PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Progress in super long loop prediction

Suwen Zhao,[1] Kai Zhu,[2] Jianing Li,[1] and Richard A. Friesner[1]*

[1] Department of Chemistry, Columbia University, New York, New York
[2] Schrödinger, Inc., New York, New York

ABSTRACT

Sampling errors are very common in super long loop (referring here to loops that have more than thirteen residues) prediction, simply because the sampling space is vast. We have developed a dipeptide segment sampling algorithm to solve this problem. As a first step in evaluating the performance of this algorithm, it was applied to the problem of reconstructing loops in native protein structures. With a newly constructed test set of 89 loops ranging from 14 to 17 residues, this method obtains average/median global backbone root-mean-square deviations (RMSDs) to the native structure (superimposing the body of the protein, not the loop itself) of 1.46/0.68 Å. Specifically, results for loops of various lengths are 1.19/0.67 Å for 36 fourteen-residue loops, 1.55/0.75 Å for 30 fifteen-residue loops, 1.43/0.80 Å for 14 sixteen-residue loops, and 2.30/1.92 Å for nine seventeen-residue loops. In the vast majority of cases, the method locates energy minima that are lower than or equal to that of the minimized native loop, thus indicating that the new sampling method is successful and rarely limits prediction accuracy. Median RMSDs are substantially lower than the averages because of a small number of outliers. The causes of these failures are examined in some detail, and some can be attributed to flaws in the energy function, such as π–π interactions are not accurately accounted for by the OPLS-AA force field we employed in this study. By introducing a new energy model which has a superior description of π–π interactions, significantly better results were achieved for quite a few former outliers. Crystal packing is explicitly included in order to provide a fair comparison with crystal structures.

Proteins 2011; 79:2920–2935.
© 2011 Wiley-Liss, Inc.

Key words: long loop build; conformational sampling; computational cost.

INTRODUCTION

One of the most important and challenging tasks in protein modeling is the prediction of loops. A number of different approaches have been developed and over time there has been a gradual improvement of predictive accuracy.[1–16] These methods can be roughly divided into two classes: *ab initio* (conformational search) and database search (or knowledge based) methods. *Ab initio* methods build sterically feasible loop conformations from scratch, while database search methods attempt to find a fragment from databases in the Protein Data Bank that packs optimally into the space available to the target loop. Recently advances in *ab initio* loop prediction have reached remarkable accuracy for loops up to thirteen residues.[6,12,13,16] Significant progress in database search has also been reported.[5,7] Despite considerable effort in both *ab initio* and database search methods, the prediction qualities of most of the methods degrade as loops get longer.[1–12] Despite the fact that the occurrences of loops longer than thirteen residues in Protein Data Bank are not as frequent as for shorter loops, long insertions or deletions are very common in template-based protein structure prediction, and many examples can be found from targets in Critical Assessment of Structure Prediction (CASP, information about CASP can be obtained from http://predictioncenter.org/). Loops of this length can also be quite important from a functional point of view, in many cases lining active site cavities and interacting extensively with ligands binding to the active site. In this situation, it is necessary to use *ab initio* methods to predict the structure in that region. Our goal is to develop an *ab initio* method for super long loops that is both accurate and efficient in its ability to predict native-like conformations.

There are two principal challenges that are common to all high resolution protein structure prediction methods: sampling and scoring. The two issues are in fact entangled with each other in many if not most cases. Many loop prediction approaches begin with the generation of a large number of loops and then use some scoring function to select those candidates that are energetically favorable. Currently, many scoring functions, including the one we use in PLOP,[6,12] can do a very good job in identifying native-like conformations if they are sampled. However, if the conformations generated are not very close to or even far from the native loop, refinement of the model fails to progress adequately. This picture highlights the importance of sampling,

Additional Supporting Information may be found in the online version of this article.
*Correspondence to: Richard A. Friesner, Department of Chemistry, Columbia University, New York, NY 10027. E-mail: rich@chem.columbia.edu.
Received 9 November 2010; Revised 6 May 2011; Accepted 15 June 2011
Published online 22 July 2011 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/prot.23129

© 2011 WILEY-LISS, INC.

especially for long loops. One of the main obstacles in long loop sampling is the combinatorial problem: the number of possible conformations of a loop grows exponentially with the length of the loop. Despite lot of efforts, such as hierarchical loop prediction algorithm with multiple stages, has been made in our group to alleviate this problem,[6,12] sampling failures appear frequently when the current methodology (i.e., that in the released version of PLOP, as described in Refs. 6 and 12) is applied to super long loops.

We have developed a new approach, based on a dipeptide segment sampling algorithm for super long loops, which is implemented in PLOP, our in-house software. When applying to the newly constructed test set with loops ranging from 14 to 17 residues, this algorithm produced substantial improvements as compared to the most recent PLOP loop sampling algorithm reported in the literature. The algorithm is similar to that which we used previously,[6,12] the most important modification is to use dipeptide segment sampling instead of single residue sampling. In the following part of this paper, a brief review of our previous method and a description of our new sampling method will be provided in Materials and Methods. In the Results and Discussion section, we present the results of applying this method to the super long loop test set, and then examine the implications of the results with regard to the using of dipeptide sampling method. We find that a high fraction of results are improved as compared to prior work, but some outliers remain. We have tested the effect of replacing our current energy and solvation model with a preliminary version of a modified model which has been optimized to improve side chain prediction accuracy, and this substantially reduces outliers due to both sampling and scoring failures. Finally, in the Conclusion, we summarize our results and discuss the future directions.

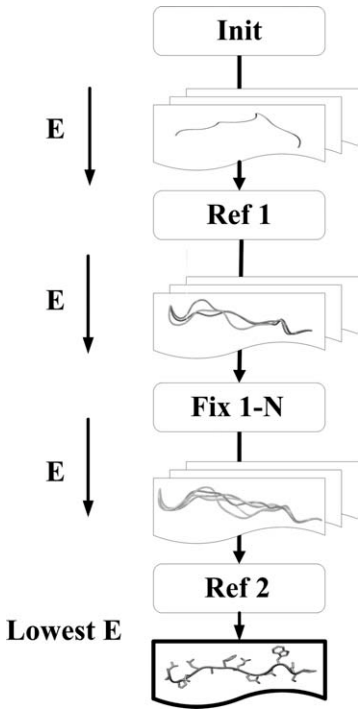

Figure 1

Full loop prediction scheme in protein local optimization program (PLOP). "Init" refers to initial stages of sampling and scoring. "Ref" refers to the refinement stages where sampling is constrained around starting loop conformations. Usually, there are several fixed stages in a full loop prediction. "Fix" refers to the fixed stages, in "Fix1" stage, totally 1 residues from both ends of a loop are fixed at the beginning coordinates during sampling, leaving a relatively easier job to sample shorter loops. Please see Materials and Methods for details.

