Published in final edited form as:

J Chem Theory Comput. 2017 January 10; 13(1): 380–388. doi:10.1021/acs.jctc.6b00845.

Distance-Guided Forward and Backward Chain-Growth Monte Carlo Method for Conformational Sampling and Structural Prediction of Antibody CDR-H3 Loops

Ke Tang[†], Jinfeng Zhang^{‡,*}, and Jie Liang^{†,*}

Department of Bioengineering, University of Illinois at Chicago, Chicago, IL, 60607, USA, and

Department of Statistics, Florida State University, Tallahassee, FL, 32306, USA

Abstract

Antibodies recognize antigens through the complementary determining regions (CDR) formed by six-loop hypervariable regions crucial for the diversity of antigen specificities. Among the six CDR loops, the H3 loop is the most challenging to predict because of its much higher variation in sequence length and identity, resulting in much larger and complex structural space, compared to the other five loops. We developed a novel method based on a chain-growth sequential Monte Carlo method, called Distance-guided Sequential chain-Growth Monte Carlo for H3 loops (DISGRO-H3). The new method samples protein chains in both forward and backward directions. It can efficiently generate low energy, near-native H3 loop structures using the conformation types predicted from the sequences of H3 loops. DISGRO-H3 performs significantly better than another *ab initio* method, RosettaAntibody, in both sampling and prediction, while taking less computational time. It performs comparably to template-based methods. As an *ab initio* method, DISGRO-H3 offers satisfactory accuracy while being able to predict any H3 loops without templates.

TOC image

1 Introduction

Protein loops are key structural components involved in recognition and binding of small molecules or other proteins.^{1–3} Their structures are difficult to determine experimentally due

*To whom correspondence should be addressed.

[†]University of Illinois at Chicago

[‡]Florida State University

to their flexibility and irregularity. Therefore, prediction of protein loop structures is an important problem, and has received considerable attention.^{4–23}

Computational prediction of loops in antibodies has the promise to determine the detailed structural basis of specific antibody-antigen interactions. Antibodies are a class of Y-shaped proteins produced by the immune system that identify and neutralize foreign pathogens. They can recognize and bind to antigens with extraordinary affinity and specificity,^{24,25} and can be used for preventing and treating various diseases, e.g. cancer,^{26,27} arthritis,²⁸ and infectious diseases.²⁹ The remarkable binding specificity and affinity are determined mainly by the six hypervariable loops, also referred to as the complementary determining regions (CDR). Three of the loops (L1, L2, and L3) belong to the variable domain of the light chain, and the other three (H1, H2, and H3) belong to the variable domain of the heavy chain. Among the six loops, the H3 loop locates at the center of the binding site and plays important roles in determining the specificity of antibody-antigen interactions.^{30,31} H3 loop is also the most diverse in terms of sequence identity and length, resulting in a larger and more complex structure space. The five non-H3 loops are usually short, with length ranges between 2 and 8 residues. In contrast, the H3 loop may have length up to 26 residues.³² The structures of the five non-H3 loops can be predicted with high accuracy using canonical classes of non-H3 CDR loops,^{32,33} which describe the relationships between the sequences and structures of those loops. However, the prediction of H3 loops has not been very satisfactory due to the increased sequence length and much larger conformational space.

A number of methods have been developed for modeling antibody H3 loops. They are either based on *ab initio* conformational search,³⁴ or rely on knowledge-based databases of structural templates of loop conformations.^{35,36} One of the widely used methods is RosettaAntibody,³⁴ which is an *ab initio* method based on fragment assembly. It achieves an average RMSD of 2.18 Å for the lowest energy conformations on a dataset of 53 loops. The 53 average minimum RMSD of the top 10 lowest energy conformations is 1.51 Å. FREAD predicts antibody H3 loops based on local sequence and geometric similarities.³⁵ H3Loopred, a recently published method uses Random Forest method to select structural templates for H3 loops from a set of candidates. After modeling the other CDR loops using canonical structure model implemented in PIGS,³⁷ it achieves an average of 2.4 Å for the same RosettaAntibody data set of 53 loops, and 2.5 Å for H3 loops using a set of 50 recently solved structures.³⁶ However, the strategy of using a knowledge-based database does not apply if there are no appropriate templates available for the loops to be modeled. As many loops may not be adequately described by a single conformation, generating an ensemble of near-native loop conformations can provide better characteristics of the dynamics of these loops, and can help to assess their entropic effects.^{38,39} As current knowledge-based methods mostly focus on generating a single conformation of low energy, they are not well suited for generating large ensembles of near-native loop conformations. Despite extensive efforts with significant progress made in recent years, conformational sampling and structural prediction of H3 loops remains a challenging problem when structural templates are not available, especially when loops are long (e.g., 1–17).

In this study, we report a novel method for conformational sampling and structural prediction of H3 loops. Our method is called Distance-guided Sequential chain-Growth

Monte Carlo for H3 loops (DISGRO-H3). It is based on chain growth with importance sampling.^{21,40–44} The strategy of sequential chain-growth sampling has been used in various of biological studies,^{42–49} for example, analysis of protein packing and void formation,⁴³ HP sequence folding problem,⁴⁴ RNA loop entropy calculation,⁴⁵ reconstruction of transition state ensemble of protein folding,⁴⁷ and reaction network sampling.⁴⁸ Prior to conformation sampling, we employ the “H3-rules” to classify CDR-H3 structure into two conformation types (kink or extended) using sequence information,^{50,51} as successfully used in RosettaAntibody.³⁴ The H3-rules can predict a loop in either kink or extended conformation type with 85% accuracy, which substantially reduces the conformational space for the subsequent sampling step.

