BIOINFORMATICS

Vol. 00 no. 00
Pages 1–7

Conformational Sampling and Structure Prediction of Multiple Interacting Loops in Soluble and β-Barrel Membrane Proteins Using Multi-Loop Distance-Guided Chain-Growth Monte Carlo Method

Ke Tang [1], Samuel W K Wong [2], Jun S Liu [3], Jinfeng Zhang [4,*], and Jie Liang [1*]

[1]Richard and Loan Hill Department of Bioengineering, University of Illinois at Chicago, Chicago, IL, United States
[2]Department of Statistics, University of Florida, Gainesville, FL, United States
[3]Department of Statistics, Harvard University, Science Center, Cambridge, MA, United States
[4]Department of Statistics, Florida State University, Tallahassee, FL, United States

ABSTRACT

Motivation: Loops in proteins are often involved in biochemical functions. Their irregularity and flexibility make experimental structure determination and computational modeling challenging. Most current loop modeling methods focus on modeling single loops. In protein structure prediction, multiple loops often need to be modeled simultaneously. As interactions among loops in spatial proximity can be rather complex, sampling the conformations of multiple interacting loops is a challenging task.

Results: In this study, we report a new method called multi-loop Distance-guided Sequential chain-Growth Monte Carlo (M-DISGRO) for prediction of the conformations of multiple interacting loops in proteins. Our method achieves an average RMSD of 1.93 Å for lowest energy conformations of 36 pairs of interacting protein loops with the total length ranging from 12 to 24 residues. We further constructed a data set containing proteins with 2, 3, and 4 interacting loops. For the most challenging target proteins with 4 loops, the average RMSD of the lowest energy conformations is 2.35 Å. Our method is also tested for predicting multiple loops in β-barrel membrane proteins. For outer-membrane protein G (OmpG), the lowest energy conformation has a RMSD of 2.62 Å for the three extracellular interacting loops with a total length of 34 residues (12, 12, and 10 residues in each loop).

Contact: jinfeng@stat.fsu.edu jliang@uic.edu

1 INTRODUCTION

Protein loops are key structural regions involved in recognition and binding of small molecules and proteins. In structure prediction, accurately predicted loop regions can provide valuable information for understanding the function and dynamics of the proteins. Prediction of loop structures, or loop modeling, has received considerable attention in the past (van Vlijmen and Karplus, 1997; Fiser *et al*., 2000; Michalsky *et al*., 2003; Jacobson *et al*., 2004; Zhu *et al*., 2006; Zhang *et al*., 2007a; Zhao *et al*., 2011; Subramani and Floudas, 2012). Among these, database search methods can make accurate loop prediction if templates of loop fragments to a query protein are found at a high level of confidence (Michalsky *et al*., 2003; Choi and Deane, 2010). With the rapid expansion of the PDB, fragment based approaches are very effective for modeling loop structures (Fernandez-Fuentes *et al*., 2006a; Fernandez-Fuentes and Fiser, 2006b; Fernandez-Fuentes *et al*., 2010). Nevertheless, fragment based loop modeling approaches are blind to the local environment of loops, and currently are not applicable for modeling multiple loops that are interacting with each other. In such cases, it is necessary to use template-free methods to predict loop structures. Recent advances have enabled template-free prediction of loops with high accuracy (Fiser *et al*., 2000; Canutescu and Dunbrack, 2003; Coutsias *et al*., 2004; Jacobson *et al*., 2004; Zhu *et al*., 2006; Zhao *et al*., 2011; Tang *et al*., 2014).

Most previous loop modeling studies have focused on predicting structure of a single loop. In reality, multiple loops often need to be modeled for a particular protein target, where loops in spatial proximity can interact in complex ways (Zaccardi *et al*., 2014; Housset *et al*., 1991). An example of interacting loops can be found in the bovine pancreatic trypsin inhibitor (BPTI, pdb `6pti`), where loop A (residues 7–16) interacts with loop B (35–46). A single substitution in loop B (Y35G, pdb `8pti`) results in substantial changes to the conformations of both loop A and loop B (Housset *et al*., 1991). In this case, the conformations of the two loops are stabilized by their interactions.

Interacting loops are also found in membrane proteins. For example, in the β-barrel assembly machinery A protein (BamA), the extracellular loops interact with one another, and form a dome over the top of the β-barrel domain (Noinaj *et al*., 2013). Given that the transmembrane strands of many β-barrel membrane proteins can be predicted (Naveed *et al*., 2012), accurate loop structure prediction will facilitate the prediction of full structures of this important class of membrane proteins.

Conventional methods for modeling interacting loops is to generate single loops sequentially one loop after another. However, these methods are prone to be trapped at local energy minima

*To whom correspondence should be addressed

formed by intra-loop interactions, since it cannot effectively model the complex interactions among the loops in spatial proximity. To resolve this issue, several methods have been developed that can model two interacting loops (Rosenbach and Rosenfeld, 1995; Danielson and Lill, 2010). In (Rosenbach and Rosenfeld, 1995), a method for simultaneous prediction of the structure of multiple loops based on the bond-scaling-relaxation loop-closure algorithm was reported. For two loops of lengths of 5-7 residues in spatial proximity, their results showed that more accurate predictions can be made than modeling individual loop sequentially. Danielson and Lill developed the CorLps method (Danielson and Lill, 2010), in which an energetically feasible ensemble of individual loops are generated using the loopyMod method (Soto *et al*., 2008) disregarding the presence of other loops. Although successful in predicting a number of interacting loop pairs, this method does not work well when both loops have $\geq$ 9 residues. Methods have also been developed to model loops in certain protein classes (e.g., antibodies (Sellers *et al*., 2010)), using information specific to these classes of proteins. To the best of our knowledge, none of these existing methods can effectively predict three or more interacting loops.

In this study, we describe the multi-loop Distance-guided chain-Growth Monte Carlo method (M-DISGRO) for simultaneously modeling of two or more interacting loops. Based on chain-growth Monte Carlo sampling, which has been applied to a number of studies of proteins (Liang *et al*., 2002; Zhang and Liu, 2006; Zhang *et al*., 2007b; Lin *et al*., 2011; Tang *et al*., 2014), M-DISGRO simultaneously constructs multiple loops, and does not require completeness of one loop before growing another loop. This sampling strategy randomizes the order of the residues being sampled to generate more diverse loop conformations, making it less likely to over-sample conformations in certain local energy minima. Consequently, inter-loop interactions are taken into account more effectively compared to sampling one loop at a time. In addition, loop growth is guided by empirical end-to-end distance functions and backbone dihedral angle distributions, allowing effective exploration of low-energy conformational space (Liu and Chen, 1998; Zhang *et al*., 2007a; Liu, 2008; Wong, 2013; Tang *et al*., 2014). To improve the sampling of long loops, we further introduce a strategy of regrowing loops using fragments.

Our paper is organized as follows. We first describe the M-DISGRO method in detail. We then present results for loop prediction of soluble proteins using two different test data sets, followed by results on predicting multiple interacting loops in β-barrel membrane proteins. We show that M-DISGRO has significant advantages in modeling native-like multi-loop compared to CorLps reported in (Danielson and Lill, 2010). The performance is further improved when loop fragments are used.

2 METHODS

2.1 Multi-loops Distance-guided chain-Growth Monte Carlo (M-DISGRO)

Based on our previous DISGRO method for sampling single loops, our task in this study is to model n $\geq$ 2 loops in spatial proximity in protein structures (Tang *et al*., 2014). M-DISGRO simultaneously samples conformations of the n loops. At each step of the chain growth process, a residue i in a chosen loop p is added upon completion of the previous residues in all of the n loops. Here loop p is chosen randomly from the n loops, regardless of loop lengths. The newly added residue is represented by three consecutive backbone atoms during the growth process: C atom of residue i, N atom of residue $i+1$, and CA atom of residue $i+1$ (Figure 1). The coordinates of the three atoms, C_i, N_{i+1} and CA_{i+1}, are denoted as $\mathbf{x}_{C,i}$, $\mathbf{x}_{N,i+1}$, and $\mathbf{x}_{CA,i+1}$, respectively. Here $\mathbf{x}_{C,i}$ and $\mathbf{x}_{N,i+1}$ are determined by sampling the dihedral angles (ϕ, ψ). We describe this sampling process in detail in Section **Sampling backbone atoms**. The ω dihedral angles that determine the coordinate of CA atoms are sampled from a normal distribution with a mean of 180° and standard deviation of 4°. Side-chains are built upon completion of backbone placement of all loops. The generated conformations of the loops are scored and ranked by our atom-based distance-dependent empirical potential function. Our potential function is an empirical function following (Miyazawa and Jernigan, 1996; Zhou and Zhou, 2002; Li *et al*., 2003; Zhang *et al*., 2005; Pokarowski *et al*., 2005). Explicit water molecules are not considered, and there is no specific solvation energy calculation beyond what is implicit in the empirical potential function. Details of side-chain construction and the atomic potential function are described in (Tang *et al*., 2014).

