Published in final edited form as:
Proteins. 2009 October ; 77(1): 220–234. doi:10.1002/prot.22434.

A novel method for predicting and using distance constraints of high accuracy for refining protein structure prediction

Tianyun Liu[1,2], **Jeremy A. Horst**[2,3], and **Ram Samudrala**[2,3,*]

[1]Department of Genetics, Stanford University, Stanford, California

[2]Department of Microbiology, School of Medicine, University of Washington, Seattle, Washington

[3]Department of Oral Biology, School of Dentistry, University of Washington, Seattle, Washington

Abstract

The principal bottleneck in protein structure prediction is the refinement of models from lower accuracies to the resolution observed by experiment. We developed a novel constraints-based refinement method that identifies a high number of accurate input constraints from initial models and rebuilds them using restrained torsion angle dynamics (rTAD). We previously created a Bayesian statistics-based residue-specific all-atom probability discriminatory function (RAPDF) to discriminate native-like models by measuring the probability of accuracy for atom type distances within a given model. Here, we exploit RAPDF to score (i.e., filter) constraints from initial predictions that may or may not be close to a native-like state, obtain consensus of top scoring constraints amongst five initial models, and compile sets with no redundant residue pair constraints. We find that this method consistently produces a large and highly accurate set of distance constraints from which to build refinement models. We further optimize the balance between accuracy and coverage of constraints by producing multiple structure sets using different constraint distance cutoffs, and note that the cutoff governs spatially near versus distant effects in model generation. This complete procedure of deriving distance constraints for rTAD simulations improves the quality of initial predictions significantly in all cases evaluated by us. Our procedure represents a significant step in solving the protein structure prediction and refinement problem, by enabling the use of consensus constraints, RAPDF, and rTAD for protein structure modeling and refinement.

Keywords

protein structure prediction; refinement; knowledge-based functions

INTRODUCTION

The accuracy of protein structure prediction

The goal of protein structure prediction is to model the three-dimensional (3D) structure for a protein sequence with an accuracy comparable to that observed by experiment. In the last 2 decades, significant progress has been made on the methods of template-based (comparative) modeling and template-free modeling (e.g., *de novo* or *ab initio*) predictions,[1–4] with the former yielding the most accurate models.[5]

*Correspondence to: Ram Samudrala, Department of Microbiology, School of Medicine, University of Washington, Seattle, WA. ram@compbio.washington.edu.

Additional Supporting Information may be found in the online version of this article.

The accuracy of template-based prediction depends on the selection of a proper template and the quality of the sequence alignments between the target and the template sequences. For modeling at low sequence similarity, template-based prediction generates low accuracy structures compared with the corresponding experimental structure, even when the best available template structures are identified.[6] Free modeling can produce accurate models of small, single-domain proteins[7]; however, these methods generally produce medium- to low-resolution models for proteins comprising more than 100 residues or multiple domains.

Achieving higher resolution requires an improvement in efficiently searching the conformational space, designing more accurate scoring functions, and an efficient refinement protocol. Although considerable advancement has been achieved in solving these problems, it has not led to an obvious improvement in the accuracy of free modeling predictions.[8] Therefore, it remains unclear whether methods for generating initial models will ever consistently produce structures of accuracy within the high-quality experimental range. Thus, there is a need for methods that increase the accuracy of initial predicted models, such that the refined models are sufficient for mechanistic studies of function and drug design.

Abilities and limitations of constraints-based modeling

Methods for constraints-based modeling have been well developed to construct and refine 3D conformations to be consistent with spatial constraints from diffraction and resonance experiments.[9] These methods have also been extensively applied to both template-based modeling[10,11] and free modeling prediction.7,12 There are two major categories of constraint-based modeling algorithms: distance geometry embedding, which uses a metric matrix of distances from atomic coordinates to their collective centroid, to project distance space to 3D space13–[15]; and minimization, which incorporates distance constraints in variable energy optimization procedures, such as restrained torsion angle dynamics (rTAD), molecular dynamics, simulated annealing, or Monte Carlo searches.[16–19]

rTAD provides at present the most efficient way to calculate a protein structure from constraints, by working with internal coordinates rather than Cartesian coordinates. The number of degrees of freedom is decreased, as the covalent structural parameters (i.e., bond lengths) are kept fixed at their optimal values during the calculation.[20] In addition, simulated interatomic forces are used to constrain protein conformations and thereby greatly reduce the effective conformational space.9

Typically, rTAD simulations start by producing a set of secondary structure elements. These elements are then assembled into compact structures using input distance constraints, with the major degrees of freedom in these calculations being the φ/ψ dihedral angles of the polypeptide chain, for which the preferences are defined by the Ramachandran plot.[21] Experimentally derived conformational constraints have proven to be sufficiently strong to guide rTAD minimization into a correct structure.[9]

Previous studies estimated the minimum number of distance constraints required to obtain effective constraints-based models. The earliest work by Smith-Brown *et al.* showed that three correct distances between each predicted secondary structure element were sufficient to generate a conformation of 3–5 Å alpha carbon root mean squared deviation (CαRMSD) to the experimental model.[22] Later work by Aszodi *et al.* showed that a conformation below 5 Å CαRMSD could be obtained when at least *N*/4 correct constraints were used, where *N* is the number of residues in the protein. However, selection of the correct fold was not achieved because of the lack of a knowledge-based force field.[14] MONSSTER by Skolnick *et al.* is able to fold small proteins using *N*/7 constraints for α proteins and *N*/4 constraints for β and α/β proteins.[18] However, the necessity to derive all distance constraints from NMR experiments or experimentally determined secondary structures is a significant limitation.[23] Later work by

the Skolnick group using restrained Monte Carlo searches showed that the generation of 3.0–6.5 Å CαRMSD models using distance constraints derived from multiple sequence alignments is possible.[19] Further work by Huang *et al.* showed that consensus distance geometry can reliably yield structural models within 6.5 Å CαRMSD for small helical proteins.[24,25] On the basis of this and previous work, we hypothesize the numbers of constraints required to achieve models of various resolutions (see Fig. 1).

Despite the significant advances represented in earlier work, effective constraints-based modeling thus far requires external information, including correct secondary structure assignment and sufficiently accurate distance constraints. Moreover, the accuracy of constraints-based modeling depends on the amount and specificity of available information. Distance constraints may be derived from initial predictions, such as secondary structure and 3D conformational predictions, without experimental information. However, constraints derived from predictions are noisy or ambiguous in nature. For example, the best performing group in CASP6 contact prediction achieved only an accuracy of 25.5% and coverage of 3.7%, [29] whereas longer distance contact prediction accuracy was limited to around 20%.[30]

It remains unclear whether rTAD modeling using a limited number of correct constraints is robust enough to successfully handle the unavoidable presence of incorrect constraints in sets derived from predictions. Therefore, identification of accurate constraints remains a major bottleneck to constraints-based predictive modeling techniques, which represents the focus of this work.

Constraints can be derived from the consensus of multiple initial predictions

For a given target protein, several different template structures are usually available. A given template/alignment combination is unique in its similarity to the target protein in different ways, thus template-based modeling using different templates/alignments produces a variety of structural models for a given target protein.[31] Even using the same template/alignment, different modeling methods sometimes yield highly dissimilar models because of variations in the side chain and loop building processes.[32] The goal of free modeling prediction is to obtain the overall polypeptide fold, often by starting with a random chain and searching for the most energetically favorable or statistically probable conformations. The process itself determines the diversity of structural information presented by different free modeling predictions.[4] Even when global similarity is not significant, initial models share structural information, particularly within functional sites.

We therefore ask the following question: given a set of initial predictions derived from multiple sources or methods, how can one take into account all of the available information in a rational way to derive accurate and sufficient constraints for rTAD? We hypothesize that dissecting the best initial models into a set of consensus interatomic distances, filtered and weighted by the Bayesian probability of being native-like, will result in constraint sets of sufficient accuracy and coverage to enable rTAD methods to produce final models more closely resembling biologically relevant conformations.

Objective

To refine predicted 3D conformations using rTAD techniques, the two major problems are (i) how to obtain constraints of good quality and quantity from initial predictions and (ii) how to make use of these constraints effectively. To answer the former question, we use a consensus-based method to derive an initial set of distance constraints from five initial template-based modeling predictions, filter and weight the set using a Bayesian approach (RAPDF), which assigns a likelihood score to each constraint using parameters derived from the observed distances in a set of 4000 experimentally determined nonredundant protein structures, then

subdivide the set by distance cutoffs, and finally evaluate the ability of this combined method to effectively produce large and accurate sets of distance constraints. To answer the latter question, we evaluate the ability of rTAD simulations using the program Combined assignment and dynamics algorithm for NMR applications (CYANA) [20] with input constraints sets of differing accuracy, to produce refined conformations of higher quality. Finally, we present the performance of the constraint derivation and rTAD simulation methods together as one contiguous pipeline, for the refinement of 30 predicted protein structures.

RESULTS AND DISCUSSION

RAPDF scores versus consensus constraint accuracy

Deriving accurate constraints from initial structure predictions is one of the major problems in using rTAD simulations for refinement. Inaccurate consensus constraints trap rTAD simulations in local energy minima and therefore should be filtered out. Our goal was to design a method to discriminate accurately predicted distance constraints from inaccurate ones. The probability of each distance constraint being accurate was estimated using the all-atom discriminatory function RAPDF. To quantify the relationship between the accuracy of a consensus constraint and the assigned RAPDF score, a total of 10 subsets of consensus constraints were derived according to their RAPDF ranking cutoffs, that is, 10% rank and better, 20% rank and better, and so forth. The accuracies of $consensus_5$ and $consensus_4$ constraints, representing constraints derived from the consensus of five and four of the give initial selected models, respectively, in each subset were plotted against the respective RAPDF ranking cutoffs. To quantify the relationship between the accuracy of consensus contacts and the RAPDF rank, and therefore the reliability of RAPDF to work as such a filter, we calculated the accuracy of consensus constraints as Accuracy = [The number of nonlocal contacts for which the selected distance constraint fell within 0.25 Å of the correct distance]/[Total number of nonlocal contacts considered as a subset].

