Modeling large regions in proteins: Applications to loops, termini and folding

Aashish N. Adhikari[1,3], Jian Peng[6], Michael Wilde[4], Jinbo Xu[6], Karl F. Freed[1,3,4] and Tobin R. Sosnick[2,4,5]
[1]Dept. of Chemistry, [2]Dept. of Biochemistry and Molecular Biology, [3] The James Franck Inst., [4]Computation Inst., [5]Inst. for Biophysical Dynamics, The University of Chicago; [6]Toyota Technology Inst. at Chicago.

Corresponding authors:

Prof. Tobin R. Sosnick
Dept. of Biochemistry & Molecular Biology
Institute for Biophysical Dynamics
Computation Institute
The University of Chicago
Chicago, Il 60637
(773)218-5950; (773)702-0439 (fax)
trsosnic@uchicago.edu,

Prof. Karl F. Freed,
Dept. of Chemistry
The James Frank Institute
Computation Institute
The University of Chicago
Chicago, Il 60637
(773)702-7202
freed@uchicago.edu

ABSTRACT

Template-based methods for predicting protein structure provide models for a significant portion of the protein but often contain insertions or chain ends ("InsEnds") of indeterminate conformation. The local structure prediction "problem" entails modeling the InsEnds onto the rest of the protein. A well known limit involves predicting loops of ≤ 12 residues in crystal structures. InsEnds, however, may contain as many as ~ 50 amino acids, and the template-based model of the protein itself may be imperfect. To address these challenges, we present a free modeling method for predicting the local structure of loops and large InsEnds in both crystal structures and template-based models. The approach uses single amino acid torsional angle "pivot" moves of the protein backbone with a C_β level representation. Nevertheless, our accuracy for loops is comparable to existing methods. We also apply a more stringent test, the blind structure prediction and refinement categories of the CASP9 tournament, where we improve the quality of several homology based models by modeling InsEnds as long as 45 amino acids, sizes generally inaccessible to existing loop prediction methods. Our approach ranks as one of the best groups in the CASP9 refinement category that involves improving template-based models so that they can function as molecular replacement models to solve the phase problem for crystallographic structure determination.

KEYWORDS: long loops, insertions, loop modeling, local protein structure prediction, molecular replacement

ABBREVIATIONS: amino acid, aa; Monte Carlo Simulated Annealing, MCSA; Ramachandran, Rama, nearest neighbor (NN), Solvent accessible surface area, SASA;

INTRODUCTION

Homology-based methods use known structures as templates and have proven extremely successful in modeling larger proteins in a computationally efficient fashion. The success of these methods, however, depends on the quality of the alignments between the target sequence and those of the templates [1]. Frequently, the sequence alignments contain gaps that correspond to regions in the sequence where no reliable structural information can be extracted from the templates. These gaps may be insertions or additions at the termini (Fig. 1). Inevitably, the model assembled from the templates lacks these local regions. In order to model the entire structure, alternative methods are required. The problem of reconstructing local regions in a protein is neither new nor exclusive to homology modeling. Experimentally determined structures from crystallography often contain regions that are difficult to characterize because they are flexible or mobile. Consequently, crystal structures can contain loops that have weak or missing electron density. This issue is particularly significant because protein function is often mediated by loops; for example, loops often act as molecular recognition or binding sites and play a crucial role in executing the protein's function [2-4]. The specificity of protein interactions as mediated by active sites and binding pockets is also a consequence of local protein structure. These issues highlight the need for reliable methods to reconstruct local regions in protein structures.

Three important problems arise in developing methods for predicting local spatial structure. First, the local regions must be modeled subject to the constraints imposed by the rest of the protein structure. , e.g., the loop termini must end at the correct anchor positions. Some approaches to this long-standing "loop closure problem" seek analytical solutions to bond angles that properly position the ends [5-7]. While exact solutions have been found for short polypeptide

segments, no general analytical solution is possible for segments containing more than a few amino acids in proteins. Other robotics-inspired algorithms for loop closure [8,9] likewise experience decreasing accuracy as the size of the loops increases. Additionally, analytical approaches to the closure problem very often yield solutions that place backbone dihedral angles in disallowed regions of Ramachandran (Rama) space and thus generate sterically forbidden conformations.

Second, irrespective of how the loop closure is performed, a procedure is required for sampling various conformations of the local region. Existing approaches for predicting local regions in protein structures can be broadly categorized into two classes: database and *de novo* (free modeling) methods [10,11]. Database methods search for loop fragments that best match the anchor geometries, [10,12] but these approaches usually are confined to short insertions because of poor database coverage for larger fragments. While these methods tend to be fast, the speed comes at the cost of the greater flexibility in exploring the conformational space of the loops permitted by free modeling methods. The applicability of these methods is further challenged in the modeling of InsEnds in template-based models because the regions are likely to correspond to parts of the sequence that are inaccessible to the homology methods.

In contrast, *de novo* methods sample sterically feasible loop conformations that are scored with physics-based or statistical potentials. For example, MODELLER places loop atoms uniformly between the anchor positions and optimizes the atom positions using conjugate gradient and MD with simulated annealing, scoring the loops using a combination of the CHARM22 force field and statistical preferences of the dihedral angles and atom contacts [13]. Other free modeling methods like RAPPER [14] and PLOP [15] build loop fragments by sampling from a dihedral angle

library for each residue, beginning from one or both anchors and eventually attempting to close the loop while avoiding steric clashes.

The third challenge is associated with the scoring of various conformations. Because the number of residues whose conformation vary between the different structural candidates is small, accurate energy functions can be crucial to guide the conformational search and score the final structures. Both statistical potentials [16] and physics-based force fields [14] have been used as scoring functions in loop modeling. Some methods use statistical potentials only for filtering, while the final ranking employs all atom force fields [17].Other methods focus on all atom energy functions designed specifically for loop modeling [18,19]. However, energy functions that are good at guiding the conformational search during the loop building stage might be inadequate for the final ranking of the decoys, especially in methods where the loop building is performed incrementally and separately from closure.

Until recently, efforts in the study of local protein structure have largely centered on predicting loops in defined crystal structures. However, InsEnds predictions are made in the context of template-based models where the structures for the remainder of the protein may be imperfect, being constructed from one or more crystal structures and relying on a sequence alignment. As a result of this imperfection, the structure prediction algorithm must be lenient, thereby fundamentally distinguishing this problem from traditional loop modeling.

Although both the treatment of loops and InsEnds involve local protein modeling, they can present different sets of challenges. While loops in crystal structures are defined as regions connecting different secondary structure elements, InsEnds are defined as regions devoid of information extracted from sequence alignments. Hence, InsEnd may include regions with

complete secondary structure elements. In addition, the length of loops is governed by the structural context, and, consequently, usually contain a limited number of residues. InsEnds, on the other hand, can be of arbitrary lengths. Furthermore, the boundaries of loops are generally well defined whereas the boundaries of InsEnds are determined by the gaps in the alignments. When multiple templates are combined to generate one model, the gap regions may appear with different boundaries in different templates, thereby rendering the InsEnds boundaries ambiguous.