Unlike the CorLps method, which is based on ensembles of individual loops with complete structures (Danielson and Lill, 2010), M-DISGRO does not require completeness of one loop before growing another loop. It generates multiple loops by randomly selecting one residue to grow from all incomplete loops. All previously built residues in different loops immediately become part of the structural environment and contribute to the calculation of the coordinates of future atoms. Compared to methods based on sampling one whole loop at a time, our method can generate more diverse conformations of multiple loops, and can effectively avoid being trapped at some local energy minima.

2.2 Loop fragment libraries

During the chain growth process, a partial conformation may have to adopt a state with very high (or infinite) energy, making further growth fruitless. For example, it may be impossible to add any atoms due to lack of available space and loops may fail to close. M-DISGRO would have to terminate the whole chain-growth process upon such a failure. This is costly if failure occurs when a significant or the full portion of other loops have been built.

To increase the efficiency of sampling, we develop a fragment-based regrowth method to repair failed loops in M-DISGRO, as the near-native regions can be more efficiently explored by using fragments to generate an adequate number of native-like loop conformations. The overall procedure of M-DISGRO with fragments is outlined in Figure 2.

For a protein with n loops to be sampled, we first build n loop fragment libraries, one for each loop. Conformations in each library are sampled independently in the absence of the other $n-1$ loops. All loop fragments are generated using the DISGRO single loop prediction method (Tang *et al*., 2014). Conformations of the individual loops generated are then ranked by the atom-based distance-dependent empirical potential function described in (Tang *et al*., 2014). For each individual loop, the top-50 energetically favorable loop conformations are retained to form the loop fragment library specific to this loop.

During the chain-growth process, when a failure occurs in loop p which begins at residue s and ends at residue t with a length of $(t-s+1)$, a randomly selected fragment f of length $(t-s-3)$ from its loop fragment library, is used to replace the residues from s to $(t-4)$ of loop p. As we apply the CSJD analytical closure method (Coutsias *et al*., 2004) at the residue $(t-2)$ to close the loop, the replacement of residues s to $(t-2)$ will lead to a high occurrence of highly similar loop conformations. In order to introduce the necessary diversity in loop conformations while at the same time achieving high sampling efficiency, we only replace the residues s to $(t-4)$ with residues from fragment f. The multi-loop chain-growth process is then continued until the rest of the residues of all loops, including residues $(t-3)$ to t of loop p, are fully built. This fragment strategy is only used for loops of length ≥ 5. For short loops of length ≤ 4, the success rate of completing loops without fragments is already sufficient.

2.3 Sampling backbone atoms

The procedure to sample backbone atoms is the same as that of the single loop method DISGRO (Tang *et al*., 2014). Below we give a brief description for completeness. Let C_t be the C-terminal anchor atom in the end residue t of a loop. We describe the sampling procedure for (C_i, O_i, N_{i+1}, and CA_{i+1}) atoms as an example (Figure 1).

C_i is generated first, followed by N_{i+1}. Denote the distance $|\mathbf{x}_{C,t} - \mathbf{x}_{CA,i}|$ between $\mathbf{x}_{CA,i}$ and $\mathbf{x}_{C,t}$ as d_{CA_i,C_t}, and the distance $|\mathbf{x}_{C,i} - \mathbf{x}_{C,t}|$ between $\mathbf{x}_{C,i}$ and $\mathbf{x}_{C,t}$ as d_{C_i,C_t}. Since the bond length l_{CA_i,C_i}, and the bond angle $\theta_{C,i}$ are fixed, C_i will be located on a circle $\mathbf{Q_C}$ (Figure 1):

$$\begin{aligned}\mathbf{Q_C} = \{\mathbf{x} \in \mathbb{R}^3 | \text{ such that } ||\mathbf{x} - \mathbf{x}_{CA,i}|| = l_{CA_i,C_i}, \\ \text{and } (\mathbf{x} - \mathbf{x}_{CA,i}) \cdot (\mathbf{x}_{CA,i} - \mathbf{x}_{N,i}) = \cos\theta_{C,i}\}.\end{aligned} \quad (1)$$

Given a fixed d_{C_i,C_t}, C_i can be placed on two positions $\mathbf{x}_{C,i}$ and $\mathbf{x}_{C',i}$ on circle $\mathbf{Q_C}$, $\mathbf{x}_{C,i}$ and $\mathbf{x}_{C',i}$ are labeled as C_i and C_i', respectively. As the probability for placing C_i on either position is equal, we randomly select one position to place atom C_i.

Sampling from the empirical distributions of d_{C_i,C_t} and mapping back to C_i should encourage the growth of loops to connect to the terminal C_t atom. Further analysis of the empirical distribution of d_{C_i,C_t} given d_{CA_i,C_t} shows that d_{CA_i,C_t} can be very informative for sampling d_{C_i,C_t} in some cases. This leads us to design a strategy of sampling $\mathbf{x}_{C_i}$ based on the conditional distribution of $\pi(d_{C_i,C_t}|d_{CA_i,C_t})$. Atom N_{i+1} is generated in a similar way as C_i. Details can be found in (Tang *et al*., 2014).

The trial positions of (C_i, N_{i+1}, CA_{i+1}) are then subject to a filtering procedure using an empirically derived backbone dihedral angle distribution. One filtered trial is selected according to its probability calculated using an atomic distance-dependent empirical potential function. The coordinate of O_i atom is determined by (N_i, CA_i, C_i). See Tang *et al*., 2014 for more details.

3 RESULTS

3.1 Test Sets

To assess the accuracy of M-DISGRO and facilitate direct comparison with the CorLps method reported in (Danielson and Lill, 2010), we use their test set, which we name as the CorLps Set. It is obtained from high-resolution (1.53 Å) structure of trypsin (pdb `1utk`) (Leiros *et al*., 2004). It contains 36 pairs of interacting loops, including 7 pairs of 6-6-residue loops, 10 pairs of 6-9-residue loops, 3 pairs of 9-9-residue loops, 9 pairs of 6-12-residue loops, 5 pairs of 9-12-residue loops, and 2 pairs of 12-12-residue loops.

We have also developed a new data set of 2-, 3- and 4-interacting loops, called MultiSet, to assess the effectiveness of M-DISGRO in more complex situations. Proteins in this set are taken from Ref. (Fiser *et al*., 2000) and Ref. (Soto *et al*., 2008). In total, MultiSet contains twenty 2-interacting loops, eleven 3-interacting loops, and eight 4-interacting loops (See Table 2). We use a strict criterion to select interacting loops: all loops must be spatially close and interact with each other. For example, each of the 4 loops have to be interacting with the other 3 loops in a 4-loop target. We use the edge simplices from the alpha-shape computed from the protein structures to detect interactions among neighboring loops (Edelsbrunner and Mücke, 1994; Liang *et al*., 1998; Li *et al*., 2003). Using alpha-shape eliminates spurious neighbor interactions that are not in physical contact (Zhang *et al*., 2005; Ouyang and Liang, 2008). Here we use the solvent radius of 0.5 Å following (Singh and Thornton, 1993). In MultiSet, loops have lengths ranging from 4 to 16. To construct a challenging data set of two interacting loops, both loops have to be at least 10-residue long. For 3-loop and 4-loop target proteins, only those with at least one long loop (length ≥ 10) are included. Multiple interacting loops are present across different protein families. Based on the classification scheme of loops of the Archdb (Bonet *et al*., 2013), 42.9% of long loops (≥ 10) and 48.6% of loops in the MultiSet have already been classified, suggesting modeling multiple loops will be applicable to a large number of proteins.