Overall, DISGRO-H3 has significant advantages in generating native-like H3 loop conformations compared to other methods. DISGRO-H3 is able to generate decoy loop sets that are enriched with near-native structures. It also achieves high accuracy in predicting loops when combined with an atom-based distance-dependent empirical potential function. This paper is organized as follows. We first describe the DISGRO-H3 methodology. We then present results for loop prediction of antibody H3 loops using two different data sets, followed by conclusion and discussion.

2 Methods

2.1 Distance-guided chain-Growth Monte Carlo for CDR H3 Loops (DISGRO-H3)

DISGRO-H3 is developed based on our previous DISGRO method.²¹ The overall procedure of DISGRO-H3 is outlined in Figure 1.

Specifically, we first use the H3-rules to predict the conformation types (as either kink or extended) of loops of length > 6 .^{50,51} Denote N as the number of residues need to be generated backward from the end of a loop of length L . If the predicted base type is “kink”, we carry out a backward chain growth from the end residue of the loop until the N residues are generated. The coordinates of these N residues are sampled according to the torsion angle distribution of kink loops. Here N is determined by the loop length L , N is 4 when $L \leq 10$, otherwise N is set to 2. For the remaining $(L - N)$ residues, Distance-guided Sequential Monte Carlo method is applied using a general dihedral angles (ϕ, ψ) distribution, followed by the CSJD closure algorithm.⁹ For short loops of length ≤ 6 , DISGRO-H3 is exactly the same as DISGRO. Specifically, conformations are generated using the Distance-guided Sequential Monte Carlo method, followed by the CSJD algorithm for loop closure. The sampling process is described in detail in the section of **Distance-guided chain-Growth Monte Carlo method**. Side-chains are then built upon completion of the backbone atoms. The generated loop conformations are scored and ranked by our atom-based distance-dependent empirical potential function, and then evaluated in RMSD. Details of side-chain construction and the atomic potential function are described in Ref.²¹

2.2 Datasets

To create the H3 loop database, we collect X-ray structures of all H3 loops from the Structural Antibody Database (SAbDab) using a sequence identity threshold of $\geq 90\%$.⁵²

After removing all antibodies that shares > 90% H3 sequence identity to the test sets which are described in later section, this data set contains a total of 810 different PDBs. We also collected another general loop database, which consists of all loops of length 4 from 6, 521 proteins in the CulledPDB database at 30% identity, 2.0 Å resolution, and with an $R = 0.25$.⁵³ This data set has been used for obtaining torsion angle distribution of general loops in a previous study.²¹

2.3 Distance-guided chain-Growth Monte Carlo method

During the chain-growth process, a new residue is added to a partially growing chain. The newly added residue is represented by three consecutive backbone atoms: the C atom of residue i (C_i), the N atom of residue $i + 1$ (N_{i+1}), and the CA atom of residue $i + 1$ (CA_{i+1}) (Figure 2). The coordinates of C_i and N_{i+1} are determined after sampling the dihedral angles (ϕ , ψ). The sampling procedure is the same as that of our previous single loop modeling method DISGRO.^{21,23} Below we give a brief description for completeness.

Let C_e be the C -terminal anchor atom in the end residue e 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 2). Specifically, C_i is generated first, followed by N_{i+1} . Denote the distance between $\mathbf{x}_{CA,i}$ and $\mathbf{x}_{C,e}$ as $d_{CA_i,C_e} = |\mathbf{x}_{C,e} - \mathbf{x}_{CA,i}|$, and the distance between $\mathbf{x}_{C,i}$ and $\mathbf{x}_{C,e}$ as $d_{C_i,C_e} = |\mathbf{x}_{C,i} - \mathbf{x}_{C,e}|$. 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 Q_C (Figure 2):

$$Q_C = \left\{ \mathbf{x} \in \mathbb{R}^3 \mid \text{such that } \|\mathbf{x} - \mathbf{x}_{CA,i}\| = b_{CA_i,C_i} \text{ and } (\mathbf{x} - \mathbf{x}_{CA,i}) \cdot (\mathbf{x}_{CA,i} - \mathbf{x}_{N,i}) = \cos \theta_{C,i} \right\}.$$

(1)

Given a fixed d_{C_i,C_e} , C_i can be placed on two positions $\mathbf{x}_{C,i}$ and $\mathbf{x}_{C',i}$ on circle Q_C . Here $\mathbf{x}_{C,i}$ and $\mathbf{x}_{C',i}$ are labeled as C_i and C'_i (yellow ball in Figure 2), respectively. As the probability for placing C_i on either position is about equal, we randomly select one position to place atom C_i . When given d_{CA_i,C_e} this distance guidance strategy effectively bias sampling $\mathbf{x}_{C_i}$ based on the conditional distribution of $\pi(d_{C_i,C_e} | d_{CA_i,C_e})$. $\pi(d_{C_i,C_e} | d_{CA_i,C_e})$ is the conditional probability distribution of the distance between the atoms C_i and C_e (d_{C_i,C_e}), given the distance between the atoms CA_i and C_e (d_{CA_i,C_e}). Atom N_{i+1} is generated in a similar way as C_i .

In general, when there are more residues between residues e and i , the conditional distance distributions become less discriminative. However, at certain distances, the conditional distance distribution can still be quite informative, even when there are a relatively large number of residues in-between. For example, when the distance is near the maximum possible range, all residues at one end will grow directly towards the other end, which is

captured in the conditional distribution. 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 (ϕ, ψ) distribution obtained from general loop database. 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 then determined by (N_i, CA_i, C_i) . The coordinates of C_{e-1} , N_e and CA_e atoms of the end residue will be generated at the last step upon closure.