MATERIALS AND METHODS

In previous work,[6] the Hierarchical Loop Prediction (HLP) algorithm, which is implemented in our in-house software PLOP, has been tested for its ability to reconstruct protein loops in crystal structures for a wide range of loops up to 10 residues. The sampling algorithm and energy function of HLP have been improved in our group for longer loops,[12] ranging in the length from 11 to 13 residues. However, both HLP and the improved method are not effective enough for super long loops. Here super long loops refer to loops longer than 13 residues. In current paper, a new algorithm has been developed, which is a modification of the improved method, introducing a more powerful and robust sampling method designed for super long loops: dipeptide segment sampling assisted by single residue sampling when necessary. A full description of the previous protocols can be found elsewhere.[6,12] We will provide an overview of our previous methods, and then discuss the features of our new method in detail, reviewing previous methods when appropriate.

Overview of previous methods

The previously published methods of a full loop prediction involve a hierarchy of stages. In each stage, there are multiple single loop predictions. The lowest energy loops generated from one stage are passed to the next, where more focused sampling is performed.

Specifically, as shown in Figure 1, the initial structure is passed to three, parallel, initial prediction stages (only one is shown) labeled "Init". The three initial stages differ by the use of five different steric overlap factors: 0.50, 0.55, 0.60, 0.65, and 0.70, respectively. The steric overlap factor is defined as the ratio of the distance between two atom centers to the sum of their van der Waals radii, and specifies the criterion for rejection of an initially constructed candidate loop (i.e., when the steric overlap factor for any atom of the loop backbone exceeds

10970134, 2011, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/prot.23129 by Cas - Shanghaitech University, Wiley Online Library on [18/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

the stipulated threshold, the candidate is discarded). Other steps in the initial structure generation algorithm, such as clustering of similar loops, side chain optimization, minimization, and final selection of low energy structures, are described in References 6 and 12. The resulting five nonredundant (RMSD between any two structures is >0.7 Å) lowest energy structures from each of the initial stages (25 in total) are passed as new starting structures to the parallel refinement stages (only one is shown in Fig. 1) labeled "Ref1".

In the first refinement stage, each model retained from initial stage is subjected to a 4 Å constraint on each Cα atom in the loop during sampling. The sampling protocol discussed above is rerun with this constraint in place, focusing the sampling effort in a substantially smaller region of phase space for each simulation seeded with a given model. Finally, the five nonredundant lowest energy loops from all "Ref1" stages are passed to the next stage, that is, the first fixed stage labeled "Fix1".

Fixed stages are optional, but they significantly improve the accuracy for long loops; the algorithm is described in more detail in Reference 12. Usually, there is more than one fixed stages in a full loop prediction, they are denoted as "Fix1", "Fix2", etc. In each fixed stage, several short fragments of the target loop will be predicted for each model whereas other parts of the loop are fixed on their given conformations. Here, we take a 14-residue loop prediction as an example. In "Fix1" stages, one terminal residue in either end of the loop is fixed, so there are two possible positions for each 13-residue loop. Then we do two 13-residue loop predictions on each model and pass a few best (that is, nonredundant and lowest energy) resulting models to "Fix2" stages. In "Fix2" stages, totally two terminal residues from both ends of the loop need to be fixed, so there are three possible positions for each 12-residue loop. Three 12-residue loop predictions will be performed on each model and then several best resulting models will be passed to the next stage, that is, stage "Fix3". In "Fix3" stages, totally three terminal residues from both ends of the loop need to be fixed, four 11-residue loop predictions will be performed for each model and so on. After the last fixed stages, a few best models are passed to the second refinement stages (only one is shown in Fig. 1) labeled as "Ref2".

In stage "Ref2", the Cα atoms in the loop are constrained to less than 2 Å from their starting coordinates. Loop prediction proceeds as in the Ref1 stage discussed above, with the tighter constraint to the initial seed providing a still greater focusing of phase space in performing the sampling. Finally, the lowest energy loop from all stages is taken as the predicted loop.

A full loop prediction algorithm involves multiple executions of PLOP, enabling additional sampling effort to be focused on loop subsections that are constrained to lie in regions of phase space that have previously identified as promising. In each stage of a full loop prediction, multiple single loop predictions are executed with different input parameters. One single loop prediction corresponds to the execution of the PLOP once. Each single loop prediction consists of four basic stages: buildup, clustering, loop side chain optimization and complete energy minimization.

Buildup refers to the generation of an initial set of loop conformations that will be passed on to the subsequent stages. The cornerstone of our previous loop sampling methodology is dihedral angle search, which is conducted via "rotamer libraries" for backbone dihedral angles (i.e., discretized version of the well-known Ramachandran plot). To obtain the dihedral angle libraries, a large (>500 structures), nonredundant database of high-resolution (<2 Å) protein crystal structures have been used and every backbone dihedral angle was recorded. The dihedral angles were then binned every 5°, and every (ϕ, ψ) combination that appeared more than five times in the database was included in the backbone library. The resultant library, at 5° resolution, contains 747 (ϕ, ψ) combinations for Gly, 215 for Pro, and 866 for all other residue types.

The extreme high resolution of the backbone libraries was chosen to ensure that discretization error does not fundamentally limit the achievable accuracy. However, it also magnifies the exponential scaling of conformational space and computational expense. In practice it is not possible to sample the backbone dihedral angle for a loop of any nontrivial length using such a large library. Effective sampling resolution is used to alleviate the problem. For a single residue, first the entire list of (ϕ, ψ) combinations is screened for steric clashes. For computational efficiency, the screening is accomplished through the use of the overlap factor. Multiplying the overlap factor by the sum of the van der Waals radii for a given pair of atoms yields a cutoff. A clash occurs when the distance between the atoms is less than the cutoff. The default overlap factor of 0.70 is used in PLOP. Then, the screened rotamers are further filtered, retaining a set of (ϕ, ψ) states in which all pairs of states obey the relation: $\Delta\phi^2 + \Delta\psi^2 > R_{\text{eff}}^2$, here R_{eff} is the "effective resolution." That is, the distance between any two states in the set should be greater than certain cutoff.

The effective sampling resolution is chosen in an adaptive manner. The strategy is to define in advance the minimum and maximum number of loops to be generated for a particular loop length. If n is the loop length in residues, for loops no longer than nine residues, the minimum number is 2^n; for loops longer than nine residues, the minimum number does not increase with the loop length any more, but fixes at 2^9, that is, 512. The maximum number is 10^6 for loops with any length. The effective resolution is then gradually decreased from a very coarse value (300°), until the number of generated loops is intermediate between the minimum and maximum numbers.