The test set of loops in β-barrel membrane proteins contains four proteins satisfying the following criteria: (1) no in-plug structure are present; (2) no incorrect loop structures are included; and (3) loops are of length $2-17$ residues. Our test set contains both extracellular and periplasmic loops, which are modeled separately. This test set is called β-Barrel Set (See Table 3).

3.2 Multi-loop structure prediction on CorLps Set

We first test M-DISGRO using the CorLps Set. The sampled loop conformations are ranked based on their energy values calculated using our atom-based distance-dependent empirical potential function (Tang *et al*., 2014) instead of the DFIRE empirical potential function as in Ref. (Danielson and Lill, 2010). For each loop pair, the RMSD of the lowest energy conformation to the native structure R_{Emin} among $10,000$ trial conformations is reported. Our results are all reported as global backbone RMSD, calculated using the N, CA, C and O atoms of the backbone. The conformation of a loop is obtained by superimposing its two ends on the flanking secondary structures, while the deviation is calculated using the backbone atoms of the loop structure.

From the results summarized in Table 1, we find that M-DISGRO performs significantly better than CorLps. Compared to CorLps, M-DISGRO has a R_{Emin} of 1.14 Å *vs* 2.35 Å for 6-6-residue loop pairs, 2.00 Å *vs* 3.39 Å for 6-9-residue loop pairs, 2.14 Å *vs* 4.38 Å for 9-9-residue loop pairs, 2.67 Å *vs* 4.43 Å for 6-12-residue loop pairs, 2.94 Å *vs* 5.47 Å for 9-12-residue loop pairs, and 3.32 Å *vs* 6.97 Å for 12-12-residue loop pairs, respectively. Overall, the value of R_{Emin} is reduced significantly using M-DISGRO compared to CorLps.

When fragments are used, M-DISGRO further improves R_{Emin} to 1.57 Å from 2.00 Å for 6-9-residue loop pairs, 1.64 Å from 2.14 Å for 9-9-residue loop pairs, 2.53 Å from 2.67 Å for 6-12-residue loop pairs, 2.41 Å from 2.94 Å for 9-12-residue loop pairs, and 2.87 Å from 3.32 Å for 12-12-residue loop pairs, respectively. Only for 6-6-residue loop pairs, M-DISGRO with fragment has a slightly larger R_{Emin} of 1.19 Å. These results indicate that using loop fragment library is effective in improving the modeling accuracy when loops are 9 residues or longer.

We also report results using a single independent loop modeling method S-DISGRO for comparison. A maximum number of 10,000 loop combinations are generated by initially combining the top-100 ranked single loop conformations of each loop region, similar to CorLps, except single loop conformations are generated here by the DISGRO method (Tang *et al*., 2014) instead of the loopyMod (Soto *et al*., 2008) used in CorLps. S-DISGRO performs much worse in modeling multiple interacting loops compared to M-DISGRO, with or without fragments (3.15 Å vs 1.93 Å / 2.22 Å), but shows improved accuracy compared to CorLps method (3.15 Å vs 4.02 Å).

The average R_{Emin} of 36 interacting loop pairs are 1.93 Å (M-DISGRO with fragment), 2.22 Å (M-DISGRO without fragment), 3.15 Å (S-DISGRO), and 4.02 Å (CorLps). Overall, M-DISGRO

Table 1. Comparison of R_{Emin} of loop conformations generated by CorLps, M-DISGRO, and M-DISGRO with fragment using the CorLps Set.

Lengths of Loop Pairs	R_{Emin} (Å)			
	CorLps	s-DiSGro	m-DiSGro	m-DiSGro + frag
6-6 (7)	2.35	1.61	1.14	1.19
6-9 (10)	3.39	2.37	2.00	1.57
9-9 (3)	4.38	3.47	2.14	1.64
6-12 (9)	4.43	4.01	2.67	2.53
9-12 (5)	5.47	4.55	2.94	2.41
12-12 (2)	6.97	4.65	3.32	2.87
all (36)	4.02	3.15	2.22	1.93

R_{Emin}: average RMSD of the lowest energy conformations compared to native loop conformation resolved experimentally. Lengths of Loop Pairs (X-Y): the pairs of interacting loops with lengths X and Y. The number of loop pairs is listed in parentheses.

with fragment shows significantly improved accuracy in modeling multiple interacting loops.

M-DISGRO is also much faster than CorLps. The reported computational time of CorLps is about 900 cpu minutes for two 12-residue interacting loops on a single core of a Intel Xeon 2.66 GHz quad-core machine (Danielson and Lill, 2010). The computation cost for M-DISGRO without fragment is only 14 cpu minutes on a single 2 GHz AMD Opteron processor. M-DISGRO with fragment takes a slightly longer time of 17 minutes, as loop fragment libraries specific to the loops need to be constructed on the fly.

3.3 Multi-loop structure prediction on MultiSet

We then test M-DISGRO on the MultiSet. Results are summarized in Table 2.

For the eight proteins with 4-interacting loops, the average R_{Emin} by M-DISGRO with and without fragment are 2.35 Å and 2.75 Å, respectively. M-DISGRO with fragment is close to the accuracy level of sub-angstrom (1.04 Å) for 3 target proteins containing only one loop with more than 10 residues. The remaining five target proteins are challenging, as they have at least two long loops with length $\geq$ 10. For these five target proteins, M-DISGRO with fragment has an R_{Emin} of 3.13 Å. Of these five, three of them have three long loops. For example, the structure of mandelate racemase (pdb `2mnr`) has four loops of length 13, 13, 14, and 9, respectively. The R_{Emin} by M-DISGRO with fragment is 2.96 Å. Another example is azurin (pdb `1nwp`), which has four loops of length 7, 10, 13 and 15, respectively. It is challenging to model a 15-residue loop in an inexact environment when three other interacting loops also need to be modeled as well. The total length of these loops is 45 residues. As can be seen from Figure 3a, M-DISGRO with fragment achieves good accuracy for `1nwp`, with an R_{Emin} of 2.77 Å.

For the eleven 3-loop target proteins, R_{Emin} by M-DISGRO with and without fragment are 2.13 Å and 2.32 Å, respectively. For the seven 3-loop target proteins with only one loop longer than 10 residues, M-DISGRO with fragment has an average R_{Emin} of 1.86 Å. The R_{Emin} of the rest of the 3-loop target proteins with at least two long loops are 2.52 Å by M-DISGRO with fragment.

For the twenty 2-loop target proteins, R_{Emin} values are 2.61 Å and 3.26 Å with and without using fragments, respectively. This 2-loop data set is challenging, as both loops have lengths $\geq$ 10-residue. Among the 3-loop and 4-loop target proteins, short loops can be modeled fairly easily with high accuracy, which contribute to the overall improved RMSD. In thirteen of the twenty 2-loops target proteins, M-DISGRO with fragment achieves good accuracy of $R_{Emin} \leq$ 3 Å, with six of them of RMSD $\leq$ 2 Å. Even when there is a loop of length $\geq$ 15, M-DISGRO with fragment also gives good R_{Emin}. An example is lactoferrin (pdb `1lcf`), which has a 16-residue and a 13-residue loops. The RMSD of the lowest energy conformation is 2.63 Å.

Overall, M-DISGRO with fragment achieves good accuracy in modeling multi-loops in the MultiSet. It also achieves an improved accuracy compared to M-DISGRO without fragment, with an average R_{Emin} of 2.35 Å *vs* 2.75 Å for eight 4-loop, 2.13 Å *vs* 2.32 Å for eleven 3-loop, and 2.61 Å *vs* 3.26 Å for twenty 2-loop target proteins.

3.4 Multi-loop structure prediction of β-Barrel membrane proteins

We also assess M-DISGRO in predicting loop conformations of β-barrel membrane proteins. Results are summarized in Table 3. For extracellular loops, the average RMSD of the lowest energy loop conformations R_{Emin} of the four proteins by M-DISGRO with fragment is 1.88 Å. Among these, 2 of the 4 extracellular loops in the Neisserial surface protein A, NspA (pdb `1p4t`) are longer than 10 residues (Vandeputte-Rutten *et al*., 2003). The predicted extracellular loop structures has an R_{Emin} of 1.54 Å. The outer membrane protein G, OmpG (pdb `2x9k`) has three long extracellular loops of length 10, 12, and 12, respectively. The R_{Emin} of this 3-loop target protein is 2.62 Å. Noticeably, there is a short helix in the middle of the 12-residue (138-149) loop. M-DISGRO with fragment successfully predicts this short helical secondary structural element without any explicit additional secondary structure information (Figure 3b).