2.4 Torsion angles of base types

The structures of H3 loops are divided into two regions, the base region, which is close to the frame regions, and the β -hairpin region which is distal to the frame regions. The base regions of the H3 loops can be further classified into two classes, a *kink base region* that contains a L-bulge at the $(n-1)$ st residue (Figure 3 A), and an *extended base region* that forms a normal anti-parallel β -strands (Figure 3 B). In DISGRO-H3, we use the “H3-rules” to predict the types of base regions of loops from the amino acid sequences. The kink and extend forms are defined by the pseudo-dihedral angles θ_{base} formed by the successive CA atoms at positions $(n-2)$, $(n-1)$, (n) , and $(n+1)$. Generally speaking, the form of the base region is kink when $-100^\circ < \theta_{base} < 100^\circ$, otherwise it is extended. However, the range of θ_{base} of the two forms are still very wide. In fact, the range of θ_{base} in kink and extend forms can be further narrowed down to a small ranges as demonstrated by the distribution of θ_{base} of all H3 loops in H3 loop data set (Figure 3 C).

The distribution of θ_{base} in antibody loops (Figure 3 C), reveals a clear separation between kink and extend loops. The range of θ_{base} in kink loops is -10° to 70° , and the range is 125° to 180° or -180° to -125° in extend loops. In contrast, there is no such pattern in general loops (Figure 3 D). The differences in θ_{base} between kinked and extended loops reflect the differences in the backbone torsion angles (ψ, ϕ) of the last few residues at the C-terminal end of the loops. Among the backbone torsion angles, the ψ_{n-1} angle has the largest difference between extended and kinked forms (Figure 3 A and B). The torsion angle distribution of ψ_{n-1} in kink antibody loops (Figure 3 E), is significantly different from the distributions in extend antibody loops (Figure 3 F), as well as general loops (Figure 3 G). The dense region of ψ_{n-1} in the plot of kink antibody loop is $0^\circ - 60^\circ$, while ψ_{n-1} are enriched in the region of $(120^\circ - 180^\circ)$ in extended loops. Both regions are dense in the plot of ψ_{n-1} in general loops. As the number of extended loops is much smaller than kink loops, the distribution of torsion angles in extended loops is sparse, especially for long loops. Therefore, DISGRO-H3 does not sample the last several residues of “extend” loop using a backbone dihedral angle distribution obtained from H3 database, and treats extended loop as the same as general loops.

2.5 Backward chain-growth

During backward chain-growth, the last few residues of the loop are generated. The number of residues grown from loop end backward depends on the length of the loop. For short loops of length ≤ 6 or loops whose base types have been predicted as extend using “H3-rules”, the backward growth is turned off, hence only forward chain-growth is used. For loops whose base types are predicted as kink, the last two residues are generated backward

from the end for loops of length 7 – 9, while the last four residues are generated from the end residue for loops of length 10.

The purposes of using forward chain-growth and backward chain-growth are different. Forward growth (N to C terminal) is used to sample general loop conformations from distance distributions and residue-specific backbone dihedral angle (ϕ , ψ) distributions, without requiring detailed information of specific classes of proteins. Backward chain-growth (C to N terminal) is used to improve the effectiveness in generating loops in kinked shapes. Unlike chain-growth in a forward direction, backward chain-growth does not require detailed distance guidance. DISGRO-H3 employs only an empirically derived backbone dihedral angle (ψ , ϕ) distribution, which was obtained from H3 kink loop database, to generate coordinates of the last several residues of loops. As it generates residues in a backward direction, the dihedral angles (ψ , ϕ) are sampled in a fashion which is different from that of forward chain-growth. The backbone dihedral angle (ψ , ϕ) distribution here is loop-length specific instead of residue-specific as is used for forward chain-growth. Specifically, the distribution of the first (ψ , ϕ) pair need to be sampled is determined only by the loop length, and is not affected by the residue type. Once the base form of the target loop is predicted as kink using “H3-rule”, the backward chain-growth function is turned on.

3 Results

Test Sets—To assess the accuracy of DISGRO-H3 and facilitate direct comparison with the other methods, we use the test set from Ref.,³⁴ which we called the RA test set. It has 53 of the 54 loops of RosettaAntibody³⁴ after excluding the H3 loop of protein 2A10, as it was incorrectly reported as a six-residue loop in that study. This test set contains 3 very short (4 – 6 residues), 22 short (7 – 9), 14 medium (10 – 11), 10 long (12 – 14) and 4 very long (17 – 22) loops. We additionally build another test set of H3 loops, which we called DISGRO test set. This set contains 30 H3 loops, whose lengths range from 4 to 22-residues. All of these loops share less than 90% sequence identity to the SAbDab set used in training. These loops are divided into 5 groups in the same way as RA set.

3.1 Sampling H3 Loops

To evaluate how effective our method is in producing native-like H3 loop conformations, we use the RA set for evaluation. We generate 20, 000 loops for each target loop in the RA set. We compare our results with those from that of RosettaAntibody³⁴ and H3Loopred.³⁶ As RosettaAntibody reports the minimum RMSD of the top 10 lowest energy conformations, we list the minimum RMSD of the top 10 lowest energy conformations ($R_{E,min,10}$) generated by DISGRO-H3 in Table 1 for a direct comparison. We also list the minimum RMSD of the top 500 lowest energy conformations ($R_{E,min,500}$), and 20,000 conformations ($R_{min,20000}$).