Figure 2 shows that the accuracy of consensus constraints correlates with the RAPDF ranking cutoffs, suggesting that RAPDF scores of the consensus contacts can be used as a standard for selecting constraints of higher accuracy. Figure 2 also shows that $consensus_5$ is almost always more accurate than $consensus_4$, with only one exception in 30 (CASP7 target T0298). The average accuracy of consensus constraints of the 30 targets was also calculated and compared (Table I), which confirm that the average accuracy of consensus constraints correlates with the RAPDF ranking. The average accuracy of the top 10% RAPDF score $consensus_5$ constraints is ~59%, whereas the average accuracy when including all constraints drops to 42%. These results indicate that the consensus-based method combined with the Bayesian approach is able to filter out many incorrect distance constraints, resulting in higher accuracy constraints.

Constraint accuracy versus rTAD model quality

How does the inaccuracy inherent to predictive modeling affect the performance of the rTAD simulations? To answer this question, we derived five constraints sets of progressively worsening accuracy for each target protein, by assigning increasingly more divergent distance values to randomly selected constraints within an increasingly larger proportion of constraints derived from the best initial predictions. These five sets share the same number of constraints for each atom type pair, with varying proportions of correct distances derived from the experimental model and incorrect distances derived from predicted models, such that the constraint accuracy is the only factor that may affect the results of the rTAD simulations. Each set of constraints was directly input to CYANA, which constructed a set of 1000 conformations satisfying the constraints using TAD for each subset for each protein. The best refined CYANA model generated in each set of restrained simulations was compared with the best initial model of that protein by calculating CαRMSDs to the corresponding experimental structure.

Figure 3(A) shows the improvement achieved by the rTAD simulations using consensus constraints of different accuracies. In all tested target proteins, the improvement of the best CYANA models is more significant when the consensus constraints of higher accuracy are used in restrained simulations. In other words, the performance of rTAD simulations is improved when the accuracy of constraints increases. In a few cases, a perfect trend between accuracy of constraints and final models was not observed, wherein better models were produced in the second or third highest accuracy set. However, in all cases, the models derived from 100% accuracy sets were better than models derived from the lowest accuracy set. A possible explanation is the tendency for models to become trapped in local energy minima: as limited time and number of seeds are used in the simulations, the result can be an insufficient sampling space for some proteins. Another explanation is a real alternate conformation. In either case, the overall result indicates that obtaining constraints of higher accuracy is important to achieve higher quality conformational predictions.

We also calculated the average accuracy for each progressively more accurate constraint set, across the 30 target proteins (Table II). The increase of constraint accuracy (lowest to highest set) produced a direct relationship with the quality of the best refined CYANA models, ranging from 4.5 to 1.9 Å average CαRMSD improvement from the original 4.8 Å average CαRMSD. Thus, for simulations using higher accuracy constraints (higher and highest sets), the average improvement of the best CYANA models to the best initial models is 2.0 to 3.0 Å CαRMSD, whereas when lower accuracy constraints are used (lowest set) the average improvement is only 0.3 Å CαRMSD. Therefore, the CαRMSD of refined CYANA models is improved when higher accuracy consensus constraints are used in rTAD simulations.

Finally, we evaluated all conformations generated from the five constraint set types for the 30 targets as one sample pool. Figure 3(B) shows that the percentage of CαRMSD improvement between the best CYANA models and the best initial models correlates highly with the consensus constraint accuracy, measured as a correlation coefficient of 0.8. Improvement was seen in only 14 of 27 refinement processes with input constraint accuracy below 40%, whereas in all but 1 of 123 cases with input constraint accuracy above 40%, the models were refined to greater accuracy.

In summary, although the randomization of inaccurate constraint distances in this experiment potentially creates a worst case scenario for each prescribed constraint accuracy set (wherein a 3 Å constraint could arbitrarily be assigned as 20 Å, rather than a closer value which might be more realistically derived from in an initial model), the quality of protein structure prediction is predictably improved when using distance constraints of more than 40% accuracy in rTAD simulations.

Maximizing constraint set coverage

Attempting to model regions of proteins without constraints, using rTAD, results in the flexible tails familiar to depictions of structures from NMR peaks. Thus, without a sufficiently large set of constraints, the simulations will result in largely unfolded regions, and therefore, we need as many accurate constraints as possible. By using a batch-by-batch process, the number of residue pairs represented by selected constraints is estimated by the equation: $N^*(N-4)/10$, where N is the total number of residues in the protein. For a protein of more than 104 residues, the number of constraints will be larger than $10N$. Compared with previous methods, this amount of constraints is considerably large.[14,18,19,24,25] The constraint accuracy corresponding to increasing coverage can be seen in Table III. The final model accuracy produced when using this compilation method in a realistic modeling situation is seen in Figure 4.

Constraint distance cutoff

Four subsets of constraints were derived using different distance cutoffs, including cutoffs of 8, 12, 16, and 20 Å (Table III). These four constraint sets were used as input for simulations using CYANA (see Fig. 4). For each subset, two different sets of dihedral angle parameters were tested to compare the influence of angle constraints on restrained simulations. The two sets are compiled from the best initial models and the best RAPDF scoring models, respectively. In essence, we sought to define the limitation caused by nonideal selection of the input model, which is a potential weak link as it relies on only one model rather than a consensus as found with other aspects of our method. In total, eight sets of different distance constraint and dihedral angle parameters were tested for each of the 30 tested target proteins. From each set of simulations, a total of 1000 refined CYANA models were generated. All these 1000 conformations were ranked by the discriminatory function RAPDF, and the top 10 scoring models were used for further analysis.

For each cutoff-derived distance constraint set, we calculated the average accuracy and coverage of the constraint set itself, and the refined CYANA models generated from different sets of simulations as well (Table III). When using a distance cutoff of 8 Å, the average accuracy of 30 tested targets was 45% within a 0.5 Å width, whereas the coverage of total residue pairs in the tested targets was only 7%. The improvement of the best refined CYANA models is about 0.6 Å CαRMSD, whereas that of the best top 10 scoring CYANA model is only 0.2 Å CαRMSD. As the distance cutoffs change from 12, 16, to 20 Å, the coverage of the constraints increases from 16, 24, to 30%, respectively. As well, the left panel of Figure 4(A) shows that the best CYANA models generated at these longer cutoffs always have lower CαRMSDs than the corresponding best initial models, whereas those produced with an 8 Å cutoff show higher CαRMSDs for 4 of the 30 targets, and, in general, the quality of the best CYANA models is not as good as those at cutoffs of 12, 16, and 20 Å. Improvements of the best refined CYANA models are on an average of 1.1 Å CαRMSD in all three cases. In addition, the improvement of the averages of the best top 10 scoring CYANA models average 1.1, 1.2, and 1.3 Å CαRMSD, respectively. This indicates that using a distance cutoff of 20 Å results in the largest improvement. Directly, restrained simulations using constraints at longer distance cutoffs result in conformations of higher quality.

For example, the best initial model of T0332 is a conformation with 2.8 Å CαRMSD, whereas the best refined CYANA models generated by simulations at distance cutoffs of 8, 12, 16, and 20 Å are 2.0, 1.9, 1.7, and 1.7 Å CαRMSD, respectively. Significant conformational improvements of the CYANA models generated with constraint sets cutoff at 20 Å are observed both in the core and at the surface regions [Fig. 5(A)]. The explanation is that nonsequential, long distance contacts are usually involved in the interactions between surface and core regions, and thus are important for correct core packing and folding of the overlying surface.

We found a few exceptions where refined CYANA models generated by simulations at a cutoff of 8 Å showed a higher quality than those at cutoffs of 12, 16, and 20 Å. By inspecting the conformations, we found that all the initial models of these exceptions contain extended N- or C-termini. The major conformational differences between the initial models and the best CYANA models are observed in the extended N- or C-termini. For example, the best initial model of T0311 [Fig. 5(B)] is a conformation with 9.0 Å CαRMSD, whereas the best refined CYANA models generated by simulations at distance cutoffs of 8 and 20 Å are 3.6 and 5.8 Å CαRMSD, respectively. Neither the core conformation nor the C-terminus is correctly folded in the best initial model of T0311, whereas the CYANA models generated at cutoffs of both 20 and 8 Å have a correctly folded core conformation. However, the extended C-terminus in the CYANA model at a cutoff of 20 Å is not properly folded, whereas the C-terminus in the CYANA model at a cutoff of 8 Å is extended in a similar manner to that C-terminus in the experimental structure. Our previous benchmark results have shown that constraints at shorter

distance cutoffs are more accurate than those at longer distance cutoffs. For target T0311, the accuracy of constraints is 38% at the 20 Å cutoff, whereas the accuracy of constraints is 60% at the 8 Å cutoff. The above observation indicates that the inaccurate long distance constraints depreciate the restrained simulations for proteins with extended termini where long distance constraints play important roles during the folding process. Therefore, when a protein contains extended termini, its folding process is more sensitive to the inaccurate constraints of long distances. Using a shorter distance cutoff helps filter out contamination of inaccurate long distance constraints and enhances the positive effects from accurate local contacts between surface motifs, thereby improving the simulation performance and producing a conformation that is closer to the experimental structure.