Our method is designed to address these issues. We demonstrate the robustness of our methods by successfully predicting the structures of long loop regions in crystal structures as well as providing blind structural predictions of InsEnds in the top homology models from our submissions to CASP9. We present a fragment free method for local structure prediction.

Approach

Our approach assumes that the principles governing the folding of proteins are equally applicable for modeling InsEnds. We have shown that single backbone (φ,ψ) pivot moves provide an effective way to sample conformations, provided the moves are contingent upon the identities and conformations of the nearest neighbors (NNs). These moves have been used successfully in the fragment-free *de novo* prediction of the structures of single domain proteins [20,21].

Our local structure prediction method generates random local conformations using the same single pivot (φ,ψ) move set as for our global structure prediction scheme (Fig. 2). The interaction energy is calculated both within the local regions and between the local region and the rest of the protein. The total energy is used to guide the conformational search, an approach that differs from many methods in which the loop fragment is constructed one residue at a time while simultaneously trying to satisfy the loop closure constraint at the end. In contrast to some

existing methods that separate loop building and closure into two subsequent stages, our approach integrates the two into a single simulated annealing Monte Carlo (MCSA) scheme, thus retaining the tertiary context of the entire protein during the simulation while attempting to rapidly find the best local conformation. This tertiary context can be critical for identifying crucial loop-protein interactions, thus greatly reducing the search space. The algorithm is designed to handle multiple loops in the same MCSA trajectory. Hence, when two loops are close enough to interact, they are modeled simultaneously.

The conformational search proceeds through MCSA scheme (described in detail in the Methods section) that is guided by a combination of the pairwise additive, orientationally dependent statistical potential DOPE-PW, along with a harmonic ligation energy term to close the loop. The relative weight of the ligation energy increases during the MCSA to enforce loop closure. Explicit side chains are absent during the sampling stage of the simulation since the DOPE-PW statistical potential and backbone torsional move set implicitly incorporates sufficient information concerning the side groups [20]. Final conformations are scored using a combination of structural clustering and accessible surface area of the hydrophobic residues to select the best solutions. The standard deviation in the positions of a given loop residue in a cluster (i.e., the tightness of the cluster) provides a metric for assessing the local quality of the predictions for the loop.

RESULTS AND DISCUSSION

Three different modeling scenarios are considered. First, we address the traditional loop modeling problem in crystal structures where the structure surrounding the loop is known. We next address InsEnds modeling as applied in the CASP9 blind prediction competition, where the

InsEnds are as large as 45 aa regions in template-based models generated by Xu's RAPTOR-X algorithm [22]. The third scenario is for the CASP9 refinement category in which the InsEnds algorithm is applied to the best structure from the server predictions and where the starting model and boundaries are specified by the organizers.

Loops in crystal structures.

In order to demonstrate the applicability to larger loops in crystal structures, 26 loops of lengths 8 to 12 have been randomly selected from standard loop benchmarking studies [15]. Loop boundaries for each target are taken as previously specified, and the loops are modeled using our method. Figure 4 illustrates the process of selecting the top 5 predictions, and Table 1a presents the best and the remaining four top predictions. After the predictions are clustered according to the RMSD between the loop structures, the largest five clusters are ranked using a linear combination of the Z-scores for the cluster tightness (RMSD between structures in the cluster), size, and average DOPE-PW energy, defined as Z_t, Z_s, and Z_E, respectively,

$$Cluster\ Score = Z_s - \ Z_d - Z_t \quad , \quad Z_i = \frac{X_i - < X_i >}{\sigma_i}$$

where the Z-score for the property of structure X, is $Z_i=(X_i-<X_i>)/\sigma_i$, and $<X_i>$ and σ_i are the mean and standard deviation, respectively. After ranking the clusters, one representative from each cluster is selected using a combination of the DOPE-PW energy and the SASA to obtain the top 5 predictions. Although DOPE-PW is very successful in guiding the protein backbone into a proper conformation based on the orientation of the C_α - C_β vectors, it is unable to resolve the details of solvation at an atomistic level because it is parameterized only at the C_β level,. Hence, explicit SASA calculations are necessary to properly account for the solvation energy.

As discussed in the Methods section, the SASA scores are determined from a combined ranking of the hydrophilic and hydrophobic ASAs. Similarly, the structures are ranked using the DOPE-PW energy function as well. The structure with the lowest total DOPE-PW + SASA rank in a given cluster is taken as the predicted structure from that cluster. Models are discarded when the distance between the free end and the anchor point fails to return to within 1.5 Å of the initial distance. If the largest cluster contains less than 5% of the structures, the scoring for the top 5 candidates uses only the sum of DOPE-PW and SASA scores As shown in Supplementary Figure 1, the inclusion of SASA to the DOPE-PW energy improves the selection of the top structure in most cases compared to simply using DOPE-PW energy to select the top structure.

A residue-specific deviation quantifies the local confidence score of the prediction for each residue individually in each of the top 5 predictions. The local confidence scores are illustrated by color and thickness in Figure 4. The thicker (redder) portions in the predicted local region correspond to residues displaying the largest deviation within the cluster.

The simulations for loops of length 12 and 8-11 residues generate conformations with global loop RMSDs of 2.76Å and 1.93 Å, respectively, where the RMSD is calculated for the loop residues after aligning the structures without the loop regions (Table 2). These results can be compared to Table II of Lee et al. [8] which presents the minimum backbone RMSDs found using different existing loop sampling protocols. Simulations for 12 and 8 residue loops in crystal structures with the cyclic coordinate descent (CCD) protocol [9] generates minimum RMSDs of 3.05 Å and 1.59 Å, the CJSD [8] method obtains 2.34 Å and 1.01 Å, the self organizing algorithm using an alternating scheme of pairwise distance adjustments(SOS)[23] yields 2.29 Å and 1.19 Å and the FALC [8] scheme finds 1.84 Å and 0.78 Å, respectively .

The average RMSDs of our top ranked predicted loops are 3.98 Å and 3.13 Å for loops with 12 and 8-11 amino acids respectively (Table 2). These results are comparable to those from two different methods, RAPPER [14] and FALC [8], ranked by DFIRE [16] as listed in Table IV in Lee et al. [8], where the average RMSDs of the top ranked 12 residue loop decoys for RAPPER and FALC are 4.32 Å and 3.84 Å respectively. Rossi et al. [24] compare four different commercial loop modeling packages - Prime (Schrödinger, LLC), MODELLER (Accelrys Software, Inc.), ICM (Molsoft, LLC) and Sybyl (Tripos, Inc.) – which obtain RMSD values ranging from 3 to 5 Å for loops with 10-12 amino acids. Our performance is comparable to these methods.