From the results summarized in Table 1, we find that DISGRO-H3 performs significantly better than RosettaAntibody in generating low energy near-native conformations for all of the five length groups. Compared to RosettaAntibody, DISGRO-H3 has an $R_{E,min,10}$ of 0.50 Å vs 0.90 Å for very short loops (4 – 6 residues), 0.95 Å vs 1.11 Å for short loops (7 – 9 residues), 1.03 Å vs 1.21 Å for loops of medium length (10 – 11 residues), 1.76 Å vs 2.43 Å

for long loops (12 – 16 residues), and 2.39 Å vs 2.95 Å for very long loops (17 – 22 residues), respectively. These RMSD values are the averaged values over loops with the same length, which may come from different proteins. The average $R_{E,min,10}$ of 53 antibody H3 loops are 1.21 Å (DISGRO-H3) vs 1.51 Å (RosettaAntibody). DISGRO-H3 has smaller $R_{E,min,10}$ compared to RosettaAntibody in 35 out of 53 H3 loops.

Although DISGRO-H3 performs better than RosettaAntibody in a direct comparison of $R_{E,min,10}$, this comparison does not give a complete picture of the differences in sampling capability as the values of $R_{E,min,10}$ depend not only on the capability of generating near-native conformation, but also on the energy scoring functions. Compared to RosettaAntibody, DISGRO-H3 only uses a simple atomic distance-dependent empirical potential function as a scoring function, and does not require performing time consuming energy minimization. The ability of conformational sampling of DISGRO-H3 can be increased if we increase the number of retained conformations. When we take the top 500 lowest energy conformations, the average minimum RMSD, $R_{E,min,500}$ of 53 loops is 1.05 Å compared to $R_{E,min,10}$ of 1.21 Å. The minimum RMSD is further decreased to 0.81 Å when we retained all 20,000 conformations. 52 out of 53 loops have $R_{min,20000} < 2.0$ Å, including very long loops of length 17 – 22. For example, $R_{min,20000}$ of the 22-residue H3 loop in pdb 2b4c is only 1.97 Å, while $R_{E,min,10}$ and $R_{E,min,500}$ are 2.65 Å and 2.50 Å, respectively (Figure 4).

Most of the loops of length < 13 have $R_{min,20000} < 1.0$ Å and $R_{E,min,10} < 2.0$ Å, except the H3 loop in pdb 2aju. This 10-residue loop has much larger $R_{E,min,10}$ of 2.43 Å vs 1.00 Å, $R_{E,min,500}$ of 2.28 Å vs 0.87 Å, $R_{min,20000}$ of 1.85 Å vs 0.72 Å compared to the other 10-residue loops, as “H3-rules” incorrectly predicts the base type of the H3 loop as “kink” instead of “extend”. When the base type of this H3 loop is corrected to “extend”, the $R_{E,min,10}$, $R_{E,min,500}$, and $R_{min,20,000}$ of generated conformations are improved to 0.78 Å, 0.72 Å, and 0.72 Å, respectively. Noticeably, the base types of three other loops in 1mlb, 2fbj, and 1fbi, are incorrectly predicted by “H3-rules” as “extend” instead of “kink”. As DISGRO-H3 treats “extend” loop as the same as general loops, the RMSDs of these loops are not as poor as that of 2aju, and can also be improved if the base type predictions are correct.

The H3Loopred method predicts the best loop conformation to a target loop in their training dataset as the best template using trained RandomForest model. Although the RMSDs of best templates in the H3Loopred study cannot be used as a direct comparisons to $R_{E,min,10}$ values obtained by RosettaAntibody and DISGRO-H3, we examined the H3Loopred results here as a reference. The average minimal RMSDs of 53 loops obtained using H3Loopred is 1.47 Å vs 1.21 Å by DISGRO-H3. DISGRO-H3 has smaller $R_{E,min,10}$ compared to H3Loopred in 33 out of 53 H3 loops. For very long loops (17 – 22 residues), 3 of 4 loops predicted by DISGRO-H3 have smaller RMSD.

As H3Loopred is a template-based method, its capability of the prediction accuracy is still limited by the number of similar structures in a database. The RMSDs of the best templates (R_{bp}) increases significantly to 2.24 Å for 19 of the 53 loops, to which H3Loopred cannot find good templates. These include many long loops: 11 of the 19 loops have lengths 12.

Unlike H3Loopred, the minimal RMSDs using DISGRO-H3 did not change much for these 19 loops compared to the rest of the 34 loops. $R_{E,min,10}$, $R_{E,min,500}$, and $R_{min,20000}$ are 1.73 Å, 1.47 Å and 1.17 Å, respectively. The increase of RMSDs using DISGRO-H3 for these 19 loops over the rest of 34 loops is mainly due to the increase of loop length, as most of the loops in these 19 loops are long loops.

The computational time of DISGRO-H3 is similar to that of H3Loopred, and is significantly faster than RosettaAntibody. The average CPU time of DISGRO-H3 is 6 cpu minutes for generating 20,000 conformations for the target H3 loops in Table 1 on a single 2.3 GHz Intel Xeon CPU, compared to 5 cpu minutes of the average computational time of H3Loopred on a 2.5 GHz CPU of different architecture. RosettaAntibody, on the other hand, can take hours and up to days for loop prediction.³⁶

3.2 H3 loop structure prediction on RA Set

In addition to generating near-native H3 loops, DISGRO-H3 can be used to predict the conformations of loops with low energy. Here we compare our predicted loop structures with structures obtained from RosettaAntibody³⁴ and FREAD.³⁵ We report the RMSDs of the lowest energy conformations to the native structure R_{Emin} among 20,000 trial conformations using DISGRO-H3.