In a real world prediction scenario, the best initial model or the best refined model cannot always be selected as the final conformation because of the lack of a perfect discriminatory function. Therefore, we also compared the accuracies of the best scoring initial model selected by RAPDF and the best of the top 10 scoring CYANA models by RAPDF [Fig. 4(A); right panel]. Again refinement processes using distance constraints at cutoffs of 12, 16, and 20 Å improve the initial predictions consistently, whereas the process using distance constraints at a cutoff of 8 Å does not as regularly improve the initial predictions (with higher CαRMSDs).

Dihedral angles derived from RAPDF selected models

We also investigated the effects of varying constraint cutoff distances on rTAD simulations from different dihedral angle sets. Figure 4(B) shows the refinement results using dihedral angles derived from the best scoring initial models, instead of the best initial models. Yet again the best top 10 scoring CYANA models generated at cutoffs of 12, 16, and 20 Å almost always have better CαRMSDs than the corresponding initial models. Refinement processes using dihedral angles derived from the best scoring initial models have similar results to those using dihedral angles derived from the best initial models [Fig. 4(A)], indicating that within our setup the input of dihedral angles does not greatly affect the rTAD performance, and that RAPDF is sufficient for selecting models with dihedral angles within the permitted tolerance range. In our work, dihedral angles were set with high tolerance for variation to allow the conformation to relax. Therefore, the input of dihedral angles does not affect the simulation process considerably. A further step of combining side chain modifications using SCWRL35 and minimization using ENCAD36,[37] was applied, but this does not substantially improve the quality of refined models (data not shown).

Application to CASP7 targets

To further test the effectiveness of our refinement method, the constraint selection and rTAD simulation methods were applied to 64 CASP7 targets. Distance constraints at a cutoff of 20 Å and dihedral angle tolerances compiled from the best initial models were directly input into CYANA, generating a set of conformations that satisfied the input constraints using rTAD. For each of the 64 test targets, a total of 1000 conformations were obtained, and all conformations were ranked by the RAPDF scoring function. The top 10 scoring models were used for further analysis.

Figure 6 compares the input to output for our refinement protocol. The left panel compares the accuracies of the best initial model and the best CYANA models. The right panel compares the accuracies of the best scoring initial model and the best of the top 10 scoring CYANA models selected by RAPDF. The average improvement between the best initial models and the best refined models is 0.6 Å CαRMSD (left panel), with the most substantial improvement of 4.4 Å CαRMSD. The average improvement between the best top 10 scoring refined models and the best scoring initial models is 0.9 Å CαRMSD (right panel), with 7.3 Å CαRMSD being the most substantial improvement. Our thorough benchmarks on proteins containing a variety

of folds and sizes demonstrate that our rTAD method is effective in improving the quality of protein structure prediction.

CONCLUSIONS

We demonstrate a multifaceted approach to derive constraints of sufficient quantity and quality for use in rTAD, which consistently refines initial predicted protein structures of low and high quality to an average improvement of 1.3 Å CαRMSD, ranging from 9.7 to 1.6 Å CαRMSD from the initial model relative to the corresponding experimental result. rTAD has been a promising method to refine predicted protein structures, but until now, the inability to identify accurate constraints from initial predictions prevented the use of constraint-based methods for model refinement. Specifically, the magnitude of improvement across a diverse set of target protein structures has not been produced by other constraint-based modeling methods.[14,18,19,24,25]

Our method succeeds in consistently improving the quality of protein structure predictions because a considerable number of high-accuracy distance constraints are derived and used in the process. This is achieved by combining consensus filtering, Bayesian scoring, batch-by-batch accumulation, and rTAD (each of which are explained in respective Methods sections). The consensus method preserves the interatomic distances modeled from multiple template structures or are otherwise reproducibly producible by contemporary initial modeling methods, resulting in constraint accuracy up to 45%. RAPDF is shown to be useful as a direct measure of the probability of a distance constraint being accurate, by identifying distances observed in other experimental structures within the PDB, and removing rare or nonob-served distances from damaging the modeling process. The knowledge-based atom pair potentials of RAPDF account for the differing propensities for stability of certain interatomic contact types in a protein conformation, and therefore select constraints more likely to be observed by experiment. The batch-by-batch selection method removes constraints for atom pairs mapped to residue pairs already incorporated by higher scoring atom pairs, removing weakening redundancy, yet maintaining coverage of roughly 10*N* for average-sized proteins. By altering the constraint distance cutoff, we are able to increase coverage without losing substantial accuracy; for example, when we increase the coverage above 30*N* the accuracy only drops to 34%. Thus, we are able to tune the balance between constraint coverage and accuracy. We further exploit the control of this balance by using multiple cutoffs to produce separate CYANA input constraint sets, which in turn produce a range of models allowing for unanticipated sensitivities across the range of naturally occurring proteins.

The cutoff distance used to limit the constraints influences the results of rTAD simulations (which seems to effect set accuracy and coverage), and dihedral angle tolerances derived from the RAPDF selected model are sufficient for use in CYANA model generation. In our study, 20 Å forms the maximal limit, as this is the upper limit of all versions of RAPDF.[38,39] For a globular protein, restrained simulations using constraints at longer distance cutoffs often result in conformations of higher quality. For a protein that contains extended termini, the folding process is more sensitive to inaccurate long distance constraints. Using a shorter distance cutoff helps filter out these contaminants, producing conformations that are closer to the experimental structure. In automated prediction procedures, when the detailed conformation of a protein is unknown, using a distance cutoff of 20 Å is the best choice.

Previous to this work, the best blindly assessed contact prediction method was exemplified as achieving an accuracy of only 25.5% and coverage of 3.7% for short range contacts,[29] whereas longer distance contact prediction accuracy was limited to around 20%.[30] Yet our Bayesian-guided consensus compilation produces higher accuracy with eight times more coverage (Table III). A recent constraint compilation method using beta carbons, hydrogen bonds, and nonlocal

short rage contacts, similarly using consensus of multiple models, was reported to produce constraint sets with an accuracy of 12.9% and a coverage of 49.4%, comparable to our sets.[29] Applying such a constraint set in a reasonable prediction category (fold recognition), the related model building method improved ~8 Å CαRMSD models to ~7 Å CαRMSD to the corresponding experimental conformation.[29] This method appears useful for course searches but it does not create models better than 6 Å CαRMSD for the targets shown, which is generally thought to begin the range of biologically relevant modeling. Additionally, bearing the combinatorial nature of the method, the success may be more a demonstration of increased sampling and a good discriminatory function than direct refinement. Nonetheless, this method could be successfully implemented for hard targets before the application of our own method, and in general, the constraint selection methods can be combined with our own.

Directly, no previously published paper known to us has demonstrated such consistent improvement across a wide range of initial accuracies. For the 64 CASP8 diffraction structure targets up to 400 residues in length, we obtain improvement for 60 targets (94%), with initial models ranging from 1 to 14 Å CαRMSD from the experimental structure. Refinement of a similar spread of initial model quality was recently demonstrated, which more commonly but not consistently improves the initial models.[40] Thus, amongst the methods not limited by local conformational searches, our method is unique in its ability to predictably make structures better, or no worse. More importantly, we do not propose this to be the only method by which to refine initial models, rather we hope that this will be used in concert with others, such as searching the nearby conformational space or iterations of template selection, fragment recompilation, or secondary structure determination. In particular, our method does nothing for refining alignments, which has great potential gains even amongst the models we assess in our manuscript. Ours does stand out by enabling control of the sampled space, by varying the number of input constraints; thus, we can potentially avoid limitations of available templates, fragments, and local minima.

Our work is also related to the modeling methods by Zhang *et al.*[41] and Moglich *et al.*,[42] which use a combination of rTAD and predicted constraints. The method by Moglich *et al.* focuses on dihedral angle constraints rather than interatomic distances, but was only tested on a single protein.[42] The method of Zhang *et al.* derives constraints using secondary structure and local contact prediction, yet was tested only on small helical proteins.[41] In our work, the dihedral constraints are relaxed but the distance constraints are very tight, with an average tolerance less than 0.5 Å. With this approach, we are able to achieve better estimations of the experimental structure. Although multiple methods may be combined to produce a superior approach, our method alone has proven to be effective in consistently improving the accuracy of structure prediction for a variety of proteins of different sizes and types.

The CASP7 assessors stated that only “seven cases of over 10% improvement” over the template C-alpha trace were observed for the 104 targets.[43] As shown in Figure 6, our new method achieves this level of improvement frequently: worse models are produced only rarely (4 of 64), whereas 10% improvement is achieved in the majority of cases, and greater than 1 Å improvements are made often (18 of 64). Finally, we assert that the improvements seen for these targets are significant by comparison to the best CASP7 participant group: the average accuracy of the best models submitted by the Zhang group for the 64 targets is 4.0 Å CαRMSD, whereas our initial models average 4.8 Å CαRMSD and our refinement method produces 3.5 Å CαRMSD.

In a real structure prediction scenario, this complete procedure of deriving and using distance constraints in restrained simulations improved the quality of structure predictions significantly and consistently. Compared with other methods for protein structure prediction and refinement, our method provides reliable improvement for a wide variety of conformations, because of the

novel ability to harness rTAD with a relatively huge abundance of sufficiently accurate distance constraints.