10970134, 2011, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/prot.23129 by Cas - Shanghaitech University, Wiley Online Library on [18/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Loop buildup starts from both ends of the loop independently. Closed loops are generated by applying a loop closure algorithm in the middle of the loop. The all closed loops are subjected to a series of screens: N-Cα-C angle, backbone dihedral angles and sufficient space for the side chain on the closure residue, and steric clashes between the two halves of the loop.

The dihedral angle-sampling buildup procedure can generate many millions of loop candidate structures. These loops have already been screened to ensure no steric clashes and sufficient space for the side chains. Then we use the K-means clustering algorithm to select representative loop candidates.

For each loop representative, the loop side chains are then added and optimized. The sampling of single side chain conformations was accomplished by using a highly detailed (10° resolution) rotamer library constructed by Xiang and Honig from a database of 297 proteins.[31] This library contains, for example, 2086 rotamers for lysine. The additional computational expense of such a detailed library was tolerated in order to ensure adequate sampling. In addition, the expense was mitigated by prescreening the rotamers using only hard sphere overlap as a criterion, allowing many rotamers to be excluded before performing any energy evaluations. All loop side chains are initially built on to the fixed backbone in a random rotamer state (each side chain is randomly picked from its pre-screened clash-free rotamer sets), and then each side chain in the loop is optimized one at a time, holding the others fixed. Here, the optimization means choosing the side chain rotamer with lowest energy. The procedure is iterated to convergence, that is, the optimization stops until no side chain(s) changing rotamer states compared to the previous iteration. The very details can be found in our previous paper.[17]

After convergence is achieved, each representative loop is completely energy minimized in Cartesian coordinates to (relax the backbone and side chains). Finally, side chain optimization described above is applied again (this time with a different backbone structure). Energy is calculated for the final structure of each representative loop. The loop with the lowest energy is chose as the predicted structure.

There are two aspects that contribute to the success of the sampling algorithm in our previous methodology. The first refers to the way that a single loop prediction builds up many candidates, screens and clusters them, and picks out the representatives for scoring and ranking; the second refers to a full loop prediction that utilizes a variety of ways to sample the low-energy region of conformational space via many PLOP calls in which the options for each execution vary. The second aspect plays a greater role as loops get longer.

An all-atom force field energy with implicit solvent is calculated for each sampled loop, and the loops are then ranked by energy. The energy is calculated using the Optimized Potential for Liquid Simulations (OPLS) all-atom force field,[18–20] the Surface Generalized Born (SGB) model of polar salvation,[21] an estimator for the nonpolar component of the solvation free energy developed by Gallicchio *et al.*,[22] and a number of correction terms as detailed in Ghosh *et al.*[21] and in Jacobson *et al.*[20] In addition, Zhu *et al.*[12] have also incorporated an additional hydrophobic term adapted from the ChemScore[23] scoring function, which has been successfully used to describe the hydrophobic contribution to the binding free energy between ligands and protein receptors. The hydrophobic term appears to approximately fix a major flaw in SGB solvation model described above.

Crystal packing is included. The simulation system consists of one asymmetric unit and all atoms from other surrounding symmetric units that are within 30 Å. Every copy of the asymmetric unit is identical at every stage of the calculation.

Dipeptide sampling method

In this work, we develop a dipeptide segment based sampling algorithm, whereas the previous single residue based sampling algorithm is still kept as an important supplement. The single residue based dihedral angle sampling (referred to as single residue sampling) algorithm has achieved great success for a wide range of loops no longer than 10 residues in PLOP. However, as loops get longer, the power of the single residue sampling method decreases gradually: The number of generated loop conformations increases explosively and a lot of redundancies are generated.[6,12] This is easy to understand, because the relation between dihedrals and backbone coordinates is nonlinear, a number of combinations of backbone dihedrals can lead to a very similar loop conformation. For example, essentially any backbone dihedral can be rotated by a large amount and the rest of the loop can stay very similar if there are a series of small compensating changes in the remaining dihedrals. This problem becomes worse as loops get longer. As a result, the following two scenarios have been frequently seen, when predicting loops long than 13 residues. First, in the buildup stage of a super long loop, the generated loop candidates can easily break through the minimum limit, which is 512. For loops longer than nine residues, it could happen even when the effective sampling resolution is rather coarse. Effective sampling resolution is a threshold to select distinguishable backbone rotamers. PLOP starts with a high value (300°) of effective sampling resolution and decreases gradually until the generated loops exceed the minimum number of conformations required by the program. A lot of cases show that the minimum can be easily broken through when the effective sampling resolution stays as high as 60–80°, which means only a few backbone rotamers are selected for each residue. Thus the overall quality and variety of

10970134, 2011, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/prot.23129 by Cas - Shanghaitech University, Wiley Online Library on [18/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table I
Number of Rotamers in 400 Dipeptide Libraries

	G	P	D	S	N	T	E	K	R	Q	H	A	L	V	I	F	Y	M	C	W
G	**1098**	516	**882**	823	679	744	636	727	575	405	280	843	803	697	548	412	408	189	156	190
P	749	377	578	475	401	363	470	307	290	217	178	442	430	372	214	282	231	93	66	88
D	**970**	486	472	428	315	337	358	369	277	163	160	388	417	315	239	227	203	83	87	93
S	763	417	473	451	352	339	279	335	237	217	156	364	313	222	168	181	153	84	75	68
N	705	449	391	352	345	318	254	290	208	181	121	314	299	269	217	156	130	62	58	84
T	594	440	444	376	326	299	249	192	177	167	102	326	240	190	171	155	118	48	73	50
E	654	401	407	295	314	272	272	300	212	125	129	276	241	185	152	112	139	56	45	47
K	550	440	441	290	341	266	279	244	195	139	88	251	260	198	174	111	130	62	59	54
R	492	344	266	269	218	238	220	180	182	103	113	204	190	140	129	124	126	61	52	46
Q	360	283	214	199	175	194	159	125	103	78	70	174	207	133	96	67	68	41	37	40
H	259	266	149	172	107	111	98	71	95	66	49	136	102	88	80	55	55	22	21	27
A	814	545	593	383	344	366	303	306	229	151	161	425	242	233	137	137	143	81	64	43
L	596	771	512	477	348	369	308	307	235	188	109	265	193	182	112	123	97	53	56	35
V	403	559	420	310	265	259	244	247	184	139	68	239	163	171	126	104	89	29	45	28
I	293	490	311	291	230	215	163	160	93	85	72	173	101	100	70	69	79	21	29	13
Y	290	299	252	192	178	138	108	133	119	97	54	146	94	85	67	73	71	15	41	26
F	295	344	299	261	180	159	146	149	94	104	79	152	109	128	71	57	64	24	31	30
M	161	157	107	86	100	88	68	64	50	33	23	47	43	33	20	40	24	20	12	6
C	156	132	105	96	65	65	47	45	52	33	49	60	49	41	30	24	20	18	13	11
W	119	105	116	89	61	54	54	51	55	36	30	54	34	24	20	33	28	11	17	24

For each number, its row amino acid is the N terminal residue in the dipeptide, and its column amino acid is the C terminal residue in the dipeptide. Except three libraries (Gly-Gly, Gly-Asp, and Asp-Gly, marked as bold), other libraries apparently have less rotamers than the old single residue backbone library shared by all residues expect Gly and Pro, which has 866 rotamers.

generated loops are very limited. Second, frequently there is a sudden explosion of loop candidates when the effective resolution is decreased by a step (1–10°), sometimes the maximum limit (10^6) could be reached. Then PLOP gets stuck there, takes forever to finish the closure and screening of loops, and often crashes the node. These situations happen to thirteen-residue loops sometimes, but even more frequently for loops that are longer.