As summarized in Table 2, DISGRO-H3 outperforms RosettaAntibody significantly, and is slightly better than FREAD. DISGRO-H3 has an R_{Emin} RMSD value much smaller than those from RosettaAntibody and slightly smaller than those from FREAD (1.44 Å vs 2.18 Å, and 1.51 Å, respectively). Compared to RosettaAntibody and FREAD, DISGRO-H3 has R_{Emin} of 0.57 Å vs 1.27 Å and 1.23 Å for very short loops, 1.10 Å vs 1.66 Å and 1.46 Å for short loops, 1.38 Å vs 1.99 Å and 1.59 Å for medium loops, 2.00 Å vs 3.14 Å and 1.62 Å for long loops, and 2.84 Å vs 4.05 Å and 2.88 Å for very long loops, respectively. DISGRO-H3 performs better in 4 of the 5 length groups compared to FREAD, and outperforms RosettaAntibody in all 5 length groups.

For loops that belong to the very short, short, and medium length groups (loop length < 12), DISGRO-H3 shows significant advantages over the other two methods. Unlike the FREAD method, which is based on determination of sequence similarities to template loop structure, DISGRO-H3 is an *ab initio* method. The values of R_{Emin} of DISGRO-H3 have stronger length-dependency compared to FREAD when template are available. Therefore, FREAD has better R_{Emin} compared to DISGRO-H3 for a number of long loops when template are found. However, for very long loops (17 – 22), results show that DISGRO-H3 is at least comparable with FREAD (2.84 Å vs 2.88 Å), and are significantly better than RosettaAntibody (2.84 Å vs 4.05 Å). Furthermore, FREAD is challenged in generating loops of low RMSD in a number of cases. There are three loops in the RA set with $R_{Emin} > 5$ Å using FREAD, while no loop has $R_{Emin} > 5$ Å using DISGRO-H3. DISGRO-H3 generate more loop conformations with $R_{Emin} < 2$ Å than FREAD (44 vs 38). These facts suggest that DISGRO-H3 is more reliable than FREAD in predicting structures of H3-loops.

3.3 H3 loop structure prediction on DISGRO test set

To further evaluate the effectiveness of DISGRO-H3 in modeling H3 loops, we tested DISGRO-H3 using the DISGRO test set. Our results are summarized in Table 3. Overall, the average minimum RMSD of the top 10, top 500, and all 20,000 conformations are 1.37 Å, 1.12 Å, and 0.96 Å, respectively. The $R_{E,min,10}$ of 25 out of the 30 loops, including all loops of length ≤ 15 are smaller than 2 Å. 27 out of the 30 loops have $R_{min,20000}$ smaller than 2 Å. Furthermore, 20 of them have $R_{min,20000}$ smaller than 1 Å. The average RMSD of the lowest energy loop conformations of the 30 loops (R_{Emin}) is 1.58 Å. 24 among 30 loops have R_{Emin} smaller than 2 Å. DISGRO-H3 achieves sub-angstrom accuracy in sampling and predicting structures of very short, short, and medium length loops. For long loops of length 12–16, the R_{Emin} , $R_{E,min,10}$, $R_{E,min,500}$, and $R_{min,20000}$ are 1.83 Å, 1.53 Å, 1.36 Å, and 1.14 Å, respectively. These results suggest that DISGRO-H3 is also able to sample and predict H3 loops of length 12–16 with high accuracy. For very long loops of length 17–22, the R_{Emin} , $R_{E,min,10}$, $R_{E,min,500}$, and $R_{min,20000}$ increase to 3.30 Å, 2.99 Å, 2.42 Å, and 2.03 Å, respectively. Noticeably, this test set contains more than one long loop of length > 20 , which make sampling and prediction more challenging than the RA set. For example, the H3 loop in 1u8n is 22-residue long, the R_{Emin} of this loop using DISGRO-H3 is 3.47 Å, and $R_{min,20000}$ is 2.61 Å. These results show that DISGRO-H3 is also successful in sampling and predicting H3 loops of the 30 antibody proteins in DISGRO test set.

4 Conclusion and Discussion

In this study, we present a novel method for generating conformations of H3-loops and for predicting their native conformations. Our method samples antibody H3 loops through sequentially growing protein chains from specific torsion angle distributions of base types based on the “H3-rules”. The calculation of the positions of newly added atoms is determined by the coordinates of the previously placed atoms. As the geometric information of the last several residues can be obtained from the loop base types, to generate near-native kink-shaped loops, we employed the backward growth strategy to generate the last few (2–4) residues. This method is able to explore the conformational space of H3 loops more efficiently. Furthermore, the success in incorporating “H3-rules” in predicting structures of H3 loops suggests that our DISGRO method can be further improved for different classes of protein loops by incorporating the knowledge specific to these classes of loops.

We show that DISGRO-H3 is able to generate loop ensembles that are enriched with near-native structures using two test sets. It has significant advantages in generating native-like H3 loop conformations over other methods, including RosettaAntibody,³⁴ H3Loopred,³⁶ and FREAD.³⁵ The minimum RMSD of the top 10 lowest energy conformations ($R_{E,min,10}$) of 53 antibody H3 loops are 1.21 Å using DISGRO-H3 vs 1.51 Å using RosettaAntibody and 1.47 Å for the best templates using H3Loopred. The average minimum RMSD can be decreased further when more conformations are generated and retained. The minimum RMSD of the top 500 lowest energy conformations ($R_{E,min,500}$) is 1.05 Å, and the minimum RMSD of 20,000 conformations ($R_{min,20000}$) is only 0.81 Å. DISGRO-H3 also performs well in identifying native-like conformations using an empirical potential function. In comparison with RosettaAntibody³⁴ and FREAD,³⁵ DISGRO-H3 shows improved

prediction accuracy, with an average RMSD of the lowest energy conformations R_{Emin} of 1.44 Å vs 2.18 Å using RosettaAntibody and 1.51 Å using FREAD for loops in RA set.