The periplasmic loops are usually short turns connecting β-strands. Most of them have lengths $<$ 10-residue. Prediction of these periplasmic loops are not as challenging as prediction of extracellular loops. The average RMSD R_{Emin} of the lowest energy periplasmic loop conformations of the four proteins by M-DISGRO with fragment is 1.35 Å. For the four interacting loops at the periplasmic end of OmpG, R_{Emin} is only 0.93 Å.

4 CONCLUSION AND DISCUSSION

In this study, we present a multi-loop chain-growth Monte Carlo method (M-DISGRO) for modeling interacting loops. Our method samples multiple loops by growing one residue at a randomly selected loop in each step. The calculation of the positions of newly added atoms is determined by the coordinates of previous placed atoms in all loops. This method is capable of exploring large conformational space of multi-loop effectively, without frequently being trapped in narrow or dead space. With further incorporation of loop fragment libraries, even failed loops can be efficiently regrown. M-DISGRO has significant advantages in predicting multiple interacting loops over previous methods, such as the CorLps method (Danielson and Lill, 2010), as interacting information among loops are taken into account during the simultaneous chain-growth process.

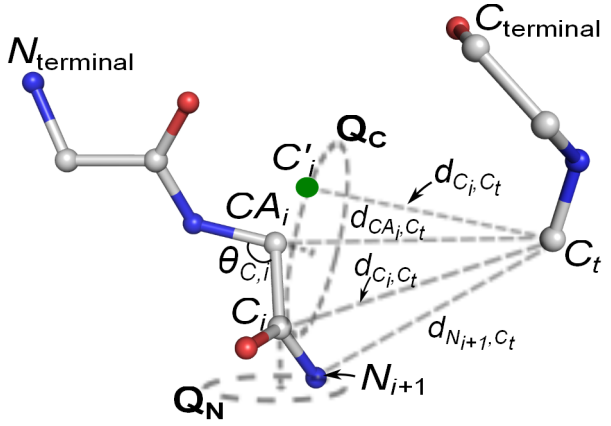

Fig. 1. Schematic illustration of placing C_i and N_{i+1} atoms. Atom C_i has to be on the circle $\mathbf{Q_C}$. The position $\mathbf{x}_{C,i}$ of the C_i atom of residue i is determined by d_{C_i,C_t}, which is based on known distance d_{CA_i,C_t} and the conditional distribution of $\pi(d_{C_i,C_t}|d_{CA_i,C_t})$. Once d_{C_i,C_t} is sampled, C_i can be placed on two positions with equal probabilities. Here $\mathbf{x}_{C,i}$ is the selected position of C_i. C'_i is placed at the position $\mathbf{x}_{C',i}$ alternative to $\mathbf{x}_{C,i}$. Similarly, the N_{i+1} atom has to be on the circle $\mathbf{Q_N}$ and its position $\mathbf{x}_{N,i+1}$ is determined by d_{N_{i+1},C_t} in a similar fashion.

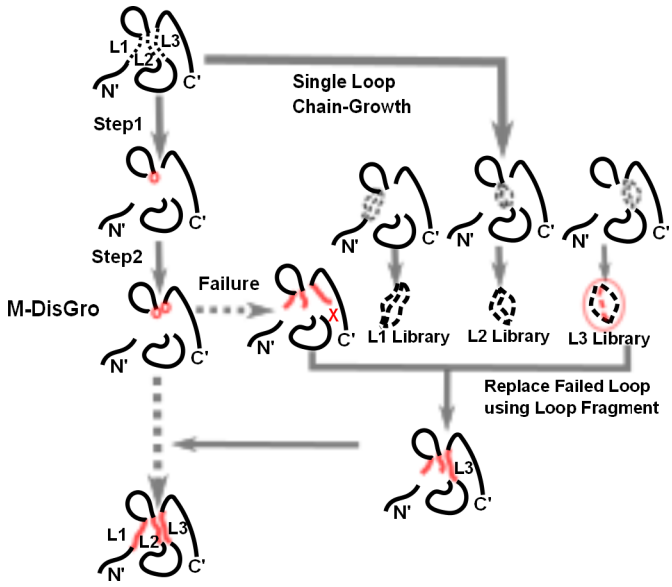

Fig. 2. The flowchart of M-DISGRO using fragments. In this example, three interacting loops (L1-L3) are to be modeled. Loop fragment libraries for them are constructed separately using the single loop method DISGRO (Tang *et al*., 2014). At each step of the chain growth process, a residue is added to a randomly chosen loop of L1-L3. This is repeated until all loops are completed. Here newly added residues are shown in red online and grey in print. In this example, a residue (red/grey circle) is added to L2 in step 1. Another residue (red/grey circle) is added to L3 in step 2. When a failure occurs (L3 in this example), a loop fragment is drawn from the corresponding library to replace the failed conformation. This process of adding residues is continued until the three missing loops are fully constructed.

Table 2. The average RMSD of the lowest energy conformations of loops in the MultiSet. R_{Emin} by M-DISGRO and M-DISGRO with fragment are listed.

PDB	Loop 1	Loop 2	Loop 3	Loop 4	R_{Emin} (Å)	
					w/o frag	w/ frag
4 Loops						
1aoz	106-115	221-225	252-259	507-510	1.35	0.93
1frd	34-49	56-67	74-86	90-94	3.61	4.29
1hnj	109-113	141-150	156-159	273-278	1.52	1.11
1ixh	85-96	106-113	120-130	186-190	2.97	2.89
1nwp	36-48	65-79	84-90	112-121	3.53	2.77
1oth	87-90	141-145	155-168	263-273	4.43	2.75
2dri	8-14	39-42	64-69	89-98	1.38	1.09
2mnr	216-228	239-251	262-275	291-299	3.23	2.96
Avg					2.75	**2.35**
3 Loops						
1bhe	194-203	216-225	248-252		2.09	1.39
1cgt	136-140	192-204	229-235		2.02	1.03
1ctt	63-73	89-102	251-256		2.99	3.23
1ddt	23-27	65-78	167-174		2.96	2.10
1ed8	198-202	233-239	245-255		1.30	2.26
1el5	79-88	230-242	257-263		2.96	3.16
1nln	44-53	70-75	109-119		2.80	2.47
1nls	11-23	98-102	201-208		2.14	1.94
1oyc	117-123	192-196	245-259		2.50	2.50
1p1m	105-108	277-288	364-369		2.25	2.18
1php	192-200	217-220	317-327		1.52	1.21
Avg					2.32	**2.13**
2 Loops						
1ads	38-50	256-265			1.39	1.14
1art	159-169	190-201			3.55	5.18
1cb0	60-73	190-199			3.58	2.35
1ctm	77-90	133-144			3.85	2.31
1lcf	100-112	232-247			3.09	2.63
1ms9	306-316	333-342			4.00	3.45
1nar	87-97	192-201			1.93	1.61
1nfp	49-64	89-99			3.60	3.45
1nhq	32-43	50-62			5.73	2.86
1nox	42-51	96-105			1.83	1.63
1ojq	88-99	141-151			2.32	2.17
1pgs	51-61	83-95			3.41	3.35
1rcf	36-48	72-81			2.06	2.24
1srp	41-55	138-150			4.66	3.64
1thg	96-106	163-174			2.26	1.72
2exo	127-139	203-213			3.75	2.96
2olb	77-89	176-190			4.95	3.04
2pia	30-42	99-112			3.36	1.89
4enl	138-147	391-402			1.85	1.43
8acn	444-458	635-644			3.94	3.23
Avg					3.26	**2.61**

Loops are designated by the beginning and the end residue positions.

M-DISGRO can model multiple interacting loops simultaneously and efficiently, especially for those target proteins with long loops. For a set of 36 interacting loop pairs, M-DISGRO with and without fragment have better average R_{Emin} of 1.93 Å and 2.22 Å, compared to CorLps of 4.02 Å. For 12-12-residue loop pairs, the R_{Emin} by M-DISGRO with fragment is 2.87 Å, and the average computing time is only 17 minutes. Furthermore, our method is

Table 3. The RMSD of the lowest energy conformations, R_{Emin} by M-DISGRO with fragment using β-Barrel Set.