Future work

The current aspects of our approach that warrant further investigation include improved selection of input and output models; iterative cycling wherein output models would be used as input models in additional rTAD rounds; further guiding torsional tolerances using consensus amongst large clusters of models; enhancing the speed of the lengthy CYANA runs; evaluating success with input free modeling models or multiple template-based modeling techniques, extending the entire RAPDF formalism to evaluate further distant constraints, and to identify key functional constraints.[44]

METHODS

Construction and selection of initial template-based models

A total of 64 targets from the Seventh Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction (CASP7)[3] were used to evaluate the effectiveness of our refinement protocol (see Supporting Information). The selection of target proteins was based on the following: (1) The target sequence is shorter than 400 residues in length; (2) the best initial model is closer than 10 Å CαRMSD to the experimental structure; (3) at least four of the five selected initial models share a reasonable similarity between each other, with respective CαRMSDs lower than 10 Å. For benchmarking the refinement protocol parameters, we selected the top 30 targets in terms of the CαRMSD of the initial model to the corresponding experimental structures.

For each target, sequence alignments were obtained from the Bioinfo 3DJury server <http://meta.bioinfo.pl/>.45 For each alignment, one initial comparative model was generated using programs in the RAMP software suite. To build conformations for the structurally conserved regions, residues that were identical in the target and the template proteins were generated by copying atomic coordinates for the main chain and the side chain atoms, whereas residues that differed in side chain type were constructed by using a minimum perturbation technique [graph theoretic paper, JMB 1998]. To build conformations for the structurally variable regions, the programs mcgen_exhaustive_loop and mcgen_semfold_loop from the RAMP suite were used for short and long loops, respectively.46·47 The former generates conformations by exhaustively enumerating all possible main chain conformations using a 14-state φ/ψ model and selecting the best using RAPDF. The latter uses fragment replacement with simulated annealing to find the best combinations of these fragments. Additional models were also obtained from the Critical Assessment of Fully Automated Structure Prediction experiment 5 (CAFASP5)48 after examining the alignments to obtain extra variability in templates/alignments and ensuring that all residues had at least one possible conformation.

Initial model sets for each target were selected by a clustering calculation of CαRMSDs between initial models, which measures the similarity between initial predictions. A cutoff of 10 Å CαRMSD was used to filter out those predictions, which were very different from other available initial models. These models were further filtered using RAPDF. From a collection of initial models, five best ranked initial models with reasonable CαRMSDs between each other were selected for deriving the consensus distance constraints and dihedral angles. The CαRMSDs of the best scoring selected initial models to the corresponding experimental structures range from 1.6 to 9.7 Å; these are noted throughout the figures as “best scoring initial models.”

In addition to testing constraints derived from models selected by RAPDF, to assess the abilities of the refinement method within the idealized case of optimal decoy selection, we test constraints derived from the closest models to the experimental conformation, which we identify by calculating the CαRMSDs of each initial model to the experimental structure. These constraint sets are noted throughout the figures as "best initial models."

Constraint selection by consensus and RAPDF

For each of the five selected initial protein models, distances between two nonlocal atoms (separated by more than four residues) were measured and binned in 0.5 Å increments (see Methods: residue-specific all-atom probability discriminatory function). For a given nonlocal interatomic contact, if the distances observed across the five initial models all fell within the same distance bin, the contact was ascribed to the $consensus_5$ group; if four of the five observed distances fell within the same distance bin, the contact was ascribed to the $consensus_4$ group. The average of the consensus distances was calculated for use as the consensus contact distance.

The RAPDF scoring function was novelly implemented here to calculate the probability of a nonlocal contact occurring within a given distance bin. Directly, we extended the application of RAPDF to evaluate one constraint at a time instead of the traditional use to evaluate whole conformations. This was enabled by the direct relevance of the formalism, to evaluate the occurrence of a given atom type pair at a given distance within a non-redundant subset of the Protein Data Bank (PDB)[49] versus random. Thus, each of the consensus contacts ($consensus_5$ and $consensus_4$) was assigned with a RAPDF score, $S(d_{ab})$. Top RAPDF ranked consensus contacts were selected for further analysis using the batch-by-batch selection method described later.

Dihedral angles were derived from the same models used to produce the interatomic distance sets. The φ/ψ parameters for each residue were obtained from the DSSP [50] calculation of initial models. For residues in helices φ/ψ parameters were each derived from the DSSP calculation ±15°; for residues in sheets, φ/ψ parameters were derived from DSSP calculations ±30°; for residues in variable regions, no dihedral restriction was assigned.

Batch-by-batch selection of consensus constraints

We considered both top ranked $consensus_5$ and $consensus_4$ constraints to be similar candidate sets, such that assigned tolerance was the only differentiation indicated. Further selection was performed in a batch-by-batch manner. The first batches of $consensus_5$ and $consensus_4$ constraints were selected according to the RAPDF ranking. The next batch of constraints was selected from the next best ranked constraints by filtering out those constraints for which the corresponding residue pair was already represented by another atom pair constituent to a common residue pair found in a previous batch. This was done in an effort to eliminate the constraints of lower accuracy. Sequential batch selection iterations continued until the residue pairs represented by the selected constraints obtained 10% of all residue pairs in each protein sequence.

The tolerance of each constraint was calculated using the standard deviation (SD) between the distances observed in different initial models. The tolerances were calculated as: Tolerance of $consensus_5$ = 0.1*(2 SD + 1.4 + (0.2*batch_number)) and Tolerance of $consensus_4$ = 0.1*(2 SD + 1.8 + (0.2*batch_number)). Therefore, the tolerance for constraints of lower ranking and less consensus (assumed to have lower accuracy) was set higher than those of higher rank and consensus.

rTAD simulations using CYANA

The distance constraints and dihedral angles were directly input to CYANA, which generated a set of conformations satisfying the input constraints using rTAD. For each parameter combination, a total of 1000 conformations were obtained. All conformations were ranked by the RAPDF scoring function. The top 10 scoring models were used for further analysis.

Residue-specific all-atom probability discriminatory function

A complete description of the RAPDF formalism and benchmark is found in the original article, 38 and subsequent study examining effects of parameterization set quality.[39] In summary, we make observations of all nonlocal interatomic distances on a dataset of experimentally determined structures. The conditional probabilities are compiled by counting frequencies of distances between pairs of atom types in a dataset of protein structures. All nonhydrogen atoms are considered and a residue-specific description of the atoms was used (e.g., the Cα of alanine is different from the Cα of glycine). This results in a total of 167 atom types. The interatomic distances observed are divided into 0.5 Å bins ranging from 2 to 20 Å. Contacts between atom types in the 0–2 Å range are placed in a separate bin, resulting in a total of 37 distance bins.

The scores $S(d_{ab})$ proportional to the negative log conditional probability of observing a native conformation given an interatomic distance of $P(C|d_{ab})$ are compiled according to the formula: $S(d_{ab}) = -\ln \frac{P(d_{ab}|C)}{P(d_{ab})} \propto -\ln P(C|d_{ab})$, where $P(d_{ab}|C)$ is the probability of observing a distance d between two atom types a and b in a correct conformation, and $P(d_{ab})$ is the probability of observing such a distance, d_{ab}, in any conformation, correct or incorrect.

For a minimally redundant dataset of high-quality experimental structures (sequence identity <30%, resolution < 2.1 Å CαRMSD), the counts of d_{ab} observations in each structure are summed to generate an overall probability. Tables of scores $S(d_{ab})$ for all possible pairs of the 167 atom types for the 37 distance ranges are compiled from a database of known structures.

Given an amino acid sequence in a particular conformation, the scores of all contacts between atom type pairs that fall within the distance cutoff are summed to yield the total "pseudo-potential" score for evaluating the probability of a conformation being native-like. The formula is $S_{(\text{conformation})} = -\sum \ln \frac{P(d_{ab}^{ij}|C)}{P(d_{ab}^{ij})} \propto -\ln P(C|d_{ab}^{ij})$, where i and j are atom indexes.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

REFERENCES

1. Moult J. Predicting protein three-dimensional structure. Curr Opin Biotechnol 1999;10:583–588. [PubMed: 10600698]
2. Schonbrun J, Wedemeyer WJ, Baker D. Protein structure prediction in 2002. Curr Opin Struct Biol 2002;12:348–354. [PubMed: 12127454]
3. Moult J. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. Curr Opin Struct Biol 2005;15:285–289. [PubMed: 15939584]
4. Krieger, E.; Nabuurs, SB.; Vriend, G. Homology modeling. In: Bourne, PE.; Weissig, H., editors. Structural bioinformatics. Hoboken, NJ: Wiley-Liss; 2003. p. 509-523.
5. Chakravarty S, Wang L, Sanchez R. Accuracy of structure-derived properties in simple comparative models of protein structures. Nucleic Acids Res 2005;33:244–259. [PubMed: 15647507]

