å with 76% coverage. Thus, better methods for identifying the correctly aligned regions need to be developed.

FR(H) Results

As shown in Table I, we analyzed 10 proteins, where T0138 is the best of our predictions compared to all groups according to the GDT_TS score, and T0193_1 is the best among all our predictions for the FR(H) category. The average RMSD is 10.4 Å, with the average RMSD from the best structural alignment of 4.4 Å and an average coverage of 76%. The average RMSD of the best continuous fragment is 7.6 Å with 72% coverage. Somewhat disappointingly, we only generated good structures in 3/10 cases: T0138, T0157, and T0193_1. Of the failures, T0134 and T0156 did not thread, T0174_1 was not recognized by our ***FASTA*** based profiles, but rather only by ***PSIBLAST***

TABLE III. Summary of Templates Hit by *PROSPECTOR* and Models Generated by Ab Initio Approach

Category	M[a]	Target	S.I.[b]	L_m[c]	L_a[d]	L_w[e]	R_a[f]	R_w[g]	R_b[h]	R_m[i]	ΔR[j]
CM	A	T0184_2[k]	0.27	72	66	58	2.3	2.1	2.4	2.8	−0.1
CM	A	T0193_2[l]	0.22	130	64	47	11.7	9.6	4.2	9.9	7.4
CM	A	average (4)[m]	0.23	163.0	136.5	106.8	6.1	5.1	3.6	5.4	2.5
CM	F	T0155[k]	0.34	117	115	107	0.8	0.7	1.0	1.1	−0.2
CM	F	T0191_2[l]	0.21	143	128	75	28.1	2.9	2.7	3.2	25.4
CM	F	average (22)[m]	0.34	165.6	152.0	129.5	3.7	2.0	2.3	3.5	1.4
CM/FR(H)	A	T0130[k]	0.14	100	70	31	5.1	2.8	3.7	4.5	1.4
CM/FR(H)	A	T0133[l]	0.15	293	220	178	15.7	15.6	6.9	6.7	8.9
CM/FR(H)	A	average (6)[m]	0.13	203.7	127.2	72.7	13.0	9.6	8.7	9.8	4.2
CM/FR(H)	F	T0159_2[k]	0.07	142	102	40	5.0	4.6	5.1	7.1	−0.1
CM/FR(H)	F	T0165[l]	0.14	318	293	212	17.5	19.5	11.5	11.8	6.1
CM/FR(H)	F	average (14)[m]	0.15	198.6	167.3	98.4	8.3	6.7	6.5	7.7	1.8
FR(H)	A	T0138[l]	0.17	135	105	59	4.2	3.1	3.9	5.2	0.2
FR(H)	A	T0134_2[k]	0.25	106	80	14	12.8	7.3	8.6	9.0	4.2
FR(H)	A	average (5)[m]	0.19	144.2	96.2	36.8	10.6	6.6	8.4	9.8	2.2
FR(H)	F	T0174[k]	0.21	352	120	36	14.5	12.3	12.6	14.2	1.9
FR(H)	F	T0174_2[l]	0.21	155	103	30	14.0	10.8	9.4	10.4	4.6
FR(H)	F	average (4)[m]	0.21	194.5	75.0	28.8	11.7	7.4	8.5	10.8	3.2
FR(A)	A	T0135[k]	0.13	106	74	38	6.3	4.2	4.0	4.8	2.3
FR(A)	A	T0148_1[l]	0.18	71	44	6	17.6	0.6	2.2	2.6	15.4
FR(A)	A	average (6)[m]	0.17	148.5	97.0	44.7	14.4	7.5	6.4	8.2	8.0
FR(A)	F	T0191_1[k, l, m]	0.21	139	125	50	17.0	15.4	11.1	11.5	5.9
FR(A)/NF	A	T0146_3[k]	0.16	56	60	13	11.7	5.7	9.4	10.0	2.3
FR(A)/NF	A	T0170[l]	0.25	69	63	33	10.8	6.6	5.0	5.3	5.8
FR(A)/NF	A	average (8)[m]	0.18	141.9	106.7	43.9	15.2	9.5	11.3	12.2	3.9
FR(A)/NF	F	T0146_3[k]	0.16	107	34	7	8.6	3.2	8.7	13.2	−0.1
FR(A)/NF	F	T0186_3[l]	0.17	36	28	10	12.0	7.6	9.0	9.4	3.0
FR(A)/NF	F	average (2)[m]	0.17	71.5	31.0	8.5	10.3	5.4	8.9	11.3	1.4
NF	A	T0129_1[k]	0.11	89	73	52	12.2	12.2	8.9	9.0	3.3
NF	A	T0139[l]	0.24	62	43	6	12.0	3.0	4.8	4.7	7.1
NF	A	average (5)[m]	0.15	106.2	86.8	46.0	14.0	9.0	9.7	9.8	4.3