We also compare our results to a recent atomic level loop modeling study which has sub-angstrom level accuracy[25]. Although our C_β level modeling certainly limits us in terms of obtaining sub-angstrom models, we still are able to obtain better or comparable models for some of the same benchmark proteins than the high resolution Kinematic Closure (KIC) protocol (Supplementary Table 1 in Ref.[26]). For instance, our top predicted model for 1hfc with 3.69 Å RMSD outperforms KIC's 8.2 Å prediction for the same loop. Similarly, our top predictions for the other targets 4i1b, 1msc, 1cyo and 1pmy from our benchmark set in Table 1a yield RMSDs of 2.03 Å, 5.5 Å, 2.47 Å and 2.97 Å that are better or comparable to the high resolution KIC method's top predictions of 3.8 Å, 3.2 Å, 5.2 Å and 2.6 Å, respectively, for the same 10-12 amino acid loops. The results demonstrate that a C_β level representation of the protein chain without a costly analytical closure constraint is sufficient to achieve accuracy comparable to existing methods for relatively long loops in the context of crystal structures.

Ends in crystal structures.

Another challenge involves modeling the termini of protein structures, a challenge that has attracted only limited study[27,28]. Unlike loops, end regions require no loop closure. To demonstrate that our method is also applicable to end regions, we have refolded the termini for six proteins (Table 1b). In each of the case, twenty residues in the C terminal end of the native proteins are first randomized while the rest of the structure is kept fixed. Starting from these pseudo-random structures, the end residues are sampled and clustered using the loop modeling protocol. Because no loop closure is required, the termini are folded using only the DOPE-PW energy function.

In 3/6 cases (1af7, 1o2f, 1r69), the best and the predicted structures have a global RMSD of under 3 Å (Table 1b), with the best local RMSD under 5 Å. Although direct comparisons are unavailable for the same proteins, the results are comparable to another method[28] for refolding of terminal secondary structures where the average RMSDs of 4.6 Å and 2.0 Å are obtained for 10-23 residue ends after three minimizations using the DFIRE and dDFIRE energy functions. We select the last 20 residues in each of the proteins for modeling irrespective of where the secondary structure boundaries lie. This protocol better mimics the situation encountered in authentic template based modeling where the number of unknown residues that require modeling is determined by the gaps in the sequence alignment and where no reliable information is often available about secondary structure type or boundaries.

CASP9 blind InsEnds predictions.

Methods designed for predicting the structure of internal loops may be inappropriate for termini of proteins because the energy functions and sampling generally used for loop modeling assume

both ends are fixed. Furthermore, InsEnds can encompass whole secondary structure elements. The existing loop modeling methods have been benchmarked for loops in crystal structures where the remaining structure and loop boundaries are known. The situation for homology modeling, however, is more complex, being highly dependent on the quality of the sequence alignments, template identification, and boundary determination. Consequently, the starting point for InsEnds modeling is imperfect and inexact.

The biannual CASP experiments present a unique platform for testing new and benchmarking developed methods through blind predictions. Our participation in CASP9 as MidwayFolding (groups TS435 and TS477) focused on testing our local structure prediction method and on improving poorly predicted local regions in template-based models. Our analysis begins with models generated by the program RAPTOR-X, which utilizes homology to identify template structures appropriate to a target sequence through sophisticated sequence/structure alignments. The templates are processed by MODELER to generate our starting model. We also use the sequence alignments of RAPTOR-X to identify the InsEnds regions in the models. Five entries may be submitted to CASP9 for each target, and Figure 5 displays the best of the 5 blind models submitted to CASP9 for each target.

The CASP9 targets serve as examples to illustrate several strengths of our method. Several of the insertion regions contain secondary structure elements in the targets. The target T0464 from CASP8 presents a case where the insertion region is a helix, which our method predicts correctly, improving the model's RMSD from 9.6 Å to 4.5 Å, as exhibited in Figure 5A. Another target, T0623 has a 25 residue gap in a region that is, in fact, a hairpin that is correctly predicted by our method as well (8.2 Å RMSD improved to 6.3 Å).

The largest InsEnds contains 45 residues (T0585), and the RAPTOR-X+MODELER programs describe them as a large loop. Our method correctly identifies that the missing region corresponds to three helices that pack into the protein core, thereby improving the model substantially from 15.1 Å to 9.1 Å overall RMSD as depicted in Figure 5H. The target TR606 presents an example where the local modeling is performed for both termini simultaneously to form a pair of beta strands, thereby improving the overall RMSD from 4.9 Å to 3.8 Å for the target as a result of modeling the ends (Fig. 5G).

Other CASP targets contain InsEnds that are loops connecting different secondary structures. For instance, the targets T0520, T0594 and T0612 yield initial models with loops containing as many as 17 residues (identified from the gap boundaries in the sequence alignments). Use of our InsEnds protocol for these three loop regions improves the overall RMSDs from 3.2 → 2.6 Å, 2.2 →1.7 Å and 7.3 → 6.6 Å for T0520, 594 and 612, respectively (Fig. 5C-D). The demonstration that we successfully model various types of InsEnds with the same protocol without any prior knowledge of whether they are loops or contain secondary structure elements highlights the robustness of the method.

Blind Prediction of refinement targets in CASP9

The judges for the refinement category in the CASP experiment select the best of all submitted (template-based) models from all participating groups. The local regions that deviate most from the native structure are identified to the predictors as the refinement targets. From our perspective, the refinement category is distinct because the starting model is guaranteed to be the best of the all CASP server models rather than one of RAPTOR-X's model and because the

boundaries for InsEnds are specified based on where the server model differs from the native structure (as identified by the organizers) rather than from RAPTOR-X's sequence alignment.

On average for the 12 refinement targets, the 24 different refinement methods in CASP8 yield no net improvement over the starting models [29]. Table 3 lists the RMSD as well as the Global Distance Test (GDT) changes from the starting models along with the ranking of our method with respect to all the other refinement methods. Our method proceeds by first initializing the InsEnds regions to a completely random conformation, so that no structural information about the InsEnds is retained from the starting model.

Unlike the RMSD which relies on a single alignment, the GDT scores reflect the structural similarities at different distance cutoffs and therefore are generally better at assessing improvements in local regions. [30] We have attempted 11 targets for refinement in CASP9 (Fig. 6) and improve the GDT scores for 7 of them. Among all groups participating in CASP9 refinement, 4 out of our 11 predictions (targets TR517, TR568, TR569 and TR517) fall in the top 10% of all submissions, and 8 out of the 11 reside in the top 25% of all submissions, thereby outperforming several of the more costly all atom refinement methods. The improvements are achieved for targets with a wide range of starting GDTs (>50). The GDT/RMSD for TR569 improves from 73.1/3.01 Å to 76.58/2.24 Å, and our method ranks 4th out of the121 total submissions for this target. The starting values for TR568 are lower at 53.35/6.39 Å, and we improve them to 56.7/5.1 Å with an overall ranking of 6th out of 127 submissions for the target. Our method performs much worse than the rest of the methods for one target, TR592, presumably because the starting structure is already extremely good (91.2/1.2 Å) that our C_β level representation is inadequate, and, consequently, an all atom side chain representation is

required to improve the model further. Moreover, we have not refined the side chains in any of the cases, something that probably would have improved the results even more.