6. Cozzetto D, Tramontano A. Relationship between multiple sequence alignments and quality of protein comparative models. Proteins 2005;58:151–157. [PubMed: 15495137]
7. Bradley P, Misura KM, Baker D. Toward high-resolution de novo structure prediction for small proteins. Science 2005;309:1868–1871. [PubMed: 16166519]
8. Valencia A. Protein refinement: a new challenge for CASP in its 10th anniversary. Bioinformatics 2005;21:277. [PubMed: 15647301]
9. de Bakker PI, Furnham N, Blundell TL, DePristo MA. Conformer generation under restraints. Curr Opin Struct Biol 2006;16:160–165. [PubMed: 16483766]
10. Sali A, Overington JP, Johnson MS, Blundell TL. From comparisons of protein sequences and structures to protein modelling and design. Trends Biochem Sci 1990;15:235–240. [PubMed: 2200167]
11. Fiser A, Sali A. Modeller: generation and refinement of homology-based protein structure models. Methods Enzymol 2003;374:461–491. [PubMed: 14696385]
12. Schueler-Furman O, Wang C, Bradley P, Misura K, Baker D. Progress in modeling of protein structures and interactions. Science 2005;310:638–642. [PubMed: 16254179]
13. Havel TF, Kuntz ID, Crippen GM. The combinatorial distance geometry method for the calculation of molecular conformation. I A new approach to an old problem. J Theor Biol 1983;104:359–381. [PubMed: 6656266]
14. Aszodi A, Gradwell MJ, Taylor WR. Global fold determination from a small number of distance restraints. J Mol Biol 1995;251:308–326. [PubMed: 7643405]
15. Aszodi A, Munro RE, Taylor WR. Protein modeling by multiple sequence threading and distance geometry. Proteins 1997;29:38–42. [PubMed: 9485493]
16. Hanggi G, Braun W. Pattern recognition and self-correcting distance geometry calculations applied to myohemerythrin. FEBS Lett 1994;344:147–153. [PubMed: 8187874]
17. Mumenthaler C, Braun W. Predicting the helix packing of globular proteins by self-correcting distance geometry. Protein Sci 1995;4:863–871. [PubMed: 7663342]
18. 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. [PubMed: 9020984]
19. 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. [PubMed: 9514747]
20. Guntert P, Mumenthaler C, Wuthrich K. Torsion angle dynamics for NMR structure calculation with the new program DYANA. J Mol Biol 1997;273:283–298. [PubMed: 9367762]
21. Ramachandran GN, Sasisekharan V. Conformation of polypeptides and proteins. Adv Protein Chem 1968;23:283–438. [PubMed: 4882249]
22. Smith-Brown MJ, Kominos D, Levy RM. Global folding of proteins using a limited number of distance constraints. Protein Eng 1993;6:605–614. [PubMed: 7694274]
23. Kolinski A, Skolnick J. Assembly of protein structure from sparse experimental data: an efficient monte carlo model. Proteins 1998;32:475–494. [PubMed: 9726417]
24. Huang ES, Samudrala R, Ponder JW. Distance geometry generates native-like folds for small helical proteins using the consensus distances of predicted protein structures. Protein Sci 1998;7:1998–2003. [PubMed: 9761481]
25. Huang ES, Samudrala R, Ponder JW. Ab initio fold prediction of small helical proteins using distance geometry and knowledge based scoring functions. J Mol Biol 1999;290:267–281. [PubMed: 10388572]
26. Qian B, Raman S, Das R, Bradley P, McCoy AJ, Read RJ, Baker D. High-resolution structure prediction and the crystallographic phase problem. Nature 2007;450:259–264. [PubMed: 17934447]
27. Hung LH, Samudrala R. An automated assignment-free Bayesian approach for accurately identifying proton contacts from NOESY data. J Biomol NMR 2006;36:189–198. [PubMed: 17016668]
28. Young MM, Tang N, Hempel JC, Oshiro CM, Taylor EW, Kuntz ID, Gibson BW, Dollinger G. High throughput protein fold identification by using experimental constraints derived from intramolecular cross-links and mass spectrometry. Proc Natl Acad Sci USA 2000;97:5802–5806. [PubMed: 10811876]

29. Latek D, Kolinski A. Contact prediction in protein modeling: scoring, folding and refinement of coarse-grained models. BMC Struct Biol 2008;8:36. [PubMed: 18694501]
30. Grana O, Baker D, MacCallum RM, Meiler J, Punta M, Rost B, Tress ML, Valencia A. CASP6 assessment of contact prediction. Proteins 2005;61:214–224. [PubMed: 16187364]
31. Godzik A. The structural alignment between two proteins: is there a unique answer? Protein Sci 1996;5:1325–1338. [PubMed: 8819165]
32. Wallner B, Elofsson A. All are not equal: a benchmark of different homology modeling programs. Protein Sci 2005;14:1315–1327. [PubMed: 15840834]
33. Kraulis P. Molscript: a program to produce both detailed and schematic plots of protein structures. J Appl Crystallogr 1991;24:946–950.
34. Merritt EA, Bacon DJ. Raster3D: photorealistic molecular graphics. Methods Enzymol 1997;277:505–524. [PubMed: 18488322]
35. Bower MJ, Cohen FE, Dunbrack RL Jr. Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool. J Mol Biol 1997;267:1268–1282. [PubMed: 9150411]
36. Park BH, Huang ES, Levitt M. Factors affecting the ability of energy functions to discriminate correct from incorrect folds. J Mol Biol 1997;266:831–846. [PubMed: 9102472]
37. Levitt M, Hirshberg M, Sharon R, Daggett V. Potential energy function and parameters for simulations of the molecular dynamics of proteins and nucleic acids in solution. Comput Phys Commun 1995;91:215–231.
38. Samudrala R, Moult J. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. J Mol Biol 1998;275:895–916. [PubMed: 9480776]
39. Liu T, Samudrala R. The effect of experimental resolution on the performance of knowledge based discriminatory functions for protein structure selection. Protein Eng Des Sel 2006;19:431–437. [PubMed: 16845128]
40. Jagielska A, Wroblewska L, Skolnick J. Protein model refinement using an optimized physics-based all-atom force field. Proc Natl Acad Sci USA 2008;105:8268–8273. [PubMed: 18550813]
41. Zhang C, Hou J, Kim SH. Fold prediction of helical proteins using torsion angle dynamics and predicted restraints. Proc Natl Acad Sci USA 2002;99:3581–3585. [PubMed: 11904420]
42. Moglich A, Weinfurtner D, Maurer T, Gronwald W, Kalbitzer HR. A restraint molecular dynamics and simulated annealing approach for protein homology modeling utilizing mean angles. BMC Bioinformatics 2005;6:91.
43. Moult J, Fidelis K, Kryshtafovych A, Rost B, Hubbard T, Tramontano A. Critical assessment of methods of protein structure prediction-Round VII. Proteins 2007;69:3–9. [PubMed: 17918729]
44. Wang K, Horst JA, Cheng G, Nickle DC, Samudrala R. Protein meta-functional signatures from combining sequence, structure, evolution, and amino acid property information. PLoS Comput Biol 2008;4:e1000181. [PubMed: 18818722]
45. Ginalski K, Elofsson A, Fischer D, Rychlewski L. 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics 2003;19:1015–1018. [PubMed: 12761065]
46. Samudrala R, Moult J. Handling context-sensitivity in protein structures using graph theory: bona fide prediction. Proteins 1997:43–49. [PubMed: 9485494]
47. Hung LH, Ngan SC, Liu T, Samudrala R. Protinfo: new algorithms for enhanced protein structure predictions. Nucleic Acids Res 2005;33:W77–W80. [PubMed: 15980581]
48. Fischer D, Rychlewski L, Dunbrack RL Jr, Ortiz AR, Elofsson A. CAFASP3: the third critical assessment of fully automated structure prediction methods. Proteins 2003;53:503–516. [PubMed: 14579340]
49. Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Bur-khardt K, Feng Z, Gilliland GL, Iype L, Jain S, Fagan P, Marvin J, Padilla D, Ravichandran V, Schneider B, Thanki N, Weissig H, Westbrook JD, Zardecki C. The protein data bank. Acta Crystallogr D Biol Crystallogr 2002;58:899–907. [PubMed: 12037327]
50. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637. [PubMed: 6667333]