[a]Method used to generate the best model, "F" indicates that the well-aligned template regions in our simulations are frozen; "A" means that ab initio folding is done.
[b]Sequence identity between target and best template over the whole aligned region (L_a).
[c]Total number of residues in the native structure whose coordinates can be compared with our best model.
[d]Length of aligned template regions hit by *PROSPECTOR*.
[e]Length of the well-aligned template regions, which are physically connectible.
[f]RMSD (Å) to native for the template over the entire aligned region (L_a).
[g]RMSD (Å) to native for the template over the well-aligned region (L_w).
[h]RMSD (Å) to native for the best model over the whole aligned region (L_a).
[i]RMSD (Å) to native of the best model over all the residues in the native structure whose coordinates can be compared with our best model (L_m).
[j]RMSD improvement by the CABS modeling simulation over the original template hit by *PROSPECTOR* (i.e., $\Delta R = R_a - R_b$).
[k]Target showing the minimum ΔR in the specified category.
[l]Target showing the maximum ΔR in the specified category.
[m]Average values (number of targets) for the specified category and method.

profiles that are prone to false positives. Thus, all failures reflect the inability of ***PROSPECTOR*** to recognize templates with corresponding very low contact predictions. Those proteins that threaded, T0138, T0157, and T0193_1, yielded acceptable models.

T0138

In Figure 3(B), we show T0138, where the RMSD of model 1, obtained from *ab initio* folding, over all residues is 5.2 Å. Interestingly, the RMSD of the best contiguous fragment is 1.9 Å over 83 residues. The sequence identity to the template, 1b00A, is 17% and the initial RMSD over 105 residues is 4.2 Å. Here, the average contact prediction accuracy is 53% with on average 2.1 restraints/residue predicted. The results are acceptable.

T0193_1

As shown in Figure 3(C), for T0193_1, the global coordinate RMSD of model 5 to native is 3.9 Å. The template is 1fseA, which has 22% sequence identity to the target over the aligned 65-residue region. Interestingly, the RMSD of the corresponding template and target is 6.5 Å, and *ab initio* folding significantly improved alignment quality. The contact prediction accuracy for this model is 35% with on average 1.9 restraints per residue predicted, a range where 55% of the templates in the benchmark have an acceptable RMSD. Again, the key to successful *ab initio* folding is have predicted contacts of reasonable number and accuracy. This coupled with the force field of the CABS model can produce significantly improved models.