A sampling algorithm that could bring down the huge number of generated loop candidates in the buildup stages as well as increase the realized effective sampling resolution is needed and has recently been developed in our group. The algorithm is mainly based on dipeptide segment sampling (referred as dipeptide sampling), assisted by single residue sampling when necessary. Dipeptide sampling is similar to single residue sampling, but differs in the step size in torsion angle phase space. The step size of single residue sampling is (ϕ, ψ), and the step size of dipeptide sampling is ($\phi1$, $\psi1$, ω, $\phi2$, $\psi2$). The first step is to build the dipeptide libraries. To build the libraries, we have used a large (3799 structures), diverse (percentage identity <30%) database of high resolution (<2 Å) protein crystal structures and recorded every backbone dihedral angle in loop regions. The dihedral angles were recorded in the format of sets, with each set having five consecutive dihedral angles ($\phi1$, $\psi1$, ω, $\phi2$, $\psi2$). That is, each set is corresponding to a dipeptide segment, and the set is labeled by the name of the dipeptide it corresponds to. We only deal with the first twenty genetically encoded amino acids (not include Selenocysteine and Pyrrolysine) in our program, so totally there are 400 (20 by 20) different dipeptides, such as Ala-Lys, Pro-Gly, and Thr-Phe, etc. Then all sets were attributed to one of the 400 groups, according to their labels. The angles in each group were then binned every 5°, and all ($\phi1$, $\psi1$, ω, $\phi2$, $\psi2$) combinations appeared in the group were included in the backbone library. The resultant 400 libraries, at 5° resolution, have 84,711 sets of ($\phi1$, $\psi1$, ω, $\phi2$, $\psi2$) combination. That is, we have added 84,711 backbone rotamers to our previous libraries, which only have 747 rotamers for Gly, 215 rotamers for Pro, and 866 rotamers for all other residues. The number of dipeptide rotamers in each library is given in Table I.

In the single residue libraries, the backbone library shared by 18 residues (except Gly and Pro) has 866 rotamers, while the average number of rotamers of the dipeptide libraries in Table I is 211. Actually, only three dipeptide libraries (Gly-Gly, Gly-Asp, and Asp-Gly) have more than 866 rotamers. The reason for less rotamers in most of the dipeptide libraries is pretty obvious. A large number of conformations for any two consecutive residues generated by the single residue sampling method have steric clashes between the two residues (including their side chains), and such sterically infeasible conformations have been excluded naturally when building the dipeptide libraries. In single residue sampling method, the correlation between side chains of two consecutive residues are not considered, remembering that 18 different residues share one backbone library, though obviously tryptophan only needs a much smaller library than alanine; furthermore, during the buildup stage, the steric clash screening on side chain is accomplished through a rather coarse approach, by testing all conformations in a 30°side chain rotamer library, until one is found that is

10970134, 2011, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/prot.23129 by Cas - Shanghaitech University, Wiley Online Library on [18/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

free of steric clash, with other side chains of the loop not included in these screens. The dipeptide libraries carry more information than our single residue libraries. Using the dipeptide libraries makes the sampling more efficient: it greatly brings down the huge number of the generated loop candidates in buildup stages; at the same time, the realized effective sampling resolution has also been improved a lot.

The 5° resolution keeps unchanged in the dipeptide libraries, in order to make sure the discretization error does not fundamentally limit the achievable accuracy. The high resolution offers a solid foundation for the accuracy of this method.

Effective sampling resolution still plays an important role in the new sampling algorithms. There is a slightly change in the effective sampling resolution by the implementation of dipeptide sampling. The old relation obeyed by filtered states is: $\Delta\phi^2 + \Delta\psi^2 > R_{\text{eff}}^2$, whereas the new relation is: $(\Delta\phi 1)^2 + (\Delta\psi 1)^2 + (\Delta\omega)^2 + (\Delta\phi 2)^2 + (\Delta\psi 2)^2 > R_{\text{eff}}^2$. With the same effective sampling resolution R_{eff}, the constraints on $\phi 1$, $\psi 1$, ω, $\phi 2$, and $\psi 2$ are tighter than that on ϕ and ψ.

Our buildup procedure continues independently from both sides of the loop up to the Cα atom on the closure residue, the length (in residues) of a half loop can be even or odd, but the step size of dipeptide sampling is fixed at two residues. If a half loop has odd length, then the residue at the end of the half loop, that is, the residue next to the closure residue will be left aside. There are several ways to solve this problem. In this work, when constructing a half loop with odd length, we choose to use the single residue sampling method to add the last residue.

The following three stages in a single loop prediction, that is, closure, clustering and scoring, are similar to that in our previous work.[6,12] The only difference is the method of adaptively increasing the number of clusters in K-means clustering algorithm.[24,25] For K-means algorithm, the number of clusters must be prespecified. We empirically use four times the number of loops residues as the default value of number of clusters. Sometimes, after the initial clustering, there are some clusters that have a much larger spread (as judged by intra-cluster RMSD) than others. At this time, a simple method to adaptively increase the number of clusters is used. Clusters with a large spread are split into three new clusters and the clustering algorithm is run again with the new clusters added. This procedure can be applied iteratively until the maximum number of cluster (sets to be four times the number of loop residues plus 30) is reached.

Test set

Loop length verses frequency of loops with the length appeared in native proteins is shown in Figure 2. Statistics are calculated from ~4000 low sequence identity (<30%), high quality (<2 Å) proteins structures.