Figure 6 displays all of our predictions for the refinement targets in CASP9. The figure illustrates how well the model aligns to the native structure before refinement (initial) and after refinement (after) when superposed using the LGA program [30]. The improvements introduced into the local region also help to align the remainder of the protein in several cases. For example, in TR614, even though the actual regions modeled are an insertion from 33-50 and the C terminal residues 106-121, the local alignment of the N terminal residues improves over the starting model as indicated with blue in the LGA alignment for TR614 in Figure 6.

Molecular Replacement Results for CASP9 refinement targets

One of the CASP9 refinement metrics assesses how well the predicted models reproduce the experimental data [31]. Recently, models generated by the structure prediction methods have been inserted into the molecular replacement likelihood algorithms for X-ray crystallographic refinement to solve the phase problem [32,33]. The assessors for CASP9 refinement judge the quality of each submitted model in this regard by calculating the Z-score of the best orientation of the model in the unit cell of the crystal compared to placing it in a set of random orientations. Only models with Z-scores above 6 are considered good enough to solve the phase problem. Table 3 in Ref. [31] summarizes how often various groups improve the Z-score of the targets from likely unrefinable (< 7) to likely refinable (> 7). Our method performs as well or better than all the other groups in this test, with positive results in 2 out of 3 cases attempted. Since our approach employs a backbone + C_β model with the side chains either missing or added simply using SCWRL4.0 with no further refinement, some of our submitted models were discarded in

the analysis by assessors. Regardless, the fact that our method ranks at the top in the molecular replacement test proves its real value in X-ray crystal structure refinement.

In contrast to most other methods that expend considerable computing resources on including all-atom interactions, our method lacks explicit side chain atoms. This difference highlights the distinction between the refinement of crystal structures and template-based models. The all atom refinement of crystal structures benefits from having high resolution information for the rest of the structure, whereas homology models are usually far from perfect. It is unclear whether the expensive modeling of all the atoms in an imperfect environment provides a computationally efficient strategy. In contrast, the first step of our approach is designed to obtain the proper backbone structure and orientation for the local region using a coarse level of modeling that is less sensitive to the atomic level details for the rest of the homology model. Once the coarse level model is obtained for the local region, side chains may be added, and more detailed all-atom refinement can proceed.

Global InsEnds RMSD vs. local InsEnds RMSD

RMSDs are calculated in three ways help to quantify the quality of the modeling of local InsEnds regions,

a. Local InsEnds RMSD: Align the loop and calculate the RMSD of only the InsEnds region.

b. Global InsEnds RMSD: Align all the residues besides the InsEnds, and then calculate the RMSD of the InsEnds region.

c. Global structure RMSD: Optimally align all the residues in the protein and calculate the RMSD of the full chain.

The local InsEnds RMSD is a measure of how well the InsEnds region itself is modeled, and the global InsEnds RMSD provides a measure of how well the modeled InsEnds is oriented with respect to the rest of the protein. The global InsEnds RMSD is the ideal measure of loop quality when predicting loops in crystal structures because the only difference between the native structure and the model can appear in the loop region. In contrast, InsEnds modeling of homology models begins from inexact structures, and, therefore, assessing the refinements requires accounting for the RMSD of the rest of the structure (besides the InsEnds) with respect to the native structure. If the starting homology model deviates significantly from the native structure, the alignment of the non-InsEnds region necessarily must skew the anchor regions, and therefore the global InsEnds RMSD would not provide as a good a metric for reporting the accuracy of InsEnds modeling than either the local InsEnds RMSD or the overall RMSD of the structure.

This utility of the different RMSDs is illustrated for six targets from CASP8 for which the initial RAPTOR models have variable RMSDs to the native structures. The 11-12 residue InsEnds regions in those models are chosen for (post-dictum) prediction using our method (Table 4). Not surprisingly, the global InsEnds RMSD is highly dependent on the quality of the initial model (i.e., the RMSD of all but the InsEnds region in the initial model). For target T0478D1, the RMSD of the non-InsEnds region in the starting model is 8.07 Å; the best local InsEnds RMSD decreases from 2.9 to 1.58 Å: whereas the best global InsEnds RMSD decreases from 12.2 to 8.4 Å.

Target T0411D1 has the non-InsEnds RMSD of the starting model much closer to native structure at 2.74 Å, and our local InsEnds RMSD improves from 3.53 to 1.85 Å, similar to the local InsEnds RMSD improvement in T0478D1(2.9 Å to 1.58 Å). However, the global InsEnds

RMSD for this target improves from 10.2 Å to 2.78 Å, which is much more remarkable than the global InsEnds RMSD in T0478D1 (12.2 Å to 8.4 Å). The difference can be attributed T0411D1's starting model having the non-InsEnds region much closer to the native structure as compared to T0478D1. Figure 7 illustrates this behavior and indicates that the local InsEnds RMSD remains relatively unaffected, whereas the global InsEnds RMSD for the same targets are quite severely affected by the RMSD of the remaining region. The successes of the modeling also support our previous contention from protein structure predictions that the neighbor dependent φ,ψ distributions capture local interactions reasonably well [20].

Applications to protein folding simulations

Although loop modeling is often called the "mini-folding problem", traditional approaches to loop modeling do not consider the folding mechanism when predicting loops. Our method on the other hand, views local modeling in a fashion that fits naturally into the larger problem of protein folding.

Experimental studies indicate that proteins fold through sequential stabilization of tertiary structure elements or foldons [34-37]. Often, long range contacts form early in the folding pathway and produce intermediate species where some entrained local regions are not yet folded. Hence, a computational scheme designed to predict structure by mimicking the natural stepwise fashion of folding pathways should encounter the problem of folding inside of loops.

Our InsEnds algorithm is well suited to address this problem because the undetermined local regions in the structure that arise during the folding pathway can correspond either to distinct secondary structures, loops, or to combinations thereof. As a proof of principle, we test our

method by predicting native structures of possible intermediates in the pathways for folding two proteins, ubiquitin and barnase.

The late folding intermediate in ubiquitin lacks the 3_{10} helix and the β_5 strand, while the rest of the structure is well formed [34,38] (Fig. 8B). Starting from a native-like structure for the intermediate, the InsEnds algorithm is used to fold the 18 residues insertion. The InsEnds refinement procedure successfully recovers the native structure to a global RMSD of 1.6Å (Fig. 8C). This illustrates an example where the local region is neither a loop nor a continuous secondary structure. Nevertheless, we still obtain the right topology, essentially completing the last step of the folding pathway to predict to the native structure.