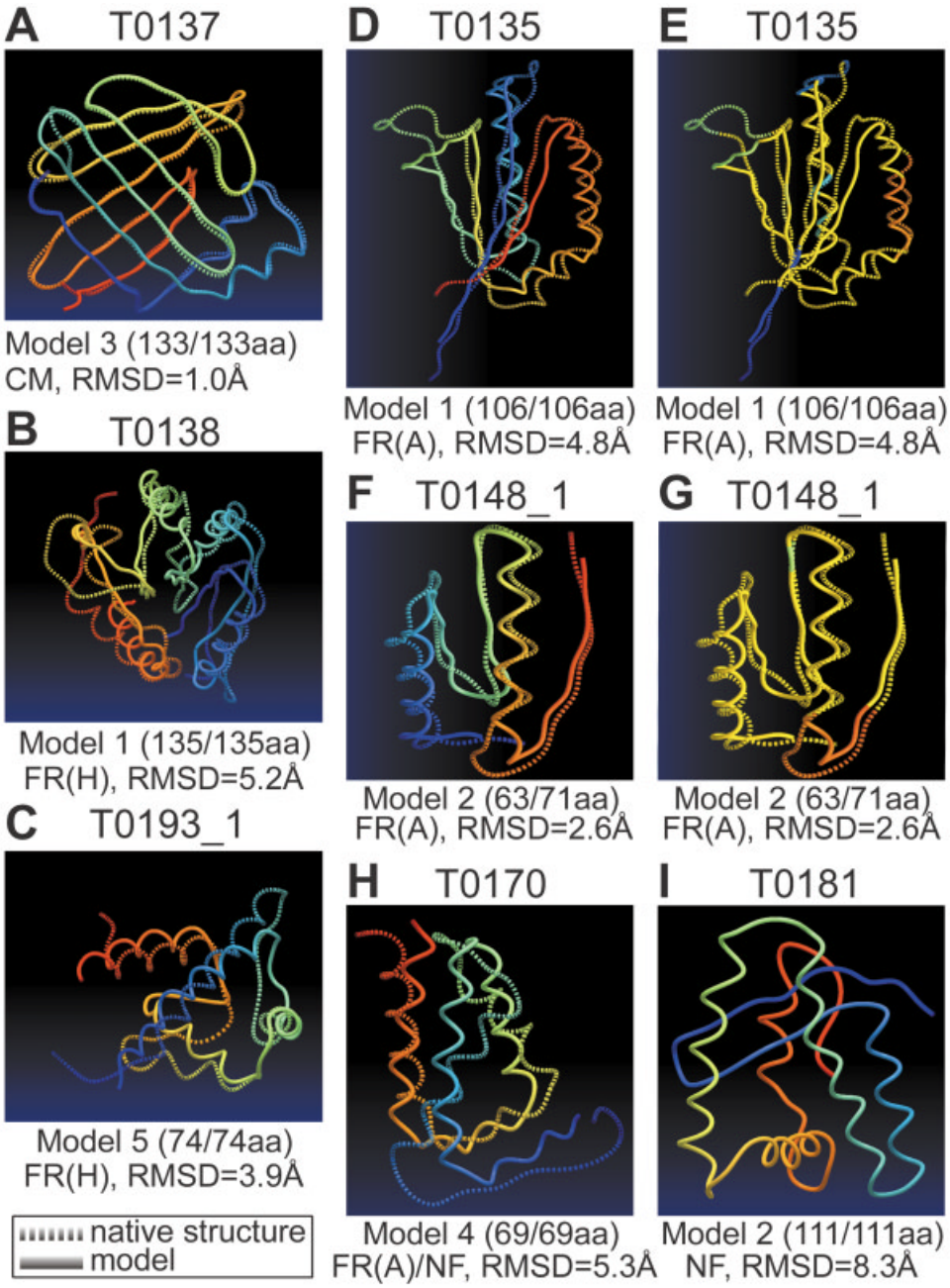

Fig. 3. Cα trace of our best models for selected CASP5 targets/domains. In all cases the model rank, model length, target/domain length, CASP category and RMSD to native are shown. **A-H:** coordinate superposition of the predicted and native structures of the indicated targets. **E** and **G:** same superposition as in **D** and **F** respectively, with the regions aligned to the template highlighted in yellow. **I:** best model for T0181 (the native structure of this target is unreleased). Blue to red runs from the N to C-terminus.