Location	PDB	Loop 1	Loop 2	Loop 3	Loop 4	M-DISGRO
Extra-cellular Loops	1p4t	18-23	51-60	95-105	136-139	1.54
	1qj8	14-21	51-57	94-98	132-135	1.26
	1thq	75-81	112-122	143-152		2.08
	2x9k	97-106	138-149	177-188		2.62
	Avg					**1.88**
Peri-plasmic Loops	1p4t	37-40	71-78	119-122		0.92
	1qj8	31-37	71-77	115-121		1.71
	1thq	60-66	93-101	133-136		1.83
	2x9k	80-83	123-126	162-165	196-204	0.93
	Avg					**1.35**

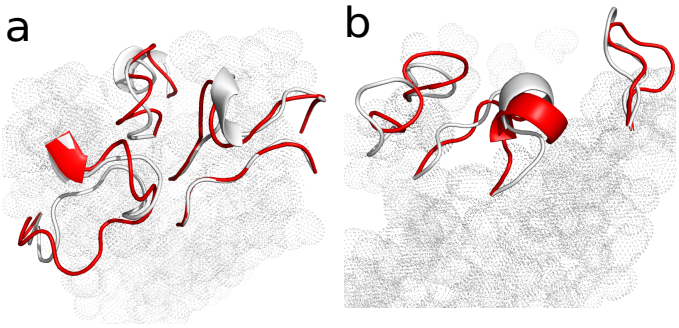

Fig. 3. Examples of prediction results of interacting loops of soluble proteins and β-barrel membrane proteins using M-DISGRO with fragment. The lowest energy modeled loops by M-DISGRO with fragment is in red/dark. The native loops are in white/light. (a) Four interacting loops of azurin (pdb 1nwp, residues 36 – 48, 65 – 79, 84 – 90 and 112 – 121). The total length of the four loops is 45. The R_{Emin} is 2.77 Å. (b) Three extracellular loops of OmpG protein (pdb 2x9k, residues 97 – 106, 138 – 149, and 177 – 188). The total length of the three loops is 34. The R_{Emin} of this 3-loop target is 2.62 Å. There exists a short α-helical secondary structure element of 4 residues which is correctly predicted.

the first that considers 3- and 4-interacting loops explicitly. It successfully predicts these loop structures, with an R_{Emin} of 2.13 Å for 3-loop target proteins, and 2.35 Å for 4-loop target proteins. Preliminary results on building loops in β-barrel membrane proteins suggests that our method can be extended to model multiple loops of other complex membrane proteins.

Our study shows that by randomizing the order of residues to be sampled, we can generate more diverse sets of loop conformations, which helps to increase the overall sampling effectiveness. However, it is possible that there exist specific orders by which loop residues are to be sampled that are especially effective for certain proteins. For example, residues in more constrained environment may need to be sampled first. One approach towards this is to look-ahead and further sample a number of states that can be grown into in the next step for each loop, and then select one state to place the next residue according to either energy or other specific bias criterion. Those sampled but unused states can be reused in future steps. This approach has been successfully applied to side chain conformations of proteins (Zhang and Liu, 2006) and reaction network sampling (Cao and Liang, 2013).

There are several directions for further improvement. The single loop method DISGRO is an important component of M-DISGRO. DISGRO is also the limiting step for the accuracy of modeled multiple interacting loops, especially when there are long loops. We envision that this can be improved by sampling dipeptide segment instead of sampling individual residue currently implemented in DISGRO. It will likely lead to improved sampling efficiency further and enable longer loops to be modeled (Zhao *et al*., 2011). Our work is also related to modeling loops in a flexible environment (Sellers *et al*., 2008; Subramani and Floudas, 2012), as other loops can be regarded as the flexible environment of the loop currently being modeled. Furthermore, the atom-based distance-dependent empirical potential function taken from Ref (Tang *et al*., 2014), can be improved by using nonlinear kernel for training (Hu *et al*., 2004), or by optimization using rapid iterations through a physical convergence function (Thomas and Dill, 1996).

ACKNOWLEDGEMENT

We thank Drs. Youfang Cao, Samuel Kou, Hsiao-Mei Lu, David Jimenez Morales, Hammad Naveed, Yun Xu, and Gamze Gursoy, Meishan Lin, and Jieling Zhao for helpful discussions.

Funding: This work is supported by NIH grants GM079804 and 1R21GM101552, NSF grant MCB-1415589, and the Chicago Biomedical Consortium with support from the Searle Funds at The Chicago Community Trust.

REFERENCES

Bonet,J. *et al*. (2013) Archdb 2014: structural classification of loops in proteins, *Nucleic acids research*, gkt1189.

Canutescu,A.A., and Dunbrack,R.L. (2003) Cyclic coordinate descent: A robotics algorithm for protein loop closure, *Protein Science*, **12**, 963-972.

Cao,Y., and Liang,J. (2013) Adaptively biased sequential importance sampling for rare events in reaction networks with comparison to exact solutions from finite buffer dCME method, *Journal of Chemical Physics*, **139**, 025101.

Choi,Y., and Deane,C.M. (2010) Fread revisited: accurate loop structure prediction using a database search algorithm, *Proteins: Structure, Function, and Bioinformatics*, **78**, 1431-1440.

Coutsias,E.A. *et al*. (2004) A kinematic view of loop closure, *Journal of computational chemistry*, **25**, 510-528.

Danielson,M.L., and Lill,M.A. (2010) New computational method for prediction of interacting protein loop regions, *Proteins: Structure, Function, and Bioinformatics*, **78**, 1748-1759.

Edelsbrunner,H., and Mücke,E.P. (1994) Three-dimensional alpha shapes. *ACM Transactions on Graphics (TOG)*, **13**, 43-72.

Fernandez-Fuentes,N., Dybas,J.M., and Fiser,A. (2010) Structural characteristics of novel protein folds, *PLoS computational biology*, **6**, e1000750.

Fernandez-Fuentes,N., Oliva,B., and Fiser,A. (2006a) A supersecondary structure library and search algorithm for modeling loops in protein structures, *Nucleic acids research*, **34**, 2085-2097.

Fernandez-Fuentes,N., and Fiser,A. (2006b) Saturating representation of loop conformational fragments in structure

databanks, *BMC structural biology*, **6**, 15.

Fiser,A., Do,R.K.G., and Šali,A. (2000) Modeling of loops in protein structures, *Protein science*, **9**, 1753-1773.

Housset,D. *et al*. (1991) Crystal structure of a y35g mutant of bovine pancreatic trypsin inhibitor, *Journal of molecular biology*, **220**, 757-770.

Hu,C., Li,X., and Liang,J. (2004) Developing optimal non-linear scoring function for protein design, *Bioinformatics*, **20**, 3080-3098.

Jacobson,M.P. *et al*. (2004) A hierarchical approach to all-atom protein loop prediction, *Proteins: Structure, Function, and Bioinformatics*, **55**, 351-367.

Leiros.H.K.S. *et al*. (2004) Trypsin specificity as elucidated by lie calculations, x-ray structures, and association constant measurements, *Protein science*, **13**, 1056-1070.

Li,X., Hu,C., and Liang,J. (2003) Simplicial edge representation of protein structures and alpha contact potential with confidence measure, *Proteins: Structure, Function, and Bioinformatics*, **53**, 792-805.

Liang,J. *et al*. (1998) Analytical shape computation of macromolecules: I. molecular area and volume through alpha shape, *Proteins Structure Function and Genetics*, **33**, 1-17.

Liang,J., Zhang,J., and Chen,R. (2002) Statistical geometry of packing defects of lattice chain polymer from enumeration and sequential monte carlo method, *The Journal of chemical physics*, **117**, 3511.

Lin,M. *et al*. (2011) Constrained proper sampling of conformations of transition state ensemble of protein folding, *Journal of Chemical Physics*, **134**, 75103.

Liu,J.S. (2008) *Monte Carlo strategies in scientific computing*. Springer Verlag.

Liu,J.S. and Chen,R. (1998) Sequential monte carlo methods for dynamic systems, *Journal of the American statistical association*, pages 1032-1044.

Miyazawa S., and Jernigan,R. (1996) Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *Journal of molecular biology*, **256**, 623-644.