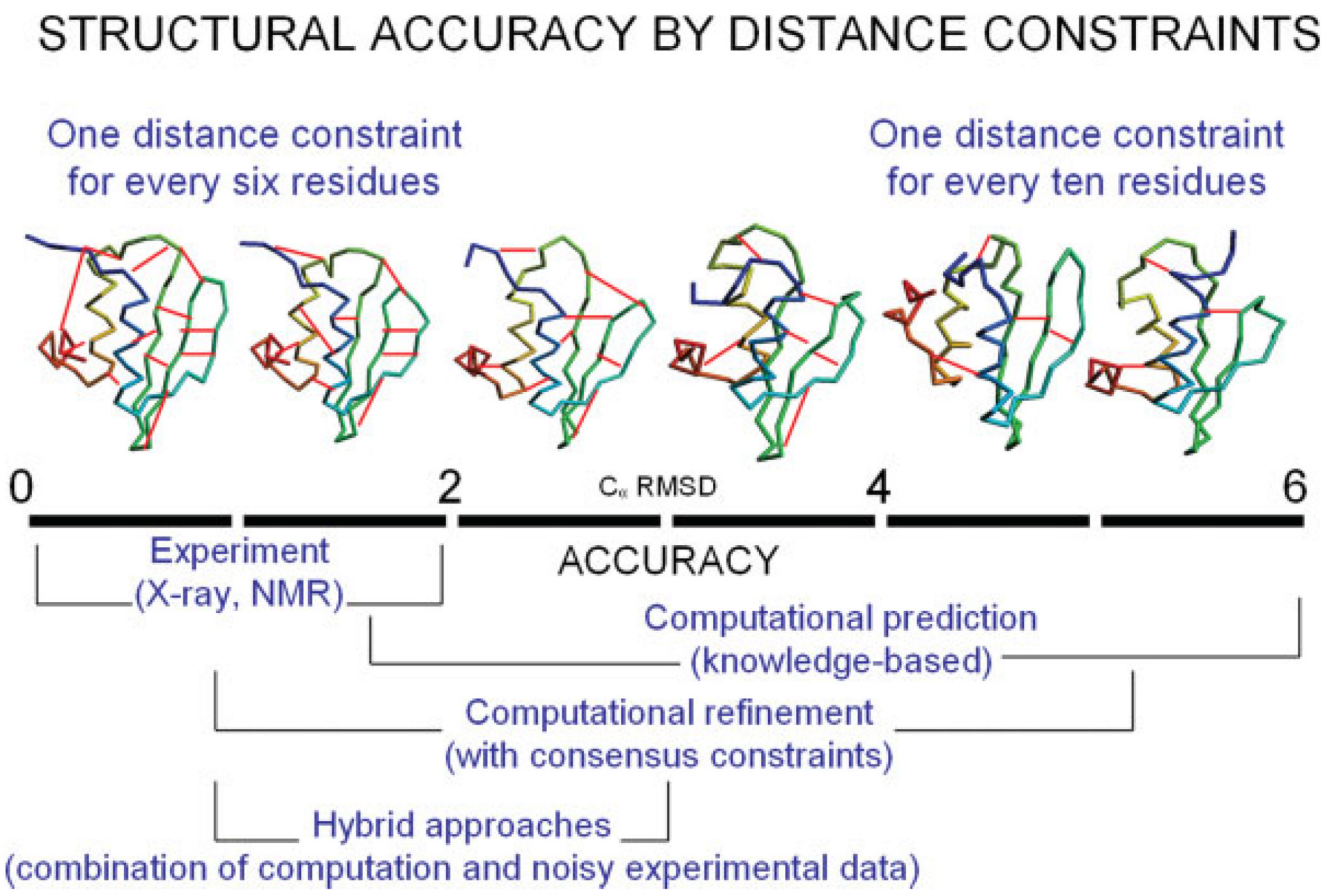

Figure 1.

Accuracy of protein structure models as a function of native-like constraints obtained. Extremely accurate models are generated when protein structure prediction methods are combined with various methods for deriving interatomic distance constraints. Our estimation here is based on a survey of such approaches (see introduction for further detail). These constraints are not only obtainable from bench methods such as NMR,[26] Bayesian interpretation of noisy NMR data,27 or crosslinkers used as molecular rulers combined with mass spectrometry,28 but also from consensus of predictive models as presented in this work. Low-resolution models are used to derive hypotheses and guide bench experiments, whereas the highest resolution models are used to enable detailed functional studies and therapeutic discovery. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

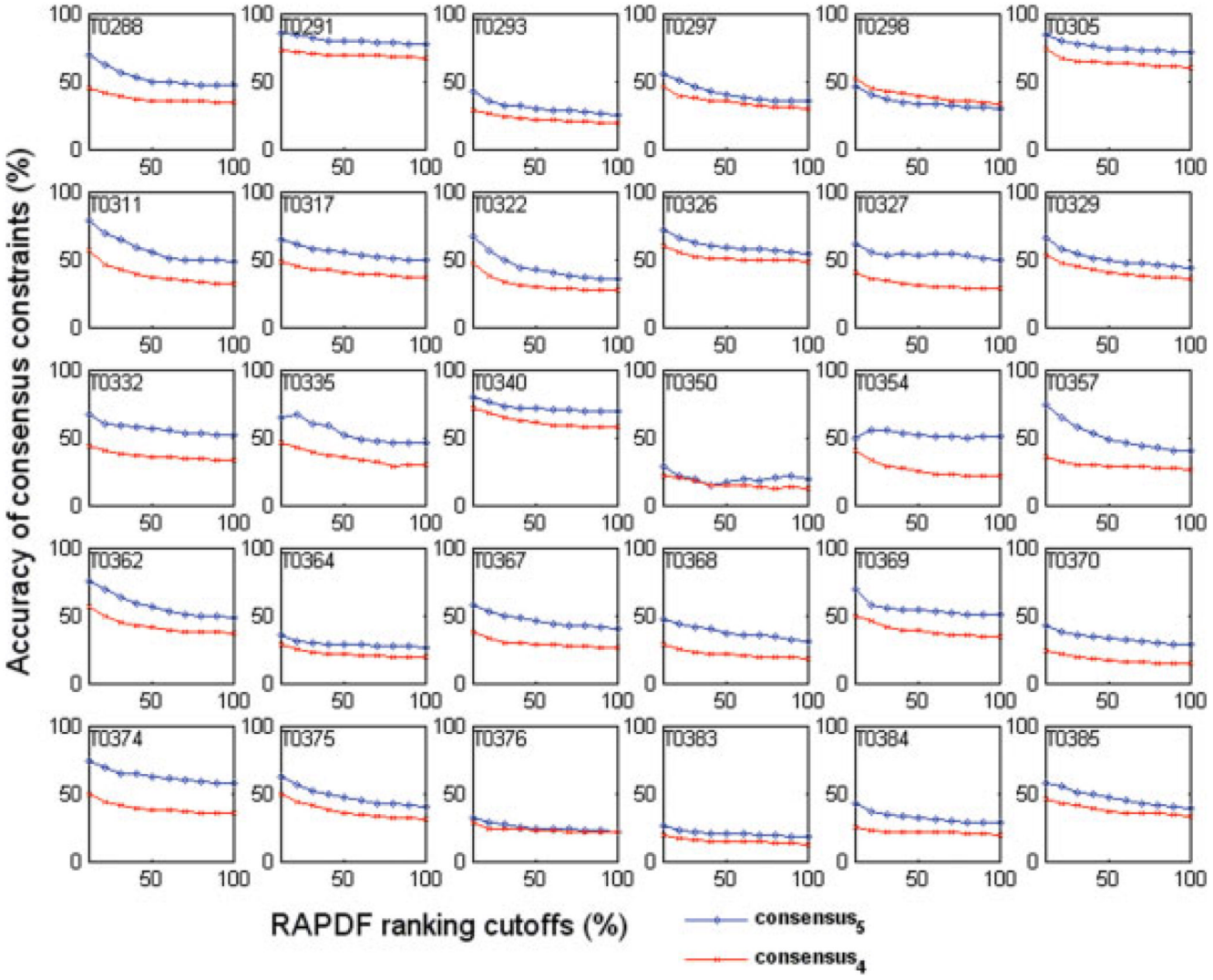

Figure 2.

The accuracy of consensus constraints correlates with the RAPDF ranking. For each of the 30 benchmark targets, 10 subsets of consensus contacts were derived according to their RAPDF rank cutoffs, as set 10%, set 20%, … set 100%. The consensus constraint accuracy is defined as Accuracy = [The number of nonlocal contacts for which the selected consensus distance constraint falls within 0.25 Å of the correct distance]/[Total nonlocal contacts]. These plots show that the consensus constraint accuracy is better for higher RAPDF ranking cutoffs, indicating that RAPDF scores can be used as a standard for selecting constraints of higher accuracy. In addition, these plots show that $consensus_5$ is almost always more accurate than $consensus_4$. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

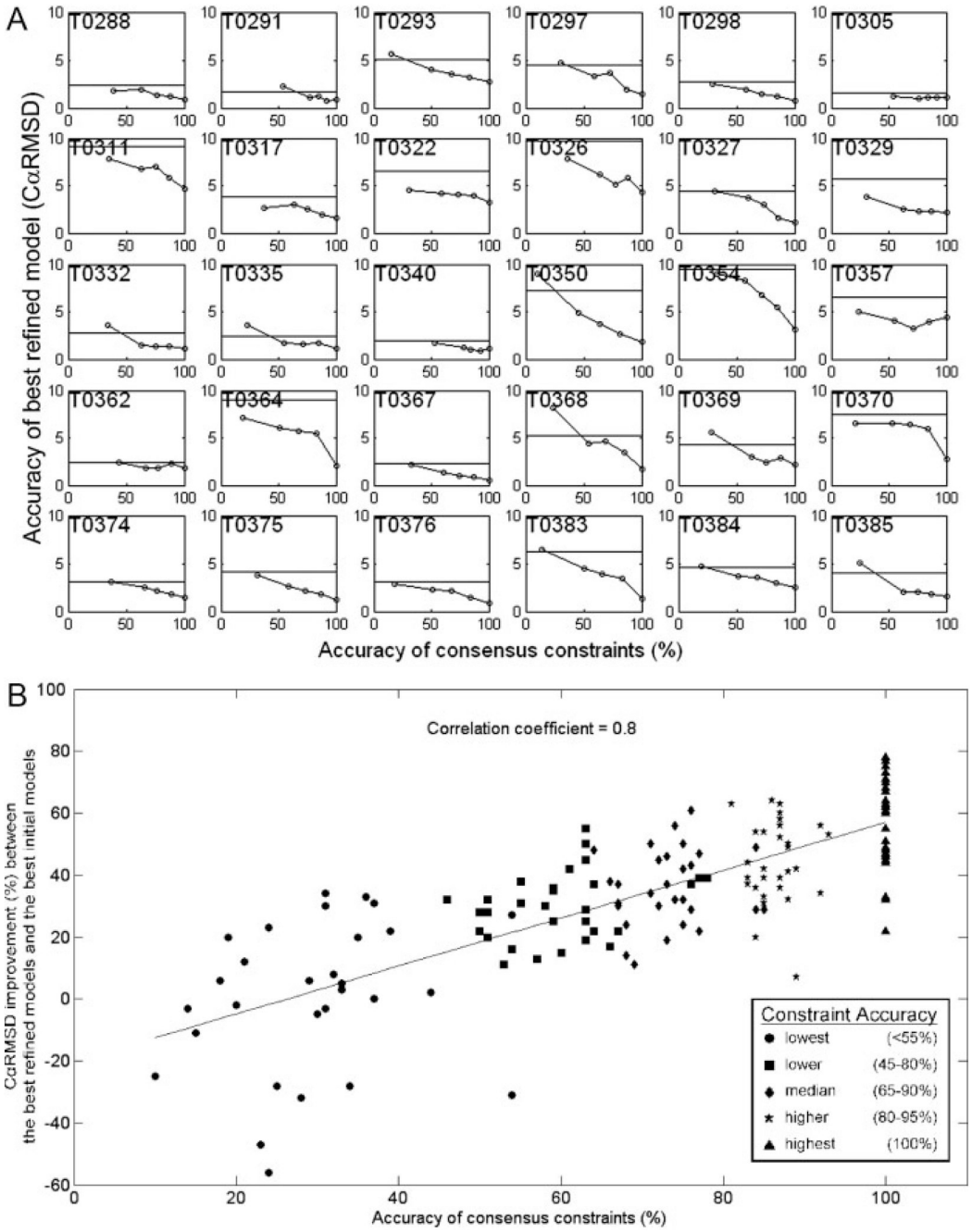

Figure 3.