FR(A) Results

Next, we turn to the FR(A) category. There are 9 targets, with 7 available PDB structures. The average RMSD of these 7 targets is 8.7 Å. 3/7 have an RMSD below 6.5 Å. The average RMSD of the best structural alignments is 4.4 Å, with 74% coverage, and the best average continuous fragment coordinate superposition has an average RMSD of 5.1 Å with 66% coverage. All 7 have identified threading templates. The average RMSD of the best structural alignment to the best threading template (assigned on the basis of its Z-score) is 6.2 Å with an average coverage of 52%. 4/7 proteins have significant structural alignments, with T0148, T0187_2 and T0191 being problematic. The second (first) best template chosen for T0148 has a structural alignment over 28% (56%) of the molecule with an RMSD of 4.6 (11.0) Å. T0187_2 has over 50% of the

sequence aligned with a RMSD of 7 Å, a poor result. f_c =0.9, a value where good alignments are only generated in 28% of the cases. T0191_1 has f_c =1.7, a value where about 50% of the selected templates have good alignments but an average contact prediction accuracy of only 19%. We did not identify two domains, rather a global alignment on the entire molecule is chosen. T0191_2 is a CM target; thus we incorrectly chose the wrong template for the C-terminus. If the second best (best) template is chosen, then a structural alignment with 35% (41%) of the molecule aligned and an RMSD from native of 5.8 (10.2) Å is generated.

T0135

Shown in Figure 3(D) is the best coordinate superposition to the native state of model 1 for T0135. Using the GDT_TS score, T0135 is the best of our models (and is the best model) as compared to all other groups. The global RMSD of model 1 is 4.8 Å for the entire 106-residue molecule. Figure 3(E) shows the same superposition as in Figure 3(D), but with the regions aligned to the template (1h6kX, sequence identity of 13%) shown in yellow. The raw target template alignment has a RMSD from native of 6.3 Å over 74 residues. The average contact prediction accuracy is 35%, with 138 contacts predicted. These results clearly demonstrate that ***PROSPECTOR*** can identify analogous global templates at least in some cases and that our *ab initio* folding algorithm can improve the global RMSD over that of the template.

T0148_1

Another interesting case is T0148_1 shown in Figure 3(F), which is the best of our FR(A) models ranked against each other by the GDT_TS score. The initial threading alignment has an RMSD from native of 17.6 Å over 44 residues. The global RMSD for model 2 over 71 residues is 2.6 Å, has a sequence identity to the template (1b9kA) of 21%, and the aligned region to the template is shown in yellow in Figure 3(G). This model is produced by *ab initio* folding, again showing the improvement over the initial alignments. The average contact prediction accuracy is 30%, with 95 contacts predicted.

FR(A)/NF Results

There are 10 proteins, and as shown in Table I, T0170 is the best of our predictions as assessed by the GDT_TS, both in terms of the our relative performance to other groups as well as the best performance among our predictions. We analyze this molecule in further detail below. The average RMSD of all 10 targets is poor, 12.0 Å, but the average RMSD obtained from structural alignments is considerably better, namely 6.1 Å, with the best continuous piece having an average RMSD of 7.4 Å. 9/10 proteins have acceptable structural alignments to their selected best template, with an average RMSD of 4.9 Å and 50% coverage. The one problematic protein is T0172_2 that has a best structural alignment RMSD of 7.6 Å over 55/101 residues.

Based on the expected likelihood of success shown in Figure 2 given the fraction of predicted contacts/residue, f_c, we would have anticipated that T0146 (f_c =1.3), T0173 (f_c =1.1), T0187_1(f_c =0.78), T0146_3 (f_c =0.79), for T0146_1 (f_c =0.51), T0146_4 (f_c =0.47) and T0172_2(f_c =0.3) would have low resolution predictions. f_c =1.66 for T0146_2, a score which for the threading validation set provides good coordinate alignments in about 40% of the cases. Thus, with the possible exception of T0146_2, these results are consistent with our benchmark.