Michalsky,E., Goede,A., and Preissner,R. (2003) Loops in proteins (lip)–a comprehensive loop database for homology modelling, *Protein engineering*, **16**, 979-985.

Naveed,H. *et al*. (2012) Predicting three-dimensional structures of transmembrane domains of β-barrel membrane proteins, *Journal of the American Chemical Society*, **134**, 1775-1781.

Noinaj,N. *et al*. (2013) Structural insight into the biogenesis of β-barrel membrane proteins, *Nature*, **501**, 385-390.

Ouyang,Z. and Liang,J. (2008) Predicting protein folding rates from geometric contact and amino acid sequence, *Protein Science*, **17**, 1256-1263.

Pieper,U. *et al*. (2014) Modbase, a database of annotated comparative protein structure models and associated resources, *Nucleic acids research*, **42**, D336-D346.

Pokarowski,P. *et al*. (2005) Inferring ideal amino acid interaction forms from statistical protein contact potentials, *Proteins: Struct. Funct. Bioinf.*, **59**, 49-57.

Rosenbach,D., and Rosenfeld,R. (1995) Simultaneous modeling of multiple loops in proteins, *Protein Science*, **4**, 496-505.

Sellers,B.D. *et al*. (2008) Toward better refinement of comparative models: predicting loops in inexact environments, *Proteins: Structure, Function, and Bioinformatics*, **72**, 959-971.

Sellers,B.D., Nilmeier,J.P., and Jacobson,M.P. (2010) Antibodies as a model system for comparative model refinement, *Proteins: Structure, Function, and Bioinformatics*, **78**, 2490-2505.

Singh,J. and Thornton,J.M. (1993) Atlas of protein side-chain interactions. vols. i and ii, *Acta Cryst*, **49**, 355-356.

Soto,C.S. *et al*. (2008) Loop modeling: Sampling, filtering, and scoring, *Proteins: Structure, Function, and Bioinformatics*, **70**, 834-843.

Subramani,A. and Floudas,C.A. (2012) Structure prediction of loops with fixed and flexible stems, *The Journal of Physical Chemistry B*, **116**, 6670-6682.

Tang,K., Zhang,J., and Liang,J. (2014) Fast protein loop sampling and structure prediction using distance-guided sequential chain-growth monte carlo method, *PLoS computational biology*, **10**, e1003539.

Thomas,P.D. and Dill,K.A. (1996) An iterative method for extracting energy-like quantities from protein structures, *Proceedings of the National Academy of Sciences*, **93**, 11628-11633.

van Vlijmen,H.W.T. and Karplus,M. (1997) Pdb-based protein loop prediction: parameters for selection and methods for optimization1, *Journal of molecular biology*, **267**, 975-1001.

Vandeputte-Rutten,L. *et al*. (2003) Crystal structure of neisserial surface protein a (nspa), a conserved outer membrane protein with vaccine potential, *Journal of Biological Chemistry*, **278**, 24825-24830.

Wong,S.W.K. (2013) Statistical computation for problems in dynamic systems and protein folding. PhD Thesis, Harvard University, Cambridge, MA, USA.

Zaccardi,M.J. *et al*. (2014) Loop-loop interactions govern multiple steps in indole-3-glycerol phosphate synthase catalysis, *Protein Science*.

Zhang,J., Chen,R., and Liang,J. (2005) Empirical potential function for simplified protein models: Combining contact and local sequence–structure descriptors, *Proteins: Structure, Function, and Bioinformatics*, **63**, 949-960.

Zhang,J., Kou,S.C., and Liu,J.S. (2007a) Biopolymer structure simulation and optimization via fragment regrowth monte carlo, *The Journal of chemical physics*, **126**, 225101.

Zhang,J. *et al*. (2007b) Monte carlo sampling of near-native structures of proteins with applications, *PROTEINS: Structure, Function, and Bioinformatics*, **66**, 61-68.

Zhang,J. and Liu,J.S. (2006) On side-chain conformational entropy of proteins, *PLoS computational biology*, **2**, e168.

Zhao,S. *et al*. (2011) Progress in super long loop prediction, *Proteins: Structure, Function, and Bioinformatics*.

Zhou,H., and Zhou,Y. (2002) Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction, *Protein Science*, **11**, 2714-2726.

Zhu,K. *et al*. (2006) Long loop prediction using the protein local optimization program, *Proteins: Structure, Function, and Bioinformatics*, **65**, 438-452.

Supporting Information

Influence of inter-loop contacts on modeling results

To examine the impact of the amount of interacting loop contacts on modeling results, we examine the results of 2-loop target proteins in MultiSet of similar lengths to control for loop lengths dependency. The average RMSD of lowest energy model structures to native structures R_{Emin} using m-DiSGro are similar for loops with interacting contacts ranging from 1 to 25, although there is a tendency of slight increase in R_{Emin} was observed when more inter-loop residue-residue interactions are formed (Figure S1a). We compare m-DiSGro with results using our single loop modeling approach (s-DiSGro). In s-DiSGro, a maximum number of 10,000 loop combinations are generated by initially combining the top-100 ranked single loop conformations of each loop region, similar to CorLps, except single loop conformations are generated here by the DiSGro method (Tang et al., 2014) instead of the loopyMod (Soto et al., 2008) used in CorLps. When the number of inter-loop contacts is less than 10, m-DiSGro shows only slight improvement over s-DiSGro, and the advantages of m-DiSGro over s-DiSGro are not significant when loops have small numbers of inter-loop contacts. However, the improvement of RMSD becomes more pronounced when the number of inter-loop contacts increases, e.g. R_{Emin} is improved from 4.85 Å using s-DiSGro to 2.60 Å using m-DiSGro for those with 15-20 inter-loop contacts. m-DiSGro gives better performance compared to s-DiSGro as large number of interacting loop contacts are not neglected and environmental information is used for loop generation.

The numbers of contacts made between the multiple interacting loops and the rest of the protein are denoted as $N_{interloop}$ and N_{rest}, respectively. The ratio $r_{i/r} = N_{interloop}/ N_{rest}$ is used to assess how $N_{interloop}$ and N_{rest} influence the modeling results. A large $r_{i/r}$ value indicates a relatively large number of contacts between interacting loops and a relatively small number of contacts between loops and rest of the protein. When $r_{i/r}$ increases, m-DiSGro shows a slight increase in $R_{Emin,}$ but has significant more improvement over s-DiSGro (Figure S1b).

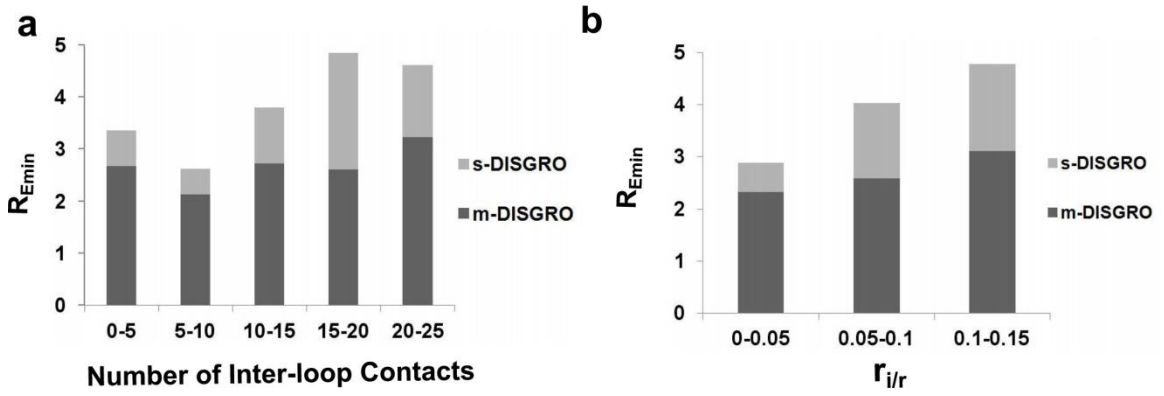

Figure S1. The influence of interacting loop contacts on modeling accuracy of 2-loop target proteins from MultiSet. Comparison in minimum energy RMSD (R_{Emin}) of loop modeling using s-DiSGro (grey bar) and m-DiSGro (black bar). (a) The histogram of the mean RMSD value of the lowest energy conformations (y-axis) of loops with different amount of inter-loop contacts (x-axis). (b) The histogram of the mean RMSD value of the lowest energy conformations (y-axis) of loops with different $r_{i/r}$ (x-axis). $r_{i/r}$ is the ratio between the number of inter-loop contacts $N_{interloop}$ and the number of non-inter-loop contacts N_{rest}. Note that R_{Emin} using m-DiSGro is consistently smaller than using s-DiSGro.