Our study focuses on generating near-native conformations of H3 loops. A conformational ensemble which is enriched with near-native conformations is essential for predicting structures of H3 loops accurately. Because of the central locations of H3 loops in the binding site and their importance in determining the specificity of antibody-antigen interactions, improved modeling of H3 loops can lead to better understanding of the antibody functions and antibody-antigen interactions. It can also provide insight into the mechanisms of antiviral innate immune response, and may facilitate the development of monoclonal antibody therapeutics or vaccines. Generating large ensemble of near-native conformations of protein structures can facilitate studies of thermodynamic properties of proteins such as those related to configuration entropy. In addition, when coupled with an improved energy function, significant improvement in protein structure prediction can be expected.

There are several directions for further improvement. The H3-rules, although very beneficial, can be a limiting factor for further improving the prediction accuracy. Incorrect H3-rules prediction will result in sampling from wrong torsion angle distributions and generation of conformations distant from the native conformation. In addition, modeling of long loops remains a challenging problem for DISGRO-H3. We envision that DISGRO-H3 can be improved by sampling dipeptide segment instead of sampling individual residue currently used in the chain-growth process.¹⁸ Furthermore, the energy function taken from Ref²¹ can be further improved by optimization using rapid iterations through a physical convergence function,⁵⁴ or by using nonlinear kernel function for training.^{55–57} The near native conformations of antibody H3 loops generated by DISGRO-H3 can be used for further refinement when a significantly improved energy function is available. In this paper, we use a backward-growth strategy to effectively take into account the geometrical information of the last several residues of H3 loops. Although the backward-growth strategy was designed specifically for H3 loops in this work, it is generally applicable in a similar fashion as that of forward-growth, which will increase the diversity of sampled conformations. Furthermore, it can be applied to sample loops of different classes of proteins as well as to sample general structure of non-loop regions. In addition, the idea of integrating specific information of H3 loops to our sampling method can also be generalized to other classes of loops. This is similar to using secondary structure and specific fragment information in protein structure prediction, which has been very successful. For example, the extracellular loops of β -barrel membrane proteins usually orient upwards toward the extracellular side as they are constrained by excluded volume of lipid polysaccharides (LPS) and cannot insert into the membrane. Such geometrical information can be converted to a pseudo-dihedral angle distribution, which can be integrated to DISGRO to enable better sampling of loop conformations of β -barrel membrane proteins.

Acknowledgments

This work was supported by grants from the National Institutes of Health (GM079804 and GM115442) and National Science Foundation (MCB-1415589).

References

1. Bajorath J, Sheriff S. *Proteins: Struct, Funct Bioinf.* 1996; 24:152–157.
2. Streaker E, Beckett D. *J Mol Biol.* 1999; 292:619–632. [PubMed: 10497026]
3. Mani M, Chen C, Amblee V, Liu H, Mathur T, Zwicke G, Zabad S, Patel B, Thakkar J, Jeffery CJ. *Nucleic Acids Res.* 2014; 43:D277–D282. [PubMed: 25324305]
4. van Vlijmen H, Karplus M. *J Mol Biol.* 1997; 267:975–1001. [PubMed: 9135125]
5. Fiser A, Do R, Šali A. *Protein Sci.* 2000; 9:1753–1773. [PubMed: 11045621]
6. Canutescu A, Dunbrack R Jr. *Protein Sci.* 2003; 12:963–972. [PubMed: 12717019]
7. de Bakker P, DePristo M, Burke D, Blundell T. *Proteins: Struct, Funct Bioinf.* 2003; 51:21–40.
8. Michalsky E, Goede A, Preissner R. *Protein Eng.* 2003; 16:979–985. [PubMed: 14983078]
9. Coutsiias E, Seok C, Jacobson M, Dill K. *J Comput Chem.* 2004; 25:510–528. [PubMed: 14735570]
10. Jacobson M, Pincus D, Rapp C, Day T, Honig B, Shaw D, Friesner R. *Proteins: Struct, Funct Bioinf.* 2004; 55:351–367.
11. Zhu K, Pincus D, Zhao S, Friesner R. *Proteins: Struct, Funct Bioinf.* 2006; 65:438–452.
12. Cui M, Mezei M, Osman R. *Protein Eng, Des Sel.* 2008; 21:729–735. [PubMed: 18957407]
13. Sellers B, Zhu K, Zhao S, Friesner R, Jacobson M. *Proteins: Struct, Funct Bioinf.* 2008; 72:959–971.
14. Spassov V, Flook P, Yan L. *Protein Eng, Des Sel.* 2008; 21:91–100. [PubMed: 18194981]
15. Liu P, Zhu F, Rassokhin D, Agrafiotis D. *PLoS Comput Biol.* 2009; 5:e1000478. [PubMed: 19696883]
16. Mandell D, Coutsiias E, Kortemme T. *Nat Methods.* 2009; 6:551–552. [PubMed: 19644455]
17. Lee J, Lee D, Park H, Coutsiias E, Seok C. *Proteins: Struct, Funct Bioinf.* 2010; 78:3428–3436.
18. Zhao S, Zhu K, Li J, Friesner R. *Proteins: Struct, Funct Bioinf.* 2011; 79:2920–2935.
19. Nilmeier J, Hua L, Coutsiias EA, Jacobson MP. *J Chem Theory Comput.* 2011; 7:1564–1574. [PubMed: 21743800]
20. Fernandez-Fuentes N, Fiser A. *Methods Mol Biol.* 2013; 932:141. [PubMed: 22987351]
21. Tang K, Zhang J, Liang J. *PLoS Comput Biol.* 2014; 10:e1003539. [PubMed: 24763317]
22. Hayward S, Kitao A. *J Chem Theory Comput.* 2015; 11:3895–3905. [PubMed: 26574470]
23. Tang K, Wong SW, Liu JS, Zhang J, Liang J. *Bioinformatics.* 2015; 31:2646–2652. [PubMed: 25861965]
24. Mian I, Bradwell A, Olson A. *J Mol Biol.* 1991; 217:133–151. [PubMed: 1988675]
25. Sliwkowski MX, Mellman I. *Science.* 2013; 341:1192–1198. [PubMed: 24031011]
26. Garrett TP, Burgess AW, Gan HK, Luwor RB. Cartwright. *Proc Natl Acad Sci U S A.* 2009; 106:5082–5087. [PubMed: 19289842]
27. Couzin-Frankel J. *Science.* 2013; 342:1432–1433. [PubMed: 24357284]
28. Campbell J, Lowe D, Sleeman MA. *Br J Pharmacol.* 2011; 162:1470–1484. [PubMed: 21182494]
29. Waldmann TA. *Nat Med.* 2003; 9:269–277. [PubMed: 12612576]
30. Alzari P, Lascombe M, Poljak R. *Annu Rev Immunol.* 1988; 6:555–580. [PubMed: 2454644]
31. Weitznier BD, Dunbrack RL, Gray JJ. *Structure.* 2015; 23:302–311. [PubMed: 25579815]
32. North B, Lehmann A, Dunbrack RL. *J Mol Biol.* 2011; 406:228–256. [PubMed: 21035459]
33. Kabat EA, Te Wu T. *Proc Natl Acad Sci U S A.* 1972; 69:960–964. [PubMed: 4502946]
34. Sivasubramanian A, Sircar A, Chaudhury S, Gray JJ. *Proteins: Struct, Funct Bioinf.* 2009; 74:497–514.
35. Choi Y, Deane CM. *Mol BioSyst.* 2011; 7:3327–3334. [PubMed: 22011953]
36. Messih MA, Lepore R, Marcatili P, Tramontano A. *Bioinformatics.* 2014; 30:2733–2740. [PubMed: 24930144]
37. Marcatili P, Rosi A, Tramontano A. *Bioinformatics.* 2008; 24:1953–1954. [PubMed: 18641403]
38. Chirikjian GS. *Methods Enzymol.* 2011; 487:99. [PubMed: 21187223]
39. Shehu A, Kavraki LE. *Entropy.* 2012; 14:252–290.