(**A**) The relationship between the accuracy of consensus constraints and the effectiveness of rTAD simulations. The qualities of the best refined models generated by simulations using five sets of constraints with different accuracies, evaluated by calculating CαRMSDs to the corresponding experimental structures. The CαRMSD of the best initial model is indicated by a line parallel to the horizontal axis. In all 30 tested target proteins, the quality of refined models is improved when the constraints of higher accuracy are used in rTAD simulations, indicating the importance of obtaining higher quality constraints. (**B**) The relationship between the accuracy of input constraints and the effectiveness of the rTAD simulations. Five constraint sets of differing accuracy were developed for each of 30 targets. The five best conformations

generated from each constraint set are taken together as a pooled sample. The magnitude of improvement from the best initial model to the best refined model correlates with the accuracy of the consensus constraint sets used in rTAD simulations. The Spearman correlation coefficient of this relationship is 0.8.

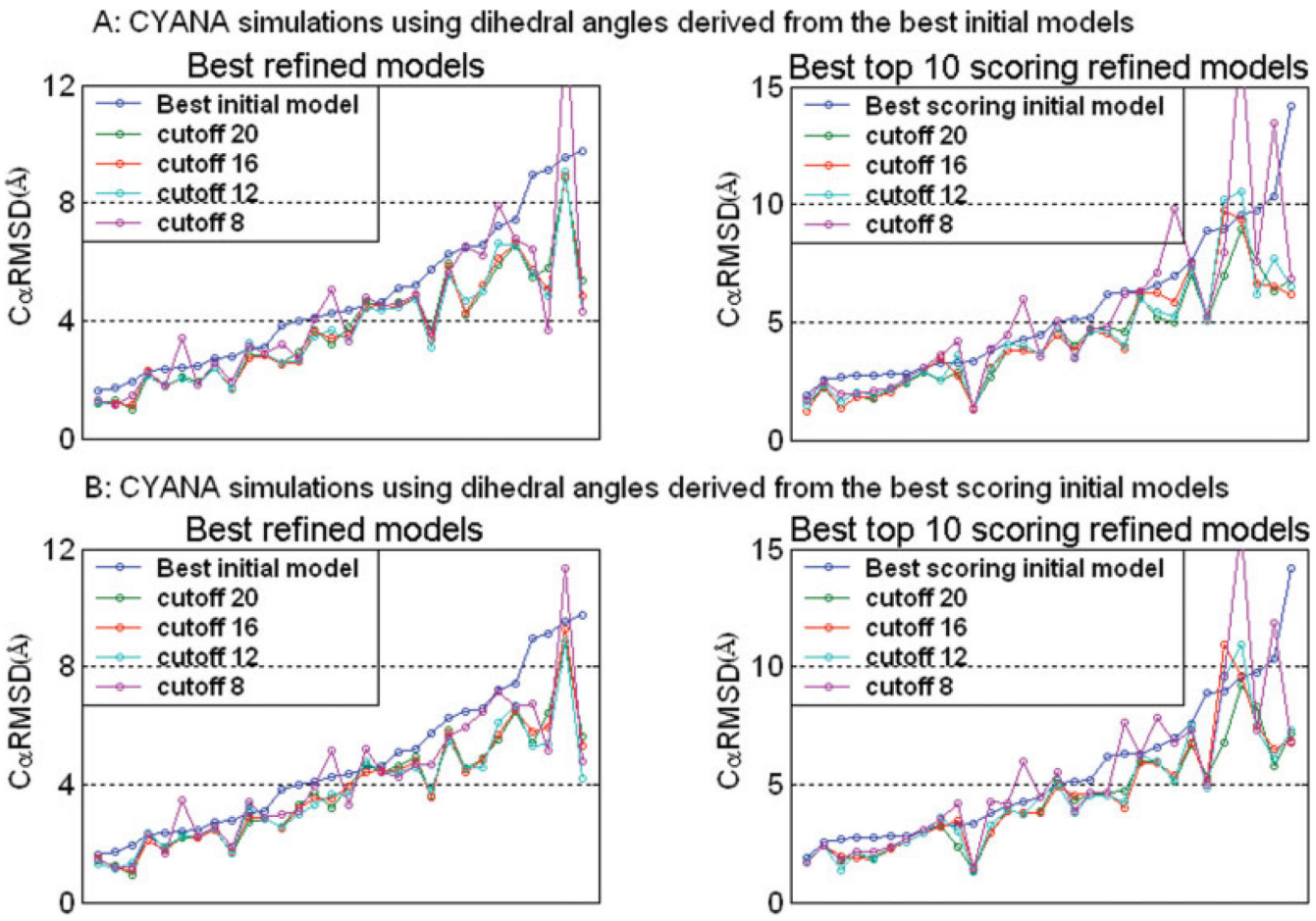

Figure 4.

The influence of distance cutoffs and dihedral angles on the rTAD simulations. The results of the simulations were evaluated by calculating CαRMSDs of the refined models to the corresponding experimental structures. A total of 30 targets from CASP7 were tested. Simulations were performed using consensus constraints at distance cutoffs of 8, 12, 16, and 20 Å, respectively. The dihedral angles were compiled from the best initial models (row A) or the best scoring initial models (row B). The left panel compares the accuracies of the best initial model to the best refined models. The right panel compares the accuracies of the best scoring initial model selected by RAPDF to the best of the top 10 scoring refined models. For all 30 targets, the best refined models generated at cutoffs of 12 Å (cyan line), 16 Å (red line), and 20 Å (green line) always show better quality (lower CαRMSDs) than the corresponding best initial models (blue line). For almost all targets, the best top 10 scoring refined models generated at cutoffs of 12, 16, and 20 Å are more accurate than the corresponding best scoring initial models. Simulations using constraints of different distance cutoffs (12, 16, and 20 Å) show similar improvement, indicating that the performance of restrained simulations results from the trade off between constraint set accuracy and coverage. Simulations using constraints at a cutoff of 8 Å are not as effective as simulations at cutoffs of 12, 16, and 20 Å The performance of simulations using dihedral angles derived from different initial models is similar, indicating that the input parameters of dihedral angles do not substantially affect the results of rTAD simulations in our protocol.

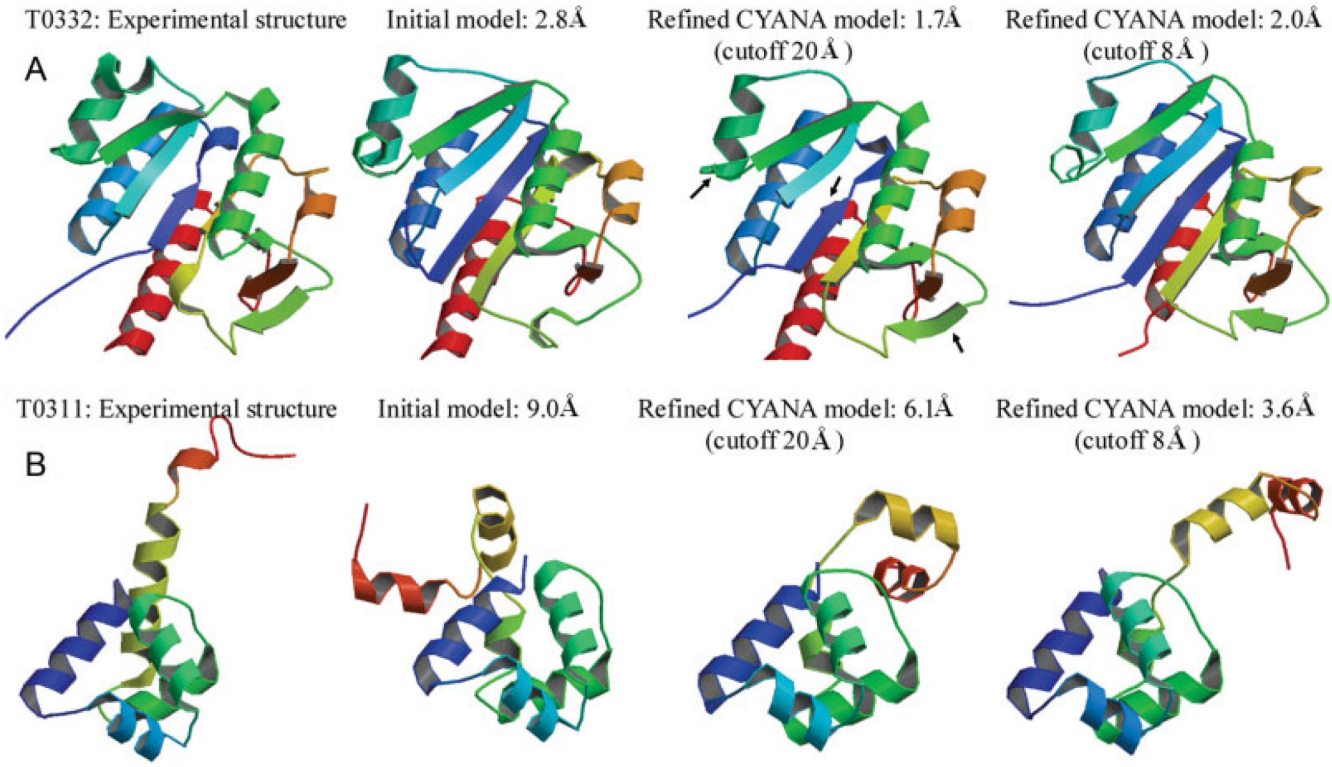

Figure 5.