T0170

The superposition of model 4 of T0170 onto its native structure is shown in Figure 3(H), where the RMSD is 5.3 Å. Interestingly, our best threading template is 1gln_ which has the mirror image topology. Thus, the RMSD of the target to this topological mirror image is 10.9 Å. Nevertheless, the contact prediction accuracy is 33% with f_c= 1.3, where about 1/3 of identified templates have good alignments. Our *ab initio* folding algorithm assembles both topologies. This again points out the utility of using Generalized Comparative Modeling and *ab initio* folding for proteins not in the CM category.

NF Results

There are a total of 10 targets of which the structures of 7 are available for analysis. For T0161, we had identified a template (1hv8A) but apparently this model is incorrect as no structure is available for comparison. For T0162_3, again no structure is available, but our threading template, 1ac7A, would predict some contacts between T0162_1 and T0162_3. For T0129, T0129_1, T0129_2, we identified these as NF targets (***PROSPECTOR*** does not find any template), and f_c=0.9, 0.6 and 0.9, respectively indicative of low foldability, and only 14%, 23% and 11% of the contacts are correct. ***PROSPECTOR*** did not identify a template for T0149_2; f_c=0.6 with only 12% of the predicted tertiary contacts are correct. Thus, this was a pure *ab initio* target that we failed to adequately treat.

T0139 is very interesting. In our CASP5 submission, we used ***PSIPRED*** [15] to predict secondary structure and found that a β-strand was incorrectly assigned to an α-helix, thereby giving for the best prediction, model 2, an RMSD of 9.9 Å. If either the correct secondary structure or the consensus of the CAFASP predictions is used, then the resulting best model RMSD from native is 4.7 Å without predicted tertiary restraints. This points out the importance of pooling secondary structure prediction schemes; if they give contradictory assignments for a given secondary structural element, both predictions should be done. This we failed to do during CASP5.

Of the 5 domains for which structural information is available, 4/5 had significant structural alignments with an average RMSD of 4.8 Å, and an average fraction of aligned residues of 54%, thereby indicating that some significant fragments are correctly predicted, even though the global fold is wrong. For one of the domains, T0149_2, the best continuous fragment has an RMSD of 6.1 Å.

At the December, 2002 CASP5 meeting, we were asked to talk about T0181. Shown in Figure 3(I) is model 2, which is topologically correct but which has an RMSD from

native of 8.3 Å. The first and third models (both wrong) have the same topology, but the latter was generated without any tertiary restraints. There are 72 contacts predicted, (f_c=0.65), for this molecule, no template is identified, so we would have expected this to be a low-resolution prediction; this is what happened.

DISCUSSION

We have presented an overview of the performance of our ***TOUCHSTONE*** algorithm on all targets/domains in all categories. ***TOUCHSTONE*** performed reasonably well in the CM, CM/FR(H), FR(H) and FR(A) categories, with significant improvement needed in the FR(A)/NF and NF categories. The overall performance can be predicted on the basis of the ratio of the number of predicted side chain contacts/number of residues. If a significant number of side chain contacts are predicted, this is indicative that there are strongly aligned regions and folding is likely to be successful. In the absence of predicted contacts, unfortunately our success rate is low.

For non NF categories, we use a two pronged approach: In one case, we apply Generalized Comparative Modeling where a significant portion of the template aligned regions identified by ***PROSPECTOR*** is frozen and the gaps filled in. An interesting result is that the process of building a continuous chain identifies the better aligned regions and in most cases, the global RMSD of the resulting model is lower that the RMSD in the original ***PROSPECTOR*** provided alignment. Second, we employ the predicted contact restraints in *ab initio* folding. In almost all cases, the global RMSD of the submitted models improves, sometimes significantly. The combination of the *ab initio* force field when guided by predicted contacts can give better results than the input threading based information. For all classes of non NF targets, the RMSD of the final model is often better than that of the initial template alignment. This represents a significant improvement over our past results.