Barnase is a 108 residue protein that is atypical for a small protein because it contains 3 distinct hydrophobic cores. The two hairpin loops depicted in Figure 8D are crucial to the structure because they are involved in formation of the protein's cores, and, therefore, the correct prediction of the loops is essential for the prediction of the global structure. Experiments indicate that loop 2 is the last structure to form in the folding pathway [36]. Applying the InsEnds method to fold both the 10 and 15 residue loops in barnase (Fig. 8E,F), our best predictions in both cases lie in the top clusters, and the best global RMSDs are 2.03 and 1.27 Å for loops 1 and 2, respectively.

The problem of folding inside of loops highlights two aspects of our method. The first is that our approach treats local and global structure prediction similarly by mimicking the natural protein folding mechanism. The second aspect is the demonstration that given the correct boundaries, our method is able to reconstruct the local structures irrespective of whether the local regions are well defined secondary structures or loops.

Simultaneous Folding of multiple InsEnds

One crucial feature of our approach is the ability of simultaneously modeling multiple local regions. When the regions are interacting, simultaneous modeling can be essential because the context provided by one local region may be important in guiding the other into place. A good example is the CASP target TR606, where the InsEnds correspond to the two termini that form a hydrogen-bonded pair of β strands. The initial template model fails to identify the ends as strands, and, therefore, the ends are wrongly placed. Accurate modeling requires that they be folded simultaneously. Guided only by the orientationally dependent DOPE-PW energy function, we have modeled the free termini and correctly predicted the pair of strands in our top submission (Fig. 5G).

Protein Structure prediction pipeline.

Here our goal is to combine the respective strengths of free modeling with templated-based modeling for an integrated structure prediction pipeline. This goal is realized through an automated server, created for CASP9 that integrates the InsEnds, RAPTOR-X and ItFix methods. Given a sequence, the pipeline begins by performing homology modeling using RAPTOR-X. If no templates are identified, the pipeline directs the sequence for free modeling using our existing ItFix algorithm for secondary and tertiary structure prediction. If RAPTOR is able to build a template-based model, the InsEnds are modeled to obtain a final structure. The pipeline has been used for the CASP9 structure predictions of the MidwayFolding group (CASP9 group numbers 435, 477).

Conclusions

Loop modeling has been an on-going challenge in protein structure prediction. With the recent surge in templated-based modeling, InsEnds modeling is a relatively new topic in need of novel approaches. Previous methods have focused on loops in the context of crystal structures and may not be generalizable to treat imprecise template-based models. InsEnds pose a more complicated situation where the poorly predicted local regions must be modeled without assumptions concerning the accuracy of the rest of the structure or the boundaries and secondary structure of the local regions being modeled. This work presents a novel free modeling method for local protein structure prediction that is applicable for modeling large local regions in both exact and inexact environments, as demonstrated by results both for loops in crystal structures and for InsEnds in template-based models. We consider this result as a step towards the generalization of the local protein structure problem. The work also presents a framework in which free and template-based modeling are integrated towards closing the final gaps in protein structure prediction.

METHODS

All backbone heavy atoms are explicitly treated, whereas the side chains are represented by single C_β atoms [20,21]. The backbone bond lengths and angles are fixed at their ideal values, and only backbone torsional angles φ,ψ are sampled during the simulation. Loop closure is achieved by ligating the free ends of the loops to the beginning of the subsequent chain with a harmonic constraint whose strength increases as 1/Temperature during the MCSA procedure (Fig 3).

Ramachandran Map (Pivot) Move Set and Sampling. The study uses our approach for sampling single residue (φ,ψ) backbone torsional angles [21]. A distribution of φ, ψ angles is generated from

a high resolution library of PDB structures for each amino acid (aa), conditional on the identity of the flanking amino acids. These nearest neighbor (NN) dependent torsional angle distributions are pre-calculated for all 20 aas, resulting in 8000 total Rama Maps that are divided into 5^0x5^0 bins. During each Monte Carlo step, a selected residue's φ, ψ angles are changed. Besides the identity of the NN, the Rama Maps can also be restricted according to secondary structure of the aa and its NNs. The data presented in the paper, however, are obtained without the imposition of this restriction, thereby enabling the exploration of all regions of torsional space allowed for a given amino acid based on its neighbor's identity. The only exception to this is the CASP8 target T0464, where 5 of the 24 residues are restricted to helical angles as the PSIPRED program [39] predicts them to be helical with high confidence.

Energy functions. The conformational search is guided through the simulation by an energy function that is a combination of the pairwise, orientation-dependent statistical potential DOPE-PW [20] and a harmonic ligation term for the closure of the loop:

$$E = E_{DOPE-PW} + \frac{T_k}{T}[(D - D_0)^2 + (L - L_0)^2]$$

where T is the simulation temperature, D/D_0 are the current/initial distances between the two anchor points, and L/L_0 are the distances between the free end and the anchor point at the site of the cut. The ligation term becomes stronger as the simulated annealing temperature decreases. The initial temperature of the simulations is set to 100, and T_k is chosen such that the contributions from the DOPE-PW and ligation energies become comparable by the end of the simulation.

The interactions in DOPE-PW are parameterized based on the observed distance distributions in the PDB, contingent on neighbors, amino acid identities, secondary structures, and side chain orientations. DOPE-PW has been demonstrated to perform well in guiding the conformational search during prediction of the structure of small proteins. The DOPE-PW term initially dominates the total energy and provides greater freedom for the conformational search, thereby aiding in properly orienting the loop with respect to the rest of the structure.

Scoring. Once the set of final conformations is generated from the MCSA simulations, the best candidate among this set of conformations is chosen using a combination of quantities computed from clustering, DOPE-PW energies, and solvent accessibility.

Clustering. Clustering based on the C_α RMSD provides a very effective means to identify dominant conformations. Hierarchical clustering proceeds with a distance cutoff of 5Å, using the minimum distance method with the Cluster module in Biopython [40]. Trials with distance cutoffs of 4 Å and 6 Å do not significantly alter the results. Clustering is used only when the largest cluster contains at least 5% of the total structures. The clusters are ranked as detailed in the Results section, while the best individual structures are selected according to the sum of the DOPE-PW energy and the solvent accessible surface area (SASA).

Loop regions reside mostly on the protein surface, and thus solvent interactions can be crucial determinants of loop structures. Hence, most successful loop scoring schemes include some approximate measure for the extent of solvation as part of the scoring function[14,15,17]. While the DOPE-PW energy function accurately describes the preferred orientations of the side chains of both hydrophilic and hydrophobic residues as being directed away and toward solvent, respectively, the interactions are still assumed to be pairwise additive between C_α-C_β bond vectors and thus do not explicitly treat the solvent accessibility. Since explicit side chains are

absent during the sampling stage, the program SCWRL 4.0 [41] is used to add side chains to enable calculating the SASA using a rapid approximation with a water radius of 1.4 Å [42]. The SASAs of each residue are assigned into hydrophobic and hydrophilic components, and the structure that minimizes the hydrophobic ASA and maximizes the hydrophilic ASA is presumed to have the best ASA score. For this purpose, the structures are ranked using both the hydrophobic and hydrophilic ASAs, and the combined rank is taken as the net ASA score.