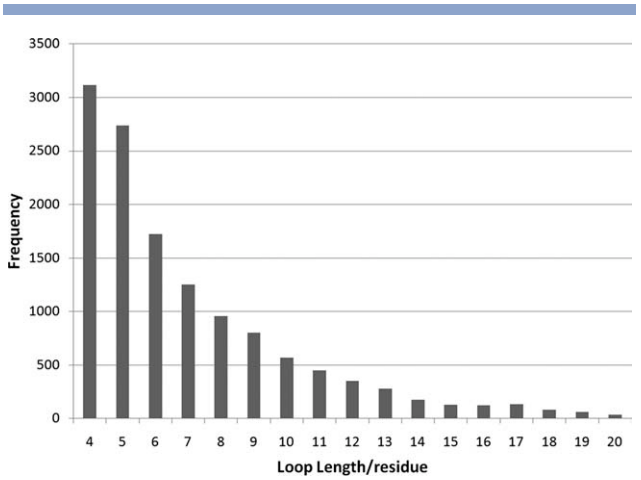

Figure 2

Distribution of loops with different lengths. Statistics are calculated from ~4000 low sequence identity (<30%), high quality (<2 Å) proteins structures. Secondary structure assignment is done by DSSP approach. Loops are defined as the regions outside helices and strands. Loops shorter than four residues and longer than 20 residues are not shown in the figure.

Secondary structure assignment is done by DSSP approach, while loops are defined as the regions outside helices and strands. The relative rareness of loops longer than 13 residues puts a cap on the size of our test set. To using inadequately small data sets, we try to collect as many loops as we can, with the selection criteria described below. Finally, a test set with 89 target loops has been constructed, in which we have 36 fourteen, 30 fifteen, 14 sixteen and nine seventeen-residue loops.

Selection criteria

The selection criteria of our test set are similar to that we used before.[6,12] The PISCES web server (http://dunbrack.fccc.edu/PISCES.php) from Dunbrack's lab is used to generate the protein list, from which loops targets are collected. The following criteria are used to ensure the selection of high-quality protein structures.

1. Low sequence identity. The percentage identity between any two proteins has to be <30%.High resolution. The resolution of a structure has to be <2.0 Å.
2. Crystal structure. Only structures obtained from X-ray crystallography are selected.
3. With both backbone and side chain information. Structures with only Cα coordinates are excluded.
4. Low *R*-factor. The *R*-factor cutoff is set to 0.25.
5. Decent pH values. Only structures obtained under pH 6–8 are selected.

Target loops are then chosen from the proteins selected out by the above criteria. More rigorous criteria than that in our previous work have been applied to the target

10970134, 2011, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/prot.23129 by Cas - Shanghaitech University, Wiley Online Library on [18/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

loop selection, which greatly increases the reliability of our data set.[6,12] The major difference is that we introduce the real space *R*-factor criterion in this work, which is a is a more sophisticated description of structure quality than B-factor. Below we list the criteria we used to select target loops:

1. The average temperature factor (B-factor) of atoms within the loop has to be <35.00.
2. The real space *R*-factor (RSR) of any residues in a selected target loop must not be greater than 0.200.
3. Any residue of a selected target loop must not have alternative structures.
4. The minimum overlap factor for any loop-atom has to be ≥0.70.
5. The minimum distance between any loop atom and any atom from a neutral ligand or organic ion in the environment has to be >4.0 Å. For a metal ion ligand, this cutoff is 6.5 Å.
6. A selected loop must not contain any chain breaks, and the protein body must not contain any chain breaks that are close to (in five residues) any end of the loop.
7. A selected loop must not contain any secondary structure content longer than three residues.
8. The presence of no less than four strand or helix residues on either end of selected loop.

The employment of temperature factor, overlap factor, and real space *R*-factor cutoffs aims at better quality of the structure in the loop region. It is for the same goal to filter away the structures with chain break and alternative structures. The distance cutoff for ligands and ions mitigates concerns as to whether the potential function is accurate for ligands and ions that potentially interact with the loop. The goal of all criteria enumerated above is to focus on evaluating the ability of the sampling algorithm and energy model to yield accurate loop predictions, without considering additional issues.

RMSD calculation

The accuracy of loop prediction is evaluated by comparing it with the native conformation of the loop. A large variety of reasonable criteria for comparing loop conformations exist. The RMSD can be calculated from the superposition of the whole structures excluding the loop ("global" superimposition) or from the superimposition of the compared loop atoms only ("local" superimposition). In this work, our results are reported as "global" backbone RMSDs, obtained by superimposing the body of the protein (excluding the loop) and calculating the RMSD of the loop, using N, Cα, C and O atoms in loop backbone.

Protonation states of titratable residues

All titratable residues are placed in their standard protonation states at pH 7.0. Instead of sampling the protonation states or varying them under specified pH conditions, this is a good approximation as shown by our results. However, the prediction accuracy for a small number of cases might be affected due to such an approximation. Possible solutions have been discussed in other works.[26]

Table II
Average and Median Loop Prediction Accuracy of the New Sampling Algorithm, Summarized by Loop Length

Loop length	Number of cases	Mean	Median	Standard deviation
14	36	1.19	0.67	1.57
15	30	1.55	0.75	1.82
16	14	1.43	0.80	1.76
17	9	2.30	1.92	2.63
All	89	1.46	0.68	1.81

The accuracy is measured by global backbone (N-Cα-C-O) RMSDs in Å.

RESULTS AND DISCUSSION

Overall performance

Table II gives the summarized results obtained by our new sampling algorithm. The methodology obtains uniformly good results for 14- to 17-residue loop data sets, with overall average/median RMSD 1.46/0.68 Å. The median RMSD for each subset ranges from 0.67 to 1.92 Å. For 36 fourteen-residue loops, only three cases (8.1%) have RMSD larger than 2.0 Å. Fifteen-, sixteen-, and seven-residue loop data sets have 22 out of 30 (73.3%), 11 out of 14 (78.6%) and 6 out of 9 (66.7%) cases with RMSD below 2.0 Å. A summary of the results is shown in Table III. Figure 3 compares the prediction loop and the native one in the context of the full-length protein in four selected cases.

Our results clearly represent a major advance as compared to our own previous loop prediction work, producing significantly lower median and mean RMSDs for 14 residue loops than we reported in Reference 30 for 13 residue loops. Direct comparison with the work of other groups in this area, even those carrying out *ab initio* prediction as opposed to using knowledge based methods, is difficult for a number of reasons. Two of the most significant are as follows: (1) we are including crystal packing effects explicitly, whereas most other studies do not include these effects; (2) we are employing specific data sets with highly restrictive criteria applied to filter the test cases; other methods might well perform better on this test set than on the test sets that were actually employed by them in prior evaluations. As our objective in this paper is to test the sampling algorithm and energy model as carefully as possible against experiment, we believe that the use of crystal packing and elimination of loops with problematic resolution leads to the fairest comparison of this type. We have also run a small subset of cases, for which we determined that crystal

10970134, 2011, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/prot.23129 by Cas - Shanghaitech University, Wiley Online Library on [18/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table III
Results of 14, 15, and 17-Residue Loops, Obtained by Using the Dipeptide Sampling Method