40. Rosenbluth M, Rosenbluth A. *J Chem Phys.* 1955; 23:356.
41. Grassberger P. *Phys Rev E.* 1997; 56:3682.
42. Liu J, Chen R. *J Am Stat Assoc.* 1998:1032–1044.
43. Liang J, Zhang J, Chen R. *J Chew Phys.* 2002; 117:3511.
44. Zhang J, Kou S, Liu J. *J Chew Phys.* 2007; 126:225101.
45. Zhang J, Lin M, Chen R, Wang W, Liang J. *J Chew Phys.* 2008; 128:125107.
46. Liu, J. *Monte Carlo strategies in scientific computing.* Springer-Verlag; New York: 2008.
47. Lin M, Zhang J, Lu H, Chen R, Liang J. *J Chem Phys.* 2011; 134:75103.
48. Cao Y, Liang J. *J Chem Phys.* 2013; 139:025101. [PubMed: 23862966]
49. Cabeza de Vaca I, Lucas MF, Guallar V. *J Chem Theory Comput.* 2015; 11:5598–5605. [PubMed: 26642982]
50. Shirai H, Kidera A, Nakamura H. *FEBS Lett.* 1999; 455:188–197. [PubMed: 10428499]
51. Kuroda D, Shirai H, Kobori M, Nakamura H. *Proteins: Struct, Funct Bioinf.* 2008; 73:608–620.
52. Dunbar J, Krawczyk K, Leem J, Baker T, Fuchs A, Georges G, Shi J, Deane CM. *Nucleic Acids Res.* 2014; 42:D1140–D1146. [PubMed: 24214988]
53. Wang G, Dunbrack R. *Bioinformatics.* 2003; 19:1589–1591. [PubMed: 12912846]
54. Huang S, Zou X. *Proteins: Struct, Funct Bioinf.* 2011; 79:2648–2661.
55. Hu C, Li X, Liang J. *Bioinformatics.* 2004; 20:3080–3098. [PubMed: 15217818]
56. Zhang J, Chen R, Liang J. *Proteins: Struct, Funct Bioinf.* 2006; 63:949–960.
57. Xu Y, Hu C, Dai Y, Liang J. *PLoS One.* 2014; 9:e104403. [PubMed: 25110986]

Figure 1. The flowchart of DISGRO-H3 applied to loops of length > 6

The H3 base type is predicted using H3-rules.⁵¹ If base type is “kink”, the coordinates of the last N residues are generated backwardly using the torsion angle distribution derived from SabDab training Set. Here N is 2 for short loops ($L < 10$), and $N = 4$ for long loops ($L \geq 10$). If base type is “extend”, conformations are generated using Distance-guided Sequential Monte Carlo method, followed by CSJD closure algorithm without backward growth process. For loops of length ≥ 6 , the growing process is as the same as that of “extend” base type.

Figure 2. Schematic illustration of placing C_i and N_{i+1} atoms

Atom C_i is placed on the circle Q_C . The position $\mathbf{x}_{C,i}$ of the C_i atom of residue i is determined by d_{C_i, C_e} which is based on known distance d_{CA_i, C_e} and the conditional distribution of $\pi(d_{C_i, C_e} | d_{CA_i, C_e})$. Once d_{C_i, C_e} 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 (yellow ball) 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 Q_N and its position $\mathbf{x}_{N,i+1}$ is determined by d_{N_{i+1}, C_e} in a similar fashion.