For the FR(A) category, we demonstrated that ***PROSPECTOR*** can identify analogous templates that cover a significant fraction of the entire target sequence. In contrast to fragment assembly algorithms, global templates are identified. Whether these are analogous or distantly homologous proteins is difficult to establish. What is very encouraging is that ***PROSPECTOR*** as well as a number of other threading algorithms have gone beyond the ***PSI-BLAST***[25] barrier in fold recognition ability, this represents qualitative improvement over previous CASPs. With respect to ***PROSPECTOR***, it is important to reemphasize that it does not use any input from metaservers (the only input information is predicted secondary structure from ***PSIPRED***[15] and ***FASTA***[17] sequence alignments), rather it is a stand-alone algorithm. But we could readily see improvements in ***PROSPECTOR*** if input from metaservers were employed to generate the initial alignments used for subsequent contact prediction/template identification. Finally, at least for the FR(A) and easier categories, ***PROSPECTOR*** side chain contact prediction has become sufficiently accurate that it is of use in structure prediction.

The clear weaknesses of ***TOUCHSTONE*** are as follows. First, for about 30% of the cases the raw alignments from ***PROSPECTOR*** are of high RMSD, even though a good template has been identified. Sometimes structure refinement using Generalized Comparative modeling or *ab initio* folding can significantly improve the alignments, but better alignments need to be generated in the first place. The use of metaservers as suggested above may assist in this regard. Second, we need to significantly improve our procedure in fold selection. It is clear that we often had very good models, but failed to rank them as number 1. Our fold selection process could perhaps be improved by building atomic models and ranking them with better knowledge based[26] or molecular mechanics potentials.[27] Alternatively, perhaps a neural network could be trained to recognize good from bad folds.[28] Along these lines, algorithms that can refine structures from 3-4 Å to 2-3 Å from native are still wanting. Third, we need to develop better domain parsing algorithms, and where the domains are well identified, develop procedures for predicting their mutual orientation. Fourth, the *ab initio* component of ***TOUCHSTONE*** relies too heavily on the accuracy of side chain contact prediction. The inherent potentials of the CABS protein model need improvement. We are attempting to do this by performing *ab initio* simulations on all representative PDB structures $\leq$ 200 residues in length. This will establish, on a very large scale NF set, the strengths and weaknesses of the algorithm and provide sufficient statistics to suggest means of improving the model. Concomitantly, we need to improve the accuracy of contact prediction in the NF regime. Sixth, we need a better way of combining secondary structure prediction schemes so that inconsistent predictions of secondary structural elements (e.g. as for T0139) are handled. This may be done either by folding with different secondary structure predictions and then selecting among the resulting predicted structures and/or by differential weighting of various secondary structure predictions according to their reliability. Seventh, for NF, ***TOUCHSTONE*** needs to be extended to treat larger proteins.

Overall, the current generation of ***TOUCHSTONE*** represents a considerable improvement over our CASP4 generation. The reasons for this improvement are that the current version of ***PROSPECTOR*** is a much more sensitive threading algorithm and the resulting alignments are longer and the contact predictions are more numerous and more accurate, we have implemented a continuous space model to treat close CM proteins so that there is no loss of accuracy due to lattice artifacts, and the CABS lattice model has a better conformational sampling scheme and its potentials are better tuned. Thus, while major improvements are necessary, the issues are well defined, and there is reason to hope for additional progress in the near future.

REFERENCES

1. Rost B, Sander C. Bridging the protein sequence-structure gap by structure predictions. Annu Rev Biophys Biomol Struct 1996;25: 113-136.