No.	PDB	Starts	Ends	Length	Sequence	RMSD (Å)	dE (kcal/mol)
1	1E6U	A:274	A:287	14	ASKPDGTPRKLLDV	1.94	−7
2	1JP4	A:153	A:166	14	PYYNYQAGPDAVLG	7.26	−43
3	1N0Q	A:24	A:37	14	AGADVNAKDKNGRT	0.22	−4
4	1N0Q	A:57	A:70	14	AGADVNAKDKNGRT	0.52	−13
5	1O97	D:156	D:169	14	RPSVFKPLEGAGSP	0.82	−13
6	1OCK	A:209	A:222	14	GRSEFSGIVPAKAP	0.93	−22
7	1P3C	A:112	A:125	14	GYRSIRQVTNLTGT	0.32	−7
8	1P3D	A:402	A:415	14	DVYAAGEAPIVGAD	1.74	6
9	1R6X	A:72	A:85	14	SRLADGTLWTIPIT	0.30	1
10	1RDQ	E:273	E:286	14	LQVDLTKRFGNLKN	1.50	−11
11	1RV9	A:225	A:238	14	GTHCTVLERDTFFS	0.26	9
12	1VYR	A:193	A:206	14	SPSSNQRTDQYGGS	0.58	4
13	1VYR	A:235	A:248	14	SPIGTFQNVDNGPN	1.17	6
14	1XU1	A:221	A:234	14	PRANAKLSLSPHGT	0.47	4
15	1ZEQ	X:53	X:66	14	ITPQTKMSEIKTGD	0.27	−15
16	2BWR	A:269	A:282	14	KDFGVNSGWRVEKH	0.44	−3
17	2BWR	B:158	B:171	14	NNFGYAQGWRLDRH	4.56	35
18	2C0H	A:40	A:53	14	QAWVNYARDFGHNQ	5.96	108
19	2EX2	A:139	A:152	14	TSIFASHDKAPGWP	0.21	−8
20	2GGC	A:79	A:92	14	HGIPDDAKLLKDGD	0.24	−8
21	2H3L	A:1360	A:1373	14	QPEGPASKLLQPGD	0.28	1
22	2O2K	A:1221	A:1234	14	SNLKSKYFAVGKIS	1.07	−29
23	2PVQ	A:139	A:152	14	LSDKNAYWLGDDFT	0.66	6
24	2VFR	A:325	A:338	14	AADAQWLSPAYGRD	0.34	−14
25	3B40	A:389	A:402	14	SDFNDGGGVDGWKD	1.28	48
26	3B64	A:44	A:57	14	DSTPMHFFGSTDPV	0.65	−25
27	3BY9	A:177	A:190	14	DLSAIEQGWQNKSS	1.13	−24
28	3BY9	A:205	A:218	14	SQPAWLFHSVADLS	0.28	−11
29	3CFZ	A:125	A:138	14	TDKSKYKDEINSTN	0.68	1
30	3CNQ	S:50	S:63	14	FVPSETNPFQDNNS	1.03	−9
31	3CSS	A:163	A:176	14	FGSDGHTASIFPDS	0.21	0
32	3DRF	A:550	A:563	14	KRVVGMTLDYGAMN	1.85	1
33	3E7H	A:67	A:80	14	GKVSAKDYVNEATG	1.78	−6
34	3EHR	A:95	A:108	14	NRVGVNGLDKAGST	0.94	7
35	3FOT	A:164	A:177	14	DVSTDSTPIPQDAT	0.30	−14
36	3HXL	A:277	A:290	14	ARVNESLTYQGYDE	0.79	28
37	1AH7	A:157	A:171	15	KVTDGNGYWNWKGTN	0.32	−1
38	1BHE	A:121	A:135	15	GQGGVKLQDKKVSWW	0.42	−1
39	1H4A	X:19	X:33	15	SSDHPNLQPYLSRCN	0.28	−4
40	1JU3	A:486	A:500	15	RETLVNPTLIEAGEI	0.35	5
41	1QAZ	A:298	A:312	15	DKSARAQASGPLRGI	1.68	−7
42	1QQF	A:1112	A:1126	15	QKPDGVFQEDGPVIH	0.31	−8
43	1RA0	A:283	A:297	15	QGRFDTYPKRRGITR	2.78	−39
44	1RA0	A:361	A:375	15	LNLQDYGIAAGNSAN	0.39	−4
45	1RYO	A:172	A:186	15	QLCPGCGCSTLNQYF	0.88	−5
46	1S95	A:477	A:491	15	TAVPHPNVKPMAYAN	0.61	−5
47	1WB4	A:1033	A:1047	15	ALPHFDYTSDFSKGN	0.21	1
48	1WUI	L:454	L:468	15	KGDNVICAPWEMPKQ	1.81	−12
49	1Y12	A:10	A:24	15	GDVKGESKDKTHAEE	0.36	−2
50	1ZHX	A:392	A:406	15	NLSTKNAPSGTLVGD	7.10	67
51	2AEB	B:156	B:170	15	IPDVPGFSWVTPCIS	2.55	25
52	2B0T	A:701	A:715	15	VQGGATDLGGYYSPN	1.23	3
53	2CJP	A:58	A:72	15	DLRGYGDTTGAPLND	0.46	12
54	2DSJ	A:354	A:368	15	GGGRKRKGEPIDHGV	0.51	−5
55	2H3L	A:1339	A:1353	15	GVGGRGNPFRPDDDG	1.10	−10
56	2O2K	A:1220	A:1234	15	FSNLKSKYFAVGKIS	1.36	−29
57	2OIT	A:290	A:304	15	FMEPCYGSCTERQHH	0.54	6
58	2PKF	A:26	A:40	15	LPEHLHKVSLSFLVD	2.34	−8
59	2V3V	A:382	A:396	15	WGLPEGRIAPEPGYH	0.35	4
60	3A3P	A:286	A:300	15	DSNDNIASFSNRQPE	0.18	5
61	3A64	A:350	A:364	15	SPASHVPAPEAGEWF	2.55	3
62	3BB7	A:231	A:245	15	SEMQYGGPNEGSGAY	6.26	−19
63	3BF7	A:49	A:63	15	DVRNHGLSPREPVMN	5.66	72
64	3CSS	A:95	A:109	15	LLRDVPSSDVISIDR	2.36	−24

(Continued)

10970134, 2011, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/prot.23129 by Cas - Shanghaitech University, Wiley Online Library on [18/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table III
Continued