(A) The experimental structure, the best initial model, and refined models of the T0332 protein. The initial model of T0332 is a conformation with a 2.8 Å CαRMSD, whereas the best refined models generated by simulations at the distance cutoffs of 20 and 8 Å are 1.7 and 2.0 Å CαRMSD, respectively. The significant conformational improvement of the refined model generated at a cutoff of 20 Å is found both in the core and at the surface regions (marked by three dark arrows). This indicates that long distance constraints influence the interactions between surface and core regions and thus are important for the correct folding of the surface and the core. (Figures were prepared with Molscript33 and Raster3D.34) (**B**) The experimental structure, the initial model, and CYANA models of the T0311 protein. The best initial model of T0311 is a conformation with 9.0 Å CαRMSD, whereas the best refined models generated by simulations at the distance cutoffs of 20 and 8 Å are 5.8 and 3.6 Å CαRMSD, respectively. Neither the core conformation nor the C-terminus is correctly folded in the best initial model of T0311. The refined CYANA models generated at the 20 and 8 Å cutoffs both present a correctly folded core conformation. The difference between these two models is at the C-terminus. The extended C-terminus in the CYANA model at the 20 Å cutoff is incorrectly folded, whereas the extended C-terminus in the CYANA model at the 8 Å cutoff is closer to the experimental structure. This indicates that inaccurate long distance constraints depreciate the quality of the rTAD simulations for proteins with extended termini in cases where long distance constraints play important roles during the folding process. The use of accurate shorter distance cutoffs helps to filter out contamination from inaccurate long distance constraints, producing a conformation that is closer to the experimental structure. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

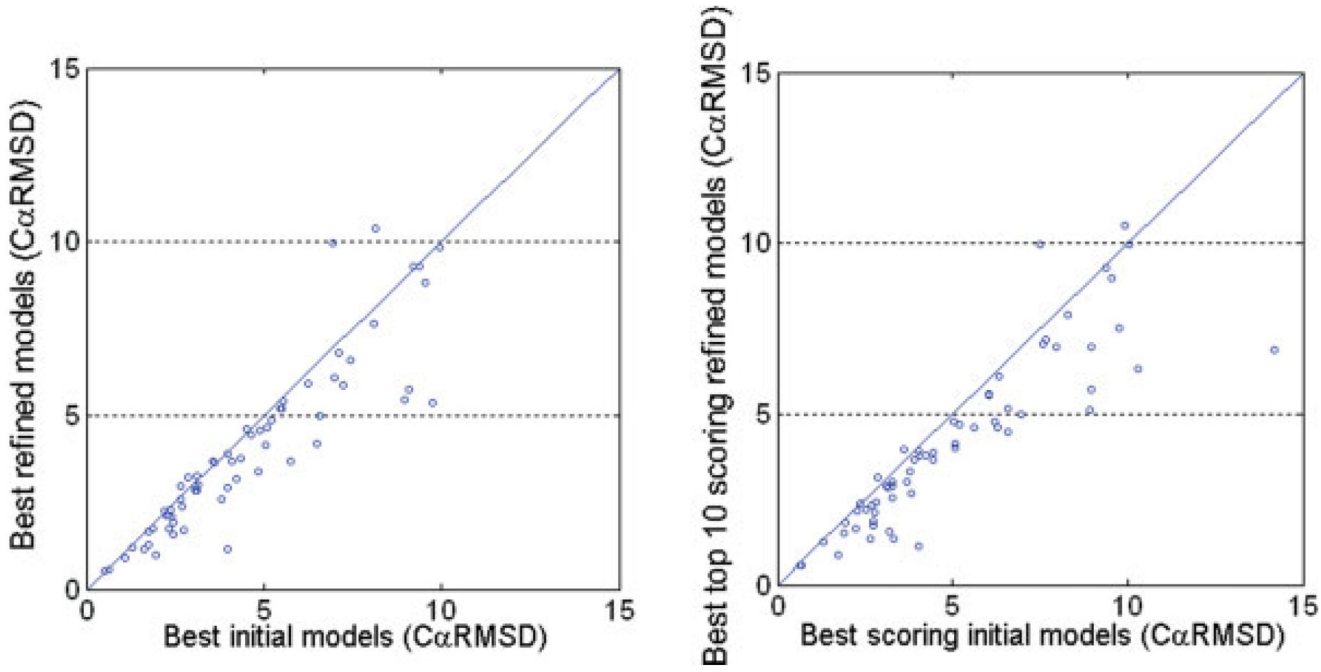

Figure 6.

Improvement of the refined CYANA models by using restrained torsion angle dynamics simulations with consensus constraints. For all 64 targets, the accuracies of the initial predictions and the refined CYANA models were evaluated with CαRMSD model measurements. The CαRMSDs of the initial models to the experimental structure were compared with those of the refined models. The left panel compares the accuracies of the best initial model to the corresponding best refined models. The right panel compares the accuracies of the best scoring initial model selected by RAPDF to the corresponding best of the top 10 RAPDF scoring refined models. The average improvement between the best refined models and the best initial models is 0.6 Å CαRMSD (left panel), with the most substantial improvement of 4.4 Å CαRMSD. The average improvement between the best top 10 scoring refined models and the best scoring initial models is 0.9 Å CαRMSD (right panel), with the most substantial improvement of 7.3 Å CαRMSD. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

Table I

The Relationship Between the Accuracy of Consensus Constraints and the RAPDF Ranking

RAPDF rank	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
$Consensus_5$	59%	54%	51%	48%	47%	46%	45%	44%	43%	42%
$Consensus_4$	44%	40%	37%	36%	34%	33%	33%	32%	32%	31%

Ten subsets of consensus constraints were derived according to their RAPDF ranking cutoffs. The accuracy of distance constraints $consensus_5$ and $consensus_4$ in each subset was calculated. The accuracy of consensus constraints correlates directly with the RAPDF ranking cutoffs. The accuracy of $consensus_5$ is always higher than that of $consensus_4$.

Table II

The Relationship Between the Accuracy of Consensus Constraints and the Improvement of the Best CYANA Models in Terms of Their CαRMSD to Their Experimental Structures

Constraint sets	Set-1	Set-2	Set-3	Set-4	Set-5
Average accuracy of constraint sets	30%	60%	73%	87%	100%
Average CαRMSD of the best CYANA models (Å CαRMSD)	4.5	3.4	3.1	2.7	1.9
Average improvement of the best CYANA models (Å CαRMSD)	0.3	1.4	1.7	2.1	2.8

Five sets of constraints of different accuracies were input into CYANA. The best refined CYANA model generated in each set of restrained simulations was compared with the best initial model of that protein by calculating their CαRMSDs to the corresponding experimental structure. The average accuracy of each set of constraints was then calculated and compared with the original average accuracy of 4.8 Å. Increasing accuracy of constraints (lowest to highest sets) improves the quality of the best refined models, corresponding to average CαRMSDs from 4.5 to 1.9 Å. The average improvement of the best refined models to the best initial models is only 0.3 Å CαRMSD when using constraints of lower accuracy (lowest set), whereas using constraints of higher accuracy (higher and highest sets), the average improvement is 2 to 3 Å CαRMSD. Thus, the quality of the refined models is improved when consensus constraints of higher accuracy are used in rTAD simulations.

Table III

The Effect of Spatial Constraints on the Restrained Torsion Angle Dynamics Simulations Is a Trade Off Between the Accuracy and the Coverage

Cutoffs	Cutoff 8	Cutoff 12	Cutoff 16	Cutoff 20
Constraints accuracy	45%	38%	36%	34%
Constraints coverage	7%	16%	24%	30%
Simulations using dihedral angles derived from best initial models				
Improvement of the best CYANA models (Å CαRMSD)	0.6	1.1	1.1	1.1
Improvement of the best top 10 scoring CYANA models (Å CαRMSD)	0.2	1.1	1.2	1.3
Simulations using dihedral angles derived from best scoring initial models				
Improvement of the best CYANA models (Å CαRMSD)	0.6	1.1	1.1	1.0
Improvement of the best top 10 scoring CYANA models (Å CαRMSD)	0.2	1.1	1.2	1.3

The rTAD simulations were performed using consensus constraints at distance cutoffs of 8, 12, 16, and 20 Å, and various input sources. For each subset of constraints, the average accuracy and coverage were calculated. As the distance cutoff lengthens from 8, 12, 16 to 20 Å, the accuracy of constraints increases from 7, 16, 24 to 30%; the improvement of the best refined model increases from 0.6 to 1.1 Å CαRMSD; and the improvement of the best top 10 scoring refined model increases from 0.2 to 1.3 Å CαRMSD. Simulations using constraints of different distance cutoffs (12, 16, and 20 Å) show similar improvement, suggesting that the efficacy of restrained simulations results from the trade off of the constraints accuracy and the constraints coverage. Improvement by the restrained simulations using dihedral angles derived from different initial models is the same, indicating that the input parameters of dihedral angles do not significantly affect the results of rTAD simulations.