2. Finkelstein AV, Reva BA. A search for the most stable folds of protein chains. Nature 1991; 351: 497-499.
3. Fischer D, Elofsson A, Rice D, Eisenberg D. Assessing the performance of fold recognition methods by means of a comprehensive benchmark. Pac Symp Biocomput 1996; 300-318.
4. Rice DW, Eisenberg D. A 3D-1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence. J Mol Biol 1997; 267: 1026-1038.
5. Fischer D. Modeling three-dimensional protein structures for amino acid sequences of the CASP3 experiment using sequence-derived predictions. Proteins 1999; Suppl: 61-65.
6. Godzik A, Skolnick J. Sequence-structure matching in globular proteins: Application to supersecondary and tertiary structure determination. Proc. Natl. Acad. Sci. USA 1992; 89: 12098-12102.
7. Jones DT. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol 1999; 287: 797-815.
8. Skolnick J, Kihara D. Defrosting the frozen approximation: PROSPECTOR—a new approach to threading. Proteins 2001; 42: 319-331.
9. Russell RB, Saqi MAS, Sayle RA, Bates PA, Sternberg MJE. Recognition of analogous and homologous protein folds: analysis of sequence and structure conservation. J. Mol. Biol. 1997; 269: 423-439.
10. Levitt M, Gerstein M, Huang E, Subbiah S, Tsai J. Protein folding: the endgame. Annu Rev Biochem 1997; 66: 549-579.
11. Bystroff C, Baker D. Prediction of local structure in proteins using a library of sequence- structure motifs. J Mol Biol 1998; 281: 565-577.
12. Ortiz AR, Kolinski A, Rotkiewicz P, Ilkowski B, Skolnick J. Ab initio folding of proteins using restraints derived from evolutionary information. Proteins 1999; 37: 177-185.
13. Sternberg MJ, Bates PA, Kelley LA, MacCallum RM. Progress in protein structure prediction: assessment of CASP3. Curr Opin Struct Biol 1999; 9: 368-373.
14. Kihara D, Lu H, Kolinski A, Skolnick J. TOUCHSTONE: an ab initio protein structure prediction method that uses threading-based tertiary restraints. Proc Natl Acad Sci U S A 2001; 98: 10125-10130.
15. McGuffin LJ, Bryson K, Jones DT. The PSIPRED protein structure prediction server. Bioinformatics 2000; 16: 404-405.
16. Kolinski A, Betancourt MR, Kihara D, Rotkiewicz P, Skolnick J. Generalized comparative modeling (GENECOMP): a combination of sequence comparison, threading, and lattice modeling for protein structure prediction and refinement. Proteins 2001; 44: 133-149.
17. Pearson WR. Rapid and sensitive sequence comparison with FASTP and FASTA. Methods Enzymol 1990; 183: 63-98.
18. Zhang Y, Kihara D, Skolnick J. Local energy landscape flattening: parallel hyperbolic Monte Carlo sampling of protein folding. Proteins 2002; 48: 192-201.
19. Swendsen RH, Wang JS. Replica Monte Carlo simulation of spin glasses. Physical Review Letters 1986; 57: 2607-2609.
20. Hansmann UHE. Parallel tempering algorithm for conformational studies of biological molecules. Chem. Phys. Lett. 1997; 281: 140-150.
21. Zhang Y, Kolinski A, Skolnick J. TOUCHSTONE II: A new approach to ab initio protein structure prediction. Biophys J 2003; 85: 1145–1164.
22. Betancourt MR, Skolnick J. Universal similarity measure for comparing protein structures. Biopolymers 2001; 59: 305-309.
23. Zemla A, Venclovas, Moult J, Fidelis K. Processing and evaluation of predictions in CASP4. Proteins 2001; Suppl 5: 13-21.
24. Skolnick J, Kolinski A, Kihara D, Betancourt M, Rotkiewicz P, Boniecki M. Ab initio protein structure prediction via a combination of threading, lattice folding, clustering, and structure refinement. Proteins 2001; Suppl 5: 149-156.
25. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997; 25: 3389-3402.
26. Lu H, Skolnick J. A distance-dependent atomic knowledge-based potential for improved protein structure selection. Proteins 2001; 44: 223-232.
27. Felts AK, Gallicchio E, Wallqvist A, Levy RM. Distinguishing native conformations of proteins from decoys with an effective free energy estimator based on the OPLS all-atom force field and the Surface Generalized Born solvent model. Proteins 2002; 48: 404-422.
28. Lundstrom J, Rychlewski L, Bujnicki J, Elofsson A. Pcons: a neural-network-based consensus predictor that improves fold recognition. Protein Sci 2001; 10: 2354-2362.