No.	PDB	Starts	Ends	Length	Sequence	RMSD (Å)	dE (kcal/mol)
65	3EA1	A:136	A:150	15	YFVDPIFLKTEGNIK	0.49	7
66	3F1L	A:99	A:113	15	NAGLLGDVCPMSEQN	1.05	−16
67	1C1K	A:31	A:46	16	NGKYDVIKYNWCMRVS	0.66	12
68	1DJ0	B:19	B:34	16	DGSKYYGWQRQNEVRS	7.08	72
69	1GPI	A:308	A:323	16	NSVANIPGVDPVNSIT	0.33	−5
70	1UG6	A:340	A:355	16	GAAYPDLWTGEAVVED	0.43	4
71	1WHI	A:88	A:103	16	RDDKSPRGTRIFGPVA	0.66	−17
72	1WM3	A:67	A:82	16	INETDTPAQLEMEDED	0.32	13
73	1ZHV	A:20	A:35	16	SASEAIPAWADGGGFV	0.64	−10
74	2BG1	A:708	A:723	16	SPSIWGNERFALDPSV	2.15	−4
75	2GGC	A:184	A:199	16	LHYDSRETNVVLKPGM	1.08	−89
76	2HKJ	A:418	A:433	16	TKIPYKSAGKESIAEV	0.42	−42
77	2PKF	B:25	B:40	16	LLPEHLHKVSLSFLVD	0.93	−93
78	2PUH	A:70	A:85	16	DSPGFNKSDAVVMDEQ	1.52	−14
79	2PYW	A:321	A:336	16	NKDGPGEAYLADIYNN	2.57	−37
80	3IFE	A:14	A:29	16	KIDTQSNEDSHTVPTT	1.18	−55
81	1KWG	A:314	A:330	17	QPGPVNWAPHNPSPAPG	1.93	−41
82	1QLW	A:145	A:161	17	FRFGPRYPDAFKDTQFP	0.41	0
83	1VJU	A:277	A:293	17	LPPRARWGYNWQPEPGT	0.63	−2
84	2FAO	A:814	A:830	17	AGDDPWADYAGTRQRIS	0.50	21
85	2HDW	A:131	A:147	17	DPSYTGESGGQPRNVAS	2.22	−126
86	2PEF	A:191	A:207	17	TNGMIDKILNKIDPEDV	1.92	44
87	3A3P	A:262	A:278	17	SGNEGAPSPSYPAAYPE	0.52	−250
88	3H2G	A:124	A:140	17	DYLGLGKSNYAYHPYLH	4.07	56
89	3HUH	A:71	A:87	17	QEMEFEPKASRPTPGSA	8.56	−39
Average						1.46	−7
Median						0.68	−4
Standard Deviation						1.81	41

RMSD is "global" backbone RMSD, dE is the energy difference between predicted and native loop.

packing effects on the target loop are unimportant, without crystal symmetry; these results are shown below. Explicit comparison between programs will require running these programs under identical conditions on the identical test set, and is reserved for future publications.

Dipeptide sampling

Considering the high accuracy of our loop prediction results, it is the dipeptide sampling method that contributes the most significantly to the improvement. The chief problem of the high resolution dihedral angle sampling method is that the number of generated conformations scales exponentially with loop length, so does the computational expense of the sampling step. For years, a lot of efforts have been made to alleviate this problem in our group, and such effort include as effective sampling resolution and hierarchical loop prediction algorithm with fixed stages previously developed. All these techniques mitigate this problem to some extent. However, our loop prediction algorithm with all these techniques still fails frequently when applied to loops longer than 13 residues.

By switching to dipeptide sampling from single residue based sampling, our full loop prediction algorithm shows its powerful search ability for super long loops. It generally eliminates significant sampling errors (defined as big outliers (RMSD >4.0 Å) with higher energy than native structure) in our results. For our 89 cases, there are eight cases (2BWR and 2C0H in 14-residue subset, 1ZHX, 2EAB and 3BF7 in 15-residue subset, 1DJ0 in 16-residue subset, 3H2G and 3HUH in 17-residue subset) showing significant sampling errors, according to an analysis of energy gaps. Further examinations are needed to exclude other factors (such as underestimated π–π interactions, and protonation states misassignment) that might cause the sampling errors.

The correlation between the two adjacent residues in dipeptide is considered in the dipeptide libraries, so some structure information not shown in the single residue libraries is now embedded in the dipeptide libraries. As a result, when searching through the information rich dipeptide libraries to build up loops, a lot of redundancies generated by single residue sampling should be removed. This is exactly what happened in our results: the huge number of generated loop candidates is greatly reduced and finer effective sampling resolution is achieved.

Figures 4 and 5 compare the dipeptide sampling method with the single residue sampling method in two aspects: realized effective sampling resolution and number of generated loop candidates. 13-Residue loops were selected as a test set for direct comparison using the criteria described above. The summary of the results is shown in Table IV.

10970134, 2011, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/prot.23129 by Cas - Shanghaitech University, Wiley Online Library on [18/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

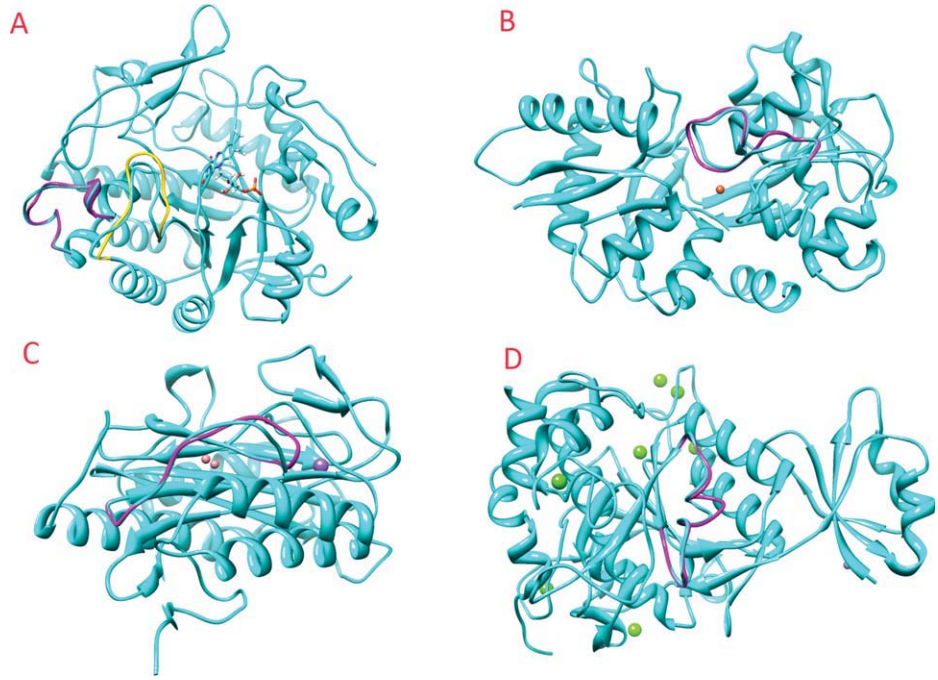

Figure 3

Four selected cases that show the predicted native loops in the context of the full length protein. Panels A, B, C, and D are 14-, 15-, 16-, and 17-residues loop, respectively. Protein body and native loops are in cyan. Predicted loops are in magenta or gold. Panel A: 1VYR A:193–206 (in magenta), with RMSD 0.58 Å; and A:235–248 (in gold), with RMSD 1.17 Å. Panel B: 1RYO A:172–186, with RMSD 0.88 Å. Panel C: 2GGC A:184–199, with RMSD 1.08 Å. Panel D: 3A3P A: 262–278, with RMSD 0.52 Å. This image and Figure 7 were generated using the program UCSF Chimera.[30]