The frequency of failed conformations out of 10,000 conformations

The frequency of failed conformations when restarts with new fragments are required is high. The average numbers are 7,104, 8,125, and 9,819 out of 10,000 for total loop length of 20-30, 30-40, and 40+, respectively (Figure S2a). For loops with similar average loop length, more failed conformations are trapped to the dead end when the number of loops increases. The average numbers of failed conformations are 5,760, 6,530 and 9,196 for 2 loops, 3 loops and 4 loops targets. The average loop length of these loops is 9 to 10 residues (Figure S2c).

As the frequency of failure to grow into full length loops and need to restart the growth process with fragments is high (Figure S2), the internal database of fragment conformations is required in most cases, when loops are longer than 6 residues. m-DiSGro with fragment outperforms m-DiSGro without fragment in 24 out of 36 cases using the CorLps Set (Table 1). Additionally, two of 36 cases have a very small RMSD differences (~ 0.02 Å). Among the remaining 10 cases, at least one loop is of short length of 6 residues. Furthermore, using the internal database of fragments improves our results

for 18 out of the 20 2-loop cases in the MultiSet. Overall, we found that the internal database of fragment conformations is important for improved performance.

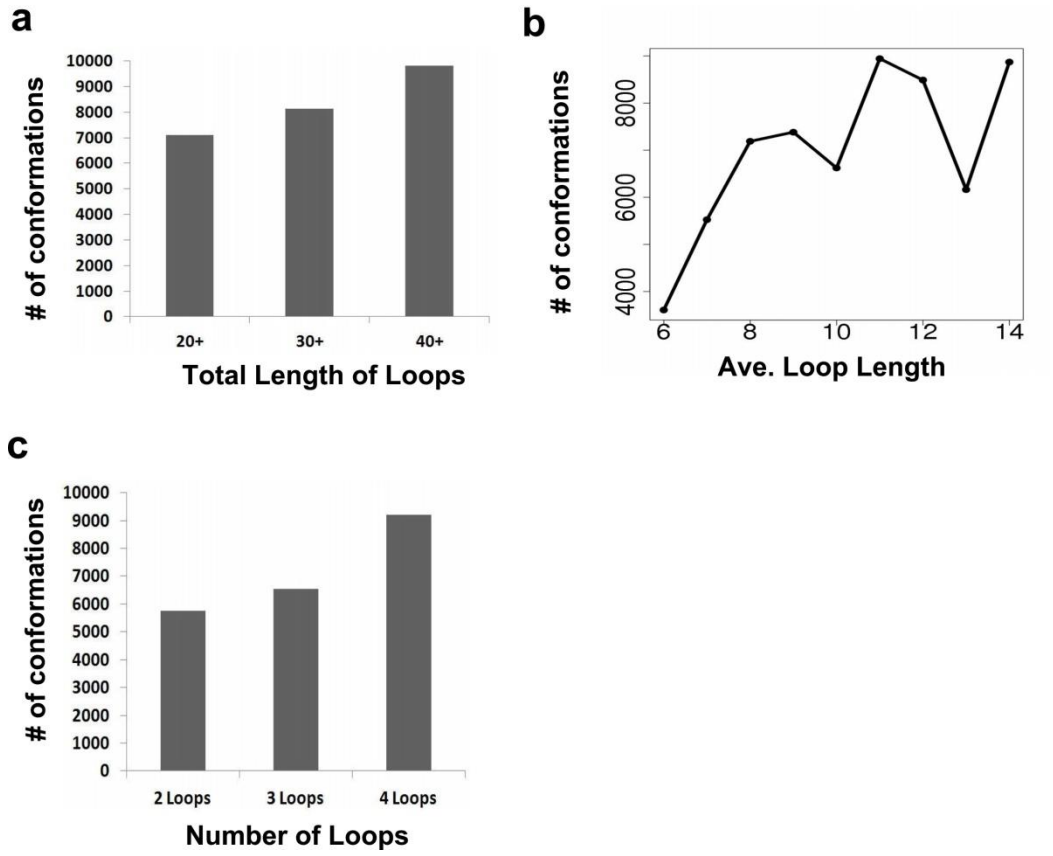

Figure S2. The dependency of frequency of failed conformations on the number and lengths of loops in the MultiSet. (a) The average number of failed conformations grouped by the total loop lengths of 20-30, 30-40, and 40+. (b) The average number of failed conformations (y-axis) is plotted against the average loop lengths (x-axis). (c) The average number of failed conformations of loops grouped by the number of loops. The average loop lengths here are 9 ~ 10 residues.

<u>Sample size dependency of modeling accuracy</u>

We test the accuracy using 1,000, 5,000 and 20,000 trial conformations on CorLps Set (Figure S3). Briefly, the accuracy of loop modeling largely depends on the number of conformations when the sample size is from 1,000 to 10,000. The RMSD of the lowest energy conformation to the native structure R_{Emin} is improved from 3.04 Å when 1,000 conformations are sampled to 1.93 Å when 10,000 conformations are sampled. However, there is no significant improvement in accuracy when the number of trial conformations is larger than 10,000. The R_{Emin} of 20,000 trials is 1.91 Å, which is very similar to 1.93 Å when 10,000 trials are sampled.

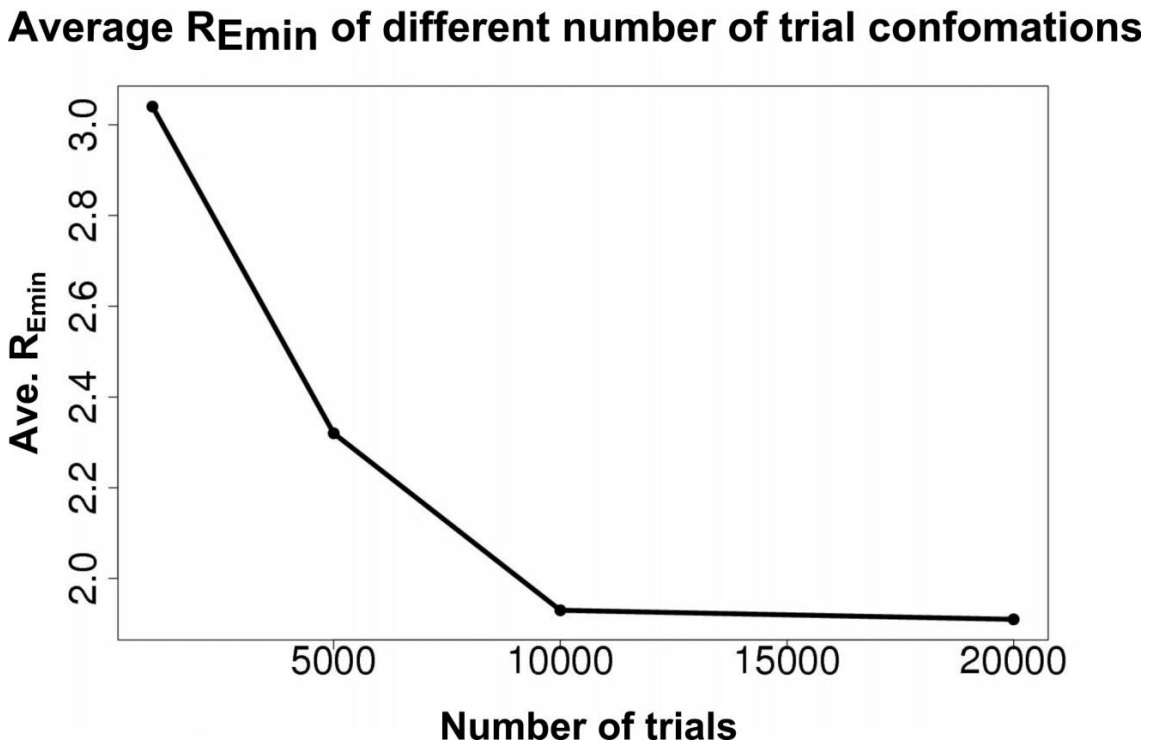

Figure S3. Mean of the backbone RMSD values of the lowest energy conformation of protein loops on CorLps Set with different number of trial conformations. The mean value of the RMSD of the lowest energy conformation R_{Emin} (y-axis) is plotted against the size of trial samples (x-axis). 1,000, 5,000, 10,000 and 20,000 trial conformations are used.