MCSA Simulation Procedure. The initial torsional angles of the InsEnds are randomly chosen so that no prior information is retained regarding its conformation, while the rest of the protein structure is kept fixed. 700-1000 independent MCSA trajectories are run using the energy functions described above. Each step of the MCSA trajectory involves selection of a random amino acid in the InsEnds whose torsional angle is modified according to the pre-generated NN dependent Rama Map for that amino acid. This results in a new InsEnds conformation whose energy is evaluated, and the conformation is either accepted or rejected based on the Metropolis criteria at that temperature using the energy functions described above. The temperature is updated every 500 Monte Carlo steps, using a polynomial time cooling schedule.[26] The simulation protocol has been implemented in a C library, called the Protein Library, and the input/output is handled using the PDB tools from the Biopython package.

Parallel Scripting. The InsEnds algorithm has been implemented for high throughput structure prediction using the parallel scripting language, Swift.[43] Swift enables the algorithm to be expressed in a high-level logical manner independent of any specific computing resources. Swift automatically parallelizes the independent invocations of the lower level protein structure manipulation programs, which are written in Python and C. Swift further provides the flexibility

of running on multiple, different, parallel architectures by automating job scheduling and error handling, and logs the provenance of all data objects produced.

ACKNOWLEGMENTS

We thank members of the Freed and Sosnick groups for helpful discussions. This work was supported by NIH grants GM081642 (TRS, KFF, JX), GM55694 (TRS), and The University of Chicago-Argonne National Laboratory Seed Grant Program (TRS, MW), the U.S. Department of Energy under Contract DE-AC02-06CH11357 (MW), and NSF grant OCI-1007115 (AA). Swift was developed, supported and applied under NSF grants OCI-721939, OCI-0944332, and OCI-1007115.

Computational results were produced using: the PADS resource (NSF grant OCI-0821678) at the Computation Institute, a joint institute of Argonne National Laboratory and the University of Chicago; NSF TeraGrid resources provided by UTexas/TACC under grant number TG-MCB090169; and resources of the Open Science Grid Engagement program.

REFERENCES

1. Moult J. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. Curr Opin Struct Biol 2005;15(3):285-9.
2. Jones S, Thornton JM. Prediction of protein-protein interaction sites using patch analysis. Journal of Molecular Biology 1997;272(1):133-143.
3. Shi L, Javitch JA. The second extracellular loop of the dopamine D2 receptor lines the binding-site crevice. Proc Natl Acad Sci U S A 2004;101(2):440-5.
4. Wu SJ, Dean DH. Functional significance of loops in the receptor binding domain of Bacillus thuringiensis CryIIIA delta-endotoxin. J Mol Biol 1996;255(4):628-40.
5. Bruccoleri RE, Karplus M. Chain Closure with Bond Angle Variations. Macromolecules 1985;18(12):2767-2773.
6. Dinner AR. Local deformations of polymers with nonplanar rigid main-chain internal coordinates. Journal of Computational Chemistry 2000;21(13):1132-1144.
7. Wedemeyer WJ, Scheraga HA. Exact analytical loop closure in proteins using polynomial equations. Journal of Computational Chemistry 1999;20(8):819-844.
8. Lee J, Lee D, Park H, Coutsias EA, Seok C. Protein loop modeling by using fragment assembly and analytical loop closure. Proteins-Structure Function and Bioinformatics 2010;78(16):3428-3436.
9. Canutescu AA, Dunbrack RL, Jr. Cyclic coordinate descent: A robotics algorithm for protein loop closure. Protein Sci 2003;12(5):963-72.
10. Michalsky E, Goede A, Preissner R. Loops In Proteins (LIP) - a comprehensive loop database for homology modelling. Protein Engineering 2003;16(12):979-985.
11. Choi Y, Deane CM. FREAD revisited: Accurate loop structure prediction using a database search algorithm. Proteins-Structure Function and Bioinformatics 2010;78(6):1431-1440.
12. Deane CM, Blundell TL. CODA: a combined algorithm for predicting the structurally variable regions of protein models. Protein Sci 2001;10(3):599-612.
13. Fiser A, Do RK, Sali A. Modeling of loops in protein structures. Protein Sci 2000;9(9):1753-73.
14. de Bakker PIW, DePristo MA, Burke DF, Blundell TL. Ab initio construction of polypeptide fragments: Accuracy of loop decoy discrimination by an all-atom statistical potential and the AMBER force field with the generalized born solvation model. Proteins-Structure Function and Genetics 2003;51(1):21-40.
15. Jacobson MP, Pincus DL, Rapp CS, Day TJF, Honig B, Shaw DE, Friesner RA. A hierarchical approach to all-atom protein loop prediction. Proteins-Structure Function and Bioinformatics 2004;55(2):351-367.
16. Zhang C, Liu S, Zhou YQ. Accurate and efficient loop selections by the DFIRE-based all-atom statistical potential. Protein Science 2004;13(2):391-399.
17. Honig B, Soto CS, Fasnacht M, Zhu J, Forrest L. Loop modeling: Sampling, filtering, and scoring. Proteins-Structure Function and Bioinformatics 2008;70(3):834-843.
18. Levy RM, Felts AK, Gallicchio E, Chekmarev D, Paris KA, Friesner RA. Prediction of protein loop conformations using the AGBNP implicit solvent model and torsion angle sampling. Journal of Chemical Theory and Computation 2008;4(5):855-868.
19. Honig B, Xiang ZX, Soto CS. Evaluating conformational free energies: The colony energy and its application to the problem of loop prediction. Proceedings of the National Academy of Sciences of the United States of America 2002;99(11):7432-7437.