Figure 3. The dihedral angle distributions of two base types of loops

(A) Kink form. The solid black lines are the base regions, and the dashed grey lines are the β -hairpin regions. (B) Extended form. (C) The distribution of θ_{base} in antibody H3 loops. (D) The distribution of θ_{base} in general loops. (E) The distribution of ψ_{n-1} in “kink” antibody loops. (F) The distribution of ψ_{n-1} in “extended” antibody loops. (G) The distribution of ψ_{n-1} in general loops.

Figure 4. Prediction results of H3 loop for antibody protein 2b4c using DISGRO-H3. The length of the H3 loop is 22. The modeled loops of lowest energy (cyan), minimum RMSD among top 10 lowest energy (blue), and the loop with the minimum RMSD among 20,000 conformations (red) are shown. The native loop is in white.

Table 1

Comparison of best RMSDs of H3 loops generated by RosettaAntibody, H3Loopred, and DISGRO-H3 using the RA set.

Length	# of Targets	RosettaAntibody $R_{E,min,10}$	H3Loopred R_H	DISGRO-H3		
				$R_{E,min,10}$	$R_{E,min,500}$	$R_{min,20000}$
4	1	1.0	1.1	0.74	0.66	0.58
5	1	0.2	0.4	0.33	0.26	0.20
6	1	1.5	1.5	0.43	0.37	0.34
avg of very short	3	0.90	1.0	0.50	0.43	0.37
7	5	1.44	0.90	1.15	0.97	0.63
8	8	1.06	0.96	0.79	0.71	0.59
9	9	0.98	1.10	0.97	0.85	0.65
avg of short	22	1.11	1.00	0.95	0.83	0.62
10	8	1.14	1.11	1.00	0.87	0.72
11	6	1.3	1.55	1.07	0.94	0.70
avg of medium	14	1.21	1.30	1.03	0.90	0.71
12	4	1.95	1.70	1.43	1.25	0.87
13	1	4.7	1.90	2.33	1.60	1.60
14	5	2.36	1.94	1.92	1.53	1.21
avg of long	10	2.43	1.84	1.76	1.43	1.11
17	1	3.50	3.80	2.42	2.22	1.57
18	1	2.7	2.40	2.91	2.77	2.30
19	1	2.3	4.90	1.58	1.58	1.38
22	1	3.3	5.30	2.65	2.50	1.97
avg of very long	4	2.95	4.10	2.39	2.27	1.81
total	53	1.51	1.47	1.21	1.05	0.81

The minimum RMSD is the backbone RMSD value of the structural closest loop conformation to the native conformation in an ensemble consist of a fixed number of conformations. The minimum RMSD of the top 10 lowest energy conformations, the top 500 lowest energy conformations, and all 20,000 conformations are denoted as $R_{E,min,10}$, $R_{E,min,500}$, and $R_{min,20000}$, respectively. R_H is the RMSD of the best template loop to a target loop conformation in H3Loopred study. Results of RosettaAntibody and H3Loopred are obtained from Table 1 of Ref.³⁴ and Table S4 of Ref.³⁶

Table 2

Comparison of R_{Emin} of the loop conformations sampled by RosettaAntibody and DISGRO-H3 using RA set taken from the RosettaAntibody study³⁴ and FREAD study.³⁵

Length	# of Targets	Average prediction accuracy (R_{Emin})		
		RosettaAntibody	FREAD-S	DiSGro-H3
very short (4 – 6)	3	1.27	1.59	0.57
short (7 – 9)	22	1.66	1.23	1.10
medium (10 – 11)	14	1.99	1.46	1.38
long (12–14)	10	3.14	1.62	2.00
very long (17–22)	4	4.05	2.88	2.84
total	53	2.18	1.51	1.44

R_{Emin} denotes the average RMSD of the lowest energy conformations of the loop ensemble. Results of RosettaAntibody and FREAD-S were obtained from Table 1 of Ref.³⁴ and Table S7 of Ref.,³⁵ respectively.

Table 3

Accuracy of the 30 antibody H3 loops modeled by DISGRO-H3 using DISGRO test set.

Length	# of Targets	R_{Emin}	$R_{E,min,10}$	$R_{E,min,500}$	$R_{min,20000}$
4	1	0.28	0.27	0.27	0.27
5	4	0.27	0.25	0.21	0.20
6	1	0.39	0.38	0.33	0.30
very short	6	0.29	0.28	0.24	0.23
7	3	0.79	0.60	0.50	0.48
8	1	1.18	1.14	0.78	0.78
9	2	1.07	0.77	0.67	0.59
short	6	0.95	0.75	0.60	0.57
10	3	1.79	1.41	1.00	0.86
11	3	1.33	1.17	0.95	0.87
medium	6	1.56	1.29	0.97	0.86
12	2	1.66	1.51	1.30	1.04
13	1	1.84	1.67	1.28	1.02
14	1	1.46	0.97	0.97	0.95
15	1	2.12	1.91	1.66	1.51
16	1	2.21	1.62	1.62	1.27
long	6	1.83	1.53	1.36	1.14
17	1	2.68	2.60	2.06	1.77
18	2	2.16	2.13	2.07	1.84
19	1	4.52	3.27	1.90	1.26
22	2	4.13	3.90	3.22	2.74
very long	6	3.30	2.99	2.42	2.03
Total	30	1.58	1.37	1.12	0.96

R_{min} denotes the minimum backbone RMSD values of the loops in an ensemble that consist of a fixed number of conformations. The minimum RMSD of the top 10 lowest energy conformations, the top 500 lowest energy conformations, and all 20,000 conformations are denoted as $R_{E,min,10}$, $R_{E,min,500}$, and $R_{min,20000}$, respectively. R_{Emin} denotes the average RMSD of the loop conformations of lowest energy in the ensemble.