Prediction results of interacting loops of bovine pancreatic trypsin inhibitor (BPTI)

We have also modeled the interacting loops in the bovine pancreatic trypsin inhibitor (BPTI). The RMSD of the lowest energy modeled structure to the native structure are 2.01 Å and 2.65 Å for the wild type BPTI (pdb 6pti) and the Y35G mutant (pdb 8pti). The changes of the conformations of both loop A (7-16) and loop B (35-46) due to the single mutation Y35G is 1.20 Å (loop A) and 2.32 Å (loop B) in RMSD of native structures from 6pti to 8pti. Our current method successfully reproduced the changes. The changes in modeled loop A and loop B between modeled wild type and modeled Y35G mutant are 1.64 Å and 3.42 Å, respectively (Figure S4).

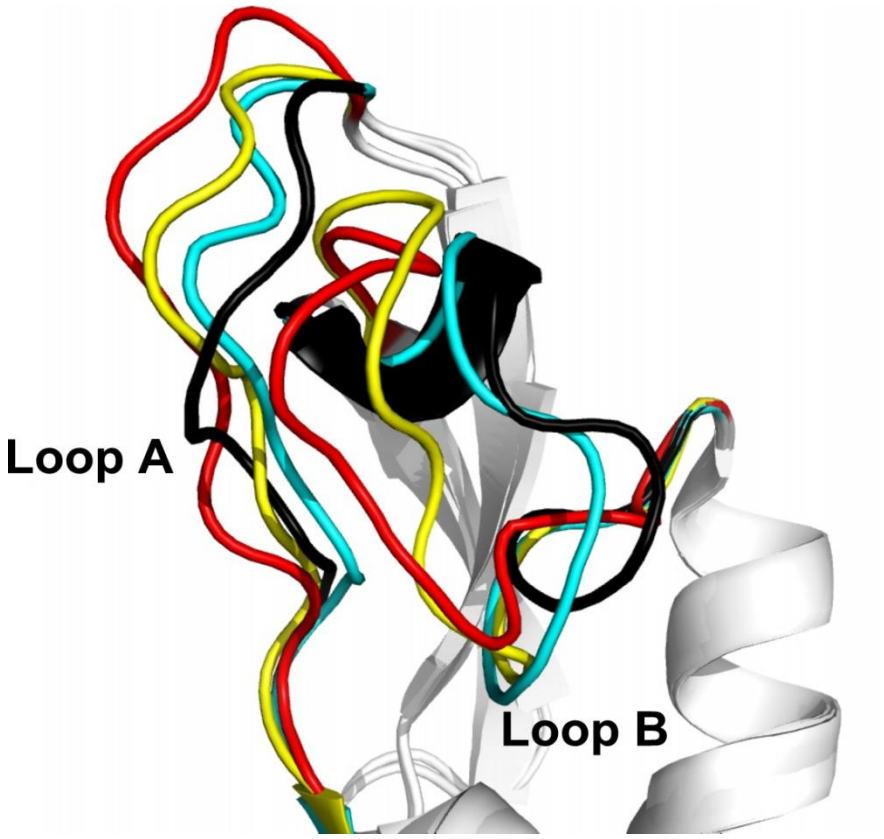

Figure S4. Prediction results of interacting loops of bovine pancreatic trypsin inhibitor (BPTI) and its mutant using M-DiSGro. Loop A (7-16) and Loop B (35-46) are two interacting loops in BPTI. The loops are in yellow, cyan, red, and black for native BPTI (pdb: 6pti), native Y35G mutant (pdb: 8pti), modeled BPTI, and modeled Y35G mutant, respectively. The total length of the two loops is 22.

Frequency of non-independent loops occur in a general database of protein models

We randomly selected 9,976 models whose sizes are 100 - 400 residues and "model score" are higher than 0.9 from MODBASE (Pieper et al, 2014). The average number of loops in these models is 14.0, each with at least one residue-residue interacting loop contacts with a distance cut-off of 6 Å. The average number of inter-loop residue-residue contacts is 96. These results suggest that non-independent loops occur with high frequencies in proteins (Figure S5).

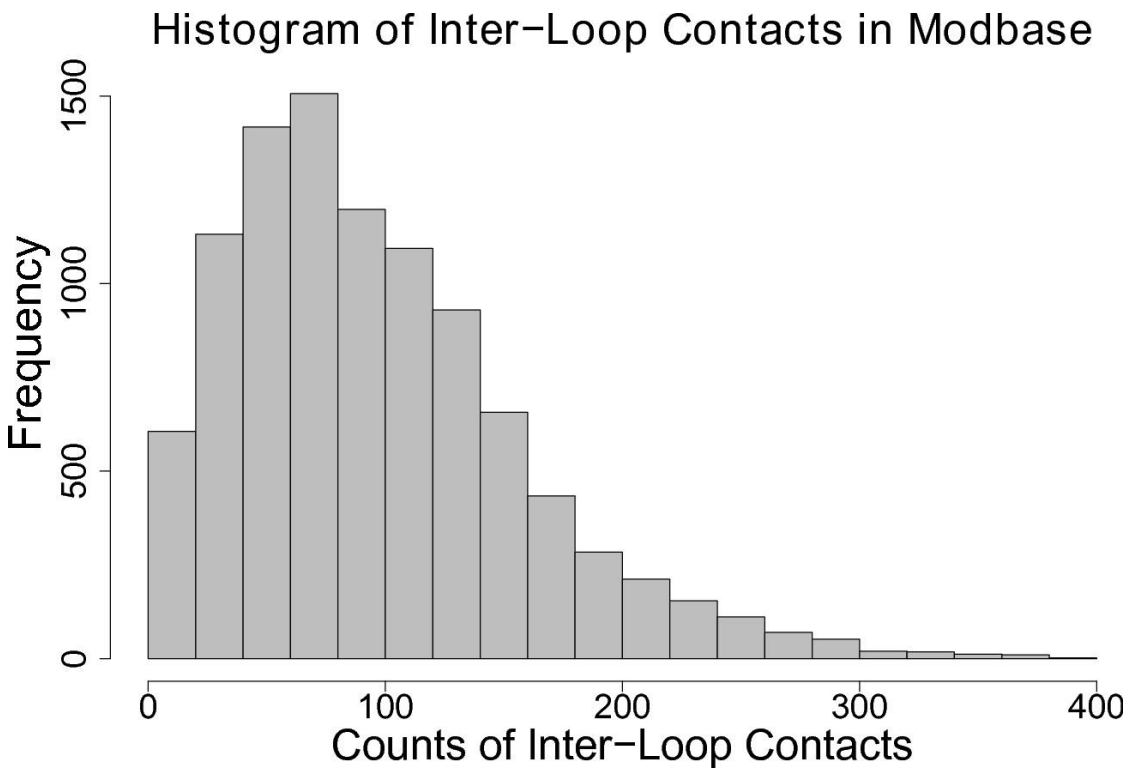

Figure S5. Histogram of the frequency of protein models from Modbase with different interacting loop contacts. 9,921 out of 9,976 models have at least one residue-residue interacting loop contacts. Most of them have >20 interacting loop contacts.

References

Bonet,J. et al. (2013) Archdb 2014: structural classification of loops in proteins, Nucleic acids research, gkt1189.

Pieper,U. et al. (2014) Modbase, a database of annotated comparative protein structure models and associated resources, Nucleic acids research, 42, D336-D346.

Sellers,B.D. et al. (2008) Toward better refinement of comparative models: predicting loops in inexact environments, Proteins: Structure, Function, and Bioinformatics, 72, 959-971.

Soto,C.S. et al. (2008) Loop modeling: Sampling, filtering, and scoring, Proteins: Structure, Function, and Bioinformatics, 70, 834-843.

Tang,K., Zhang,J., and Liang,J. (2014) Fast protein loop sampling and structure prediction using distance-guided sequential chain-growth monte carlo method, PLoS computational biology, 10, e1003539.