20. DeBartolo J, Colubri A, Jha AK, Fitzgerald JE, Freed KF, Sosnick TR. Mimicking the folding pathway to improve homology-free protein structure prediction. Proc Natl Acad Sci U S A 2009;106(10):3734-9.
21. DeBartolo J, Hocky G, Wilde M, Xu JB, Freed KF, Sosnick TR. Protein structure prediction enhanced with evolutionary diversity: SPEED. Protein Science 2010;19(3):520-534.
22. Peng J, Xu J. Boosting Protein Threading Accuracy. Lect Notes Comput Sci 2009;5541:31.
23. Liu P, Zhu F, Rassokhin DN, Agrafiotis DK. A self-organizing algorithm for modeling protein loops. PLoS Comput Biol 2009;5(8):e1000478.
24. Rossi KA, Weigelt CA, Nayeem A, Krystek SR, Jr. Loopholes and missing links in protein modeling. Protein Sci 2007;16(9):1999-2012.
25. Mandell DJ, Coutsias EA, Kortemme T. Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. Nature Methods 2009;6(8):551-552.
26. Aarts EHL, Korst J. Simulated annealing and Boltzmann machines : a stochastic approach to combinatorial optimization and neural computing. Chichester [England] ; New York: Wiley; 1989. xii, 272 p. p.
27. Yang Y, Zhou Y. Ab initio folding of terminal segments with secondary structures reveals the fine difference between two closely related all-atom statistical energy functions. Protein Sci 2008;17(7):1212-9.
28. Yang YD, Zhou YQ. Specific interactions for ab initio folding of protein terminal regions with secondary structures. Proteins-Structure Function and Bioinformatics 2008;72(2):793-803.
29. MacCallum JL, Hua L, Schnieders MJ, Pande VS, Jacobson MP, Dill KA. Assessment of the protein-structure refinement category in CASP8. Proteins-Structure Function and Bioinformatics 2009;77:66-80.
30. Zemla A. LGA: a method for finding 3D similarities in protein structures. Nucleic Acids Research 2003;31(13):3370-3374.
31. MacCallum JL, Perez A, Schnieders MJ, Hua L, Jacobson MP, Dilly KA. Assessment of protein structure refinement in CASP9. Proteins: Structure, Function, and Bioinformatics: doi: 10.1002/prot.23131.
32. Giorgetti A, Raimondo D, Miele AE, Tramontano A. Evaluating the usefulness of protein structure models for molecular replacement. Bioinformatics 2005;21 Suppl 2:ii72-6.
33. Das R, Baker D. Prospects for de novo phasing with de novo protein models. Acta Crystallogr D Biol Crystallogr 2009;65(Pt 2):169-75.
34. Krantz BA, Dothager RS, Sosnick TR. Discerning the structure and energy of multiple transition states in protein folding using psi-analysis. J Mol Biol 2004;337(2):463-75.
35. Sosnick TR. Kinetic barriers and the role of topology in protein and RNA folding. Prot. Sci. 2008;17:1308–1318.
36. Vu ND, Feng H, Bai Y. The folding pathway of barnase: the rate-limiting transition state and a hidden intermediate under native conditions. Biochemistry 2004;43(12):3346-56.
37. Maity H, Maity M, Krishna MM, Mayne L, Englander SW. Protein folding: The stepwise assembly of foldon units. Proc. Natl. Acad. Sci. U S A 2005;102(13):4741-6.
38. Sosnick TR, Krantz BA, Dothager RS, Baxa M. Characterizing the protein folding transition state using psi analysis. Chem Rev 2006;106(5):1862-76.
39. McGuffin LJ, Bryson K, Jones DT. The PSIPRED protein structure prediction server. Bioinformatics 2000;16(4):404-5.
40. Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B and others. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 2009;25(11):1422-3.

41. Krivov GG, Shapovalov MV, Dunbrack RL. Improved prediction of protein side-chain conformations with SCWRL4. Proteins-Structure Function and Bioinformatics 2009;77(4):778-795.
42. Legrand SM, Merz KM. Rapid Approximation to Molecular-Surface Area Via the Use of Boolean Logic and Look-up Tables. Journal of Computational Chemistry 1993;14(3):349-352.
43. Wilde M, Hategan M, Wozniak JM, Clifford B, Katz DS, Foster I. Swift: A language for distributed parallel scripting. Parallel Computing 2011;37(9):633-652.

Table 1a: Prediction of loops of 8-12 residues in crystal structures

		Local Loop RMSD(Align Loop, RMSD loop)						Global Loop RMSD(Align rest, RMSD loop)					
Target	Loop Length	Best	Pred 1	Pred 2	Pred 3	Pred 4	Pred 5	Best	Pred 1	Pred 2	Pred 3	Pred4	Pred 5
1rcf	12	1.81	**2.67**	3.39	3.27	3.38	2.79	2.61	**3.79**	5.29	4.63	4.64	4.53
1thw	12	1.84	4.61	**2.4**				4.53	5.94	**5.87**			
2cpl	12	1.09	**1.09**	4.14	4.02	4.72	4.37	2.43	**2.43**	6.34	6.94	6.84	8.37
1cyo	12	0.75	1.97	3.46	**1.67**	3.94	5.11	1.23	**2.47**	4.95	4.09	5.15	7.03
1hfc*	12	1.91	**2.08**	2.42	3.38	3.06	3.39	3.69	**3.69**	4.58	5.6	4.29	7.01
1onc	12	2.19	3.24	**2.19**	2.95	2.94	2.99	2.91	**3.83**	5.26	5.36	5.66	5.22
1pmy	12	0.57	2.47	**1.79**	2.28	2.64	3.79	1.24	**2.97**	3.13	2.98	4.07	7.02
1rro	12	1.84	**2.65**	2.98	3.79	2.83	2.82	4.63	6.94	7.78	8.44	**6.52**	7.34
1scs	12	2.27	**3.48**	5.2	4.91			3.1	**4.22**	7.87	6.95		
1bkf	12	1.55	2.26	2.55	**2.9**	2.69	2.35	2.48	3.37	4.82	**6.82**	4.85	3.98
2tgi	12	2.44	**2.85**	3.35	3.35	3.48	3.22	3.64	**4.19**	5.32	5	4.66	4.33
1eco	12	0.52	**3.43**	3.85	5.09	4.16	4.12	0.85	2.65	3.01	**1.99**	3.45	3.09

1msc	12	1.96	3.64	5.08	**3.57**	3.96	5.17	2.57	**5.5**	7.3	11.9	12.6	7.9
1acf	11	1.8	**2.05**	3.59	3.52	3.88	2.22	2.34	**2.86**	4.21	4.09	4.41	3.19
1cid	11	1.16	**1.17**	2.19	1.65	3.34	3.32	1.39	**1.83**	3.16	2.17	5.47	4.83
1noa	11	1.68	4.14	4.35	4.7	**3.43**	4.35	2.59	7.22	6.7	8.3	**6.51**	7.7
1plc*	11	1.8	2.05	2	**1.8**	4.59	1.84	3.42	**3.42**	4.89	4.08	8.35	3.57
1xnb	11	0.91	**1.62**	2.67	3.3	2.72	3.17	1.39	**2.41**	3.42	4.45	3.67	4.6
4i1b	11	1.17	**1.13**	3.72	3.64	3.11		2.03	**2.03**	9.71	10.3	10.83	
8dfr*	11	1.71	**1.71**	2.04	3.2	3.85	2.96	2.89	**2.89**	4.79	6.19	6.37	6.39
1aaj	11	1.71	**2.5**	2.69	2.81	3.14	2.62	2.79	**3.59**	4.66	5.9	7.08	6.03
5p21	10	1.74	2.78	2.99	2.47	2.67	**2.38**	2.62	4.27	3.87	**3.24**	3.94	3.71
5fx2	10	1.93	2.79	3.42	**2.3**	3.77	3.92	2.29	3.23	3.55	**2.6**	5.32	4.7
1cbs	8	0.74	3.65	**0.74**				1.74	5.41	**1.74**			
1xnb	8	0.28	1.49	1.21	**0.53**	1.16	1.13	0.77	3.53	1.83	**0.85**	1.86	1.94
1poa	8	0.44	**0.54**	0.97	1.01	3.94	2.71	0.76	**1.23**	2.1	1.59	5.56	4.84

* refers to cases where the top cluster contain less than 5% of the total structures. In those cases, the top 5 predictions are selected using DOPEPW+SASA instead; units are in Å.

- The bold font indicates the best out of the top 5 predictions.

Table 1b: Prediction of 20 end residues in crystal structures

			Local Ends RMSD(Align Ends, RMSD Ends)						**Global Ends RMSD(Align rest, RMSD Ends)**					
Target	**Type**	**Ends Length**	**Best**	**Pred 1**	**Pred 2**	**Pred 3**	**Pred 4**	**Pred 5**	**Best**	**Pred 1**	**Pred 2**	**Pred 3**	**Pred4**	**Pred 5**
1af7	α	20	1.33	2.36	**2.21**	3.74	3.8	3.07	2.24	**2.24**	3.8	4.36	5.58	5.42

1o2f	αβ	20	1.76	**2.01**	3.26	3.06	3.07	4.72	2.37	**2.80**	5.94	4.04	4.29	11.9
1mky	αβ	20	3.47	3.82	4.6	**3.71**	4.3	4.15	4.11	5.68	**5.58**	7.78	9.66	8.72
1b72	α	20	4.06	5.08	4.95	**4.69**	5.19	5.11	5.2	5.8	6.71	**5.6**	6.75	6.15
1r69	α	20	2.17	**2.50**	3.28	7.46	8.04	8.08	2.72	**2.99**	5.04	8.35	9.6	8.34
1tif	αβ	20	4.85	**5.50**	8.75	9.38	9.19	9.68	6.31	**7.02**	10.6	12.1	11.5	11.6

Table 2: Statistics for loops in crystal structures.

Length	Local loop RMSD (Å)				Global loop RMSD (Å)			
	Best		Pred1		Best		Pred1	
	Average	Stdev	Average	Stdev	Average	Stdev	Average	Stdev
12	1.59	0.62	2.80	0.85	2.76	1.14	3.98	1.3
8 to 11	1.31	0.54	1.97	1.09	1.93	0.93	3.13	1.74

Table 3: Blind InsEnds prediction of refinement targets in CASP9.

CASP9 refinement target	GDT starting	RMSD starting	GDT MidwayFolding	RMSD MidwayFolding	Rank of MidwayFolding
TR569	73.1	3.01	**76.58**	2.249	4/121
TR568	53.35	6.963	**56.7**	5.108	6/127
TR517	71.38	4.646	**72.17**	4.638	11/119
TR622	67.42	7.47	**69.47**	5.773	12/120
TR606	71.95	4.85	**72.56**	3.915	19/128
TR594	87.32	1.805	86.07	1.957	23/134
TR567	78.34	3.435	**78.52**	3.46	28/107
TR557	67.6	4.074	**68.2**	3.74	28/118
TR614	75.21	4.1	67.36	4.895	43/118
TR624	54.71	5.529	52.9	5.577	47/122
TR592	91.43	1.204	82.38	3.415	111/131
TR576	48.91	10.926	Ignored since initial GDT<50		

The numbers reported are the GDT and RMSD values from the CASP9 website. The values in bold indicate targets with an improvement in the GDT score from the starting model.

Table 4: Prediction of InsEnds in CASP8 structures (post-diction).

CASP8 Target	InsEnds Length	RMSD of InsEnds region	Local InsEnds RMSD to native (align InsEnds, RMSD of InsEnds)			Global InsEnds RMSD to native (align non InsEnds, RMSD of InsEnds)			Global protein RMSD to native (align all, RMSD of all)		
			Initial model (RAPTOR + Modeler)	InsEnds Best	InsEnds Predicted	Initial model (RAPTOR + Modeler)	InsEnds Best	InsEnds Predicted	Initial model (RAPTOR + Modeler)	InsEnds Best	InsEnds Predicted
T0431D1	11	5.21	0.22	0.52	0.91	5.34	4.14	4.7	4.76	4.04	4.81
T0456D2	12	2.09	0.49	0.47	0.47	1.44	1.22	2.03	2.04	2.12	2.16
T0478D1	12	8.07	2.9	1.58	3.05	12.2	8.4	11.8	8.26	8.21	8.31
T0443D1	12	3.42	4.04	3.7	3.48	11.13	7.3	7.67	5.33	4.42	4.62
T0411D1	11	2.74	3.53	1.85	3.53	10.2	2.78	4.98	3.87	2.85	3.11
T0479D1	11	1.54	0.75	0.48	0.97	2.3	1.59	1.69	1.62	1.59	1.59

Figure Legends

Figure 1. The InsEnds modeling problem. A multiple sequence alignment of a target sequence to template sequences can contain insertion regions at the same location.

Figure 2: Local structure prediction algorithm.

Figure 3. The ligation terms close the loop. Constraints are placed on the distance between the two ends of the loop and the distance between the free end of the loop and the anchor residue.

Figure 4. Selection of the top 5 loop predictions for 1xnb. After clustering, the largest 5 clusters are ranked based on Z-scores with respect to cluster tightness, size, and average DOPE-PW energy. Once ranked, a selection is made from each of the 5 clusters using DOPE-PW + SASA.

Figure 5. CASP9 Ins&Ends blind predictions. Numbers indicate improvement from MODELER (Red) to our model (blue), as compared to the native structure (green) after modeling the regions enclosed by the boxes. RMSD changes are for the whole structure.

Figure 6. InsEnds predictions for CASP9 refinement targets. Difference between CA/CA distance across the sequence of the initial (starting) /native and final (refined using InsEnds method)/native after superposition using sequence-dependent LGA protocol. Official data from CASP9 official website (http://predictioncenter.org/casp9/) **.** For each target, the arrows indicate the regions where the InsEnds modeling has been performed. The blue to green color change designates regions where the InsEnds modeling improves upon the given target based on LGA superposition to native structures.

Figure 7. Global versus local RMSD. The RMSD of non InsEnds region is plotted against the global InsEnds RMSD (red) and local InsEnds RMSD (blue) for six CASP8 targets. The global InsEnds RMSD is affected severely by the quality of the homology model.

Figure 8. InsEnds algorithm applied to protein folding pathways. A) The β5 and 3_{10} helix in ubiquitin that are the last structures to form in the pathway. Their structures are depicted as disordered in the model B) of the folding intermediate and C) predicted using the InsEnds algorithm. D) Barnase native structure highlighting the two hairpin loops that are part of two different cores, and E) and F) predictions of the loops using InsEnds algorithm, respectively.

Supplementary Figure 1. Improvement in model selection after including SASA : The combination of DOPE-PW+SASA improves the selection of the top prediction in most cases compared to using only DOPE-PW.