Effective sampling resolution is the threshold to select distinguishable rotamers, it is directly related to the quality and variety of generated loop candidates. Finer realized effective sampling resolution often implies better results. Figure 4 shows the average realized effective sampling resolutions in each sampling stage for the 13-residue loop set, using single residue and dipeptide sampling, respectively. In each sampling stage, when using the dipeptide sampling method, the average realized effective sampling resolution is improved by 25–45°, compared to that when using single residue sampling. One should keep in mind that longer loops are much more difficult to predict, and usually have coarser realized effective sampling resolution. So, it is not difficult to imagine that the discrepancies in Figure 4 would be even larger if the two sampling methods were applied to loops longer than 13 residues. We chose 13-residue loops to do the comparison, simply because single residue sampling method has terrible performance on loops longer than that length: either the required CPU time is extremely high, or the job could crash in the middle, due to the huge number of generated conformations.

Meanwhile, Figure 5 shows that, in each sampling stage, the average number of generated loops for the 13-residue loop set, using single residue and dipeptide sampling, respectively. In each sampling stage, when

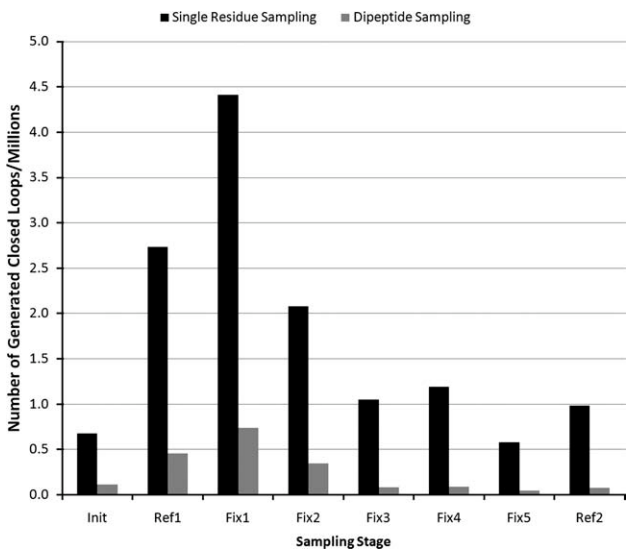

Figure 4

Comparison of the average realized effective sampling resolution of seventeen 13-residue loops for the two sampling methods, in each sampling stage. The blue columns are results of single residue sampling method, whereas the red columns are results of dipeptide sampling method.

10970134, 2011, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/prot.23129 by Cas - Shanghaitech University, Wiley Online Library on [18/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

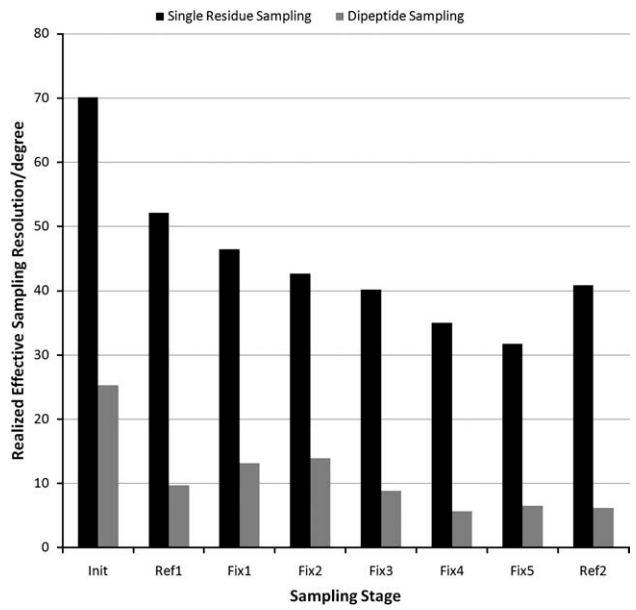

Figure 5

Comparison of the average number of generated loop conformations of seventeen 13-residue loops for two sampling methods, in each sampling stage. The blue columns are results of single residue sampling method, whereas the red columns are results of dipeptide sampling method. For each case, the number of generated loop conformations in each sampling stage is calculated by simply summing over the generated loop conformations in all single loop prediction jobs in this stage.

using dipeptide sampling method, the average number of generated loops is only about 8–17% of that when using single residue sampling method. Again, if the comparison in Figure 4 were made on longer loops, the improvement should be even bigger, since the average number of generated loops increases much faster with loop length when using single residue sampling method. With a much smaller number of generated loops, computational cost can be dramatically reduced by using dipeptide sampling method.

With significantly improved realized effective sampling resolution, better prediction results have been achieved. In each sampling stage, the sampling algorithm generates many loop candidates and ranks them according to the energy function. Several lowest energy models are passed to the next stage, in which the lowest energy model is called the predicted conformation of this stage. Figure 6 illustrate the average RMSD (to native) of the best-predicted conformations in each sampling stage, for loops with 14, 15, 16, and 17 residues, respectively. For RMSDs of the predicted conformations in each sampling stage, there is a clear decreasing tendency, which tells that our hierarchical scheme still works very well in the super long loop regime. However, occasionally, the first refinement stage even gets worse results than the previous stage (eg. first refinement stage in 17-residue loop set). The first refinement stage puts a 4 Å Cartesian constraint on each Cα atom in the target loop, aiming at to sample more finely in the promising region identified in the initial stage. However, for super long loops, lowest energy

Table IV

Results for a Test Set of 13-Residue Loops, Obtained by Using the Single Residue Sampling and Dipeptide Sampling Methods, Respectively[a]

				Single residue sample			Dipeptide sample		
PDB	Starts	Ends	Sequence	RMSD	dE	CPU time	RMSD	dE	CPU time
1AVW	A:91	A:103	HPNFNGNTLDNDI	5.45	1	73	2.10	−11	59
1JA1	A:515	A:527	KSQFRLPFKSTTP	0.80	−5	196	0.77	−6	177
1OWL	A:394	A:406	ASSGMDPKPLRIF	3.00	−9	158	2.43	21	179
1Q8F	A:151	A:163	GGAYGTGNFTPSA	1.66	−20	90	1.71	−17	74
2C0H	A:40	A:52	QAWVNYARDFGHN	0.40	15	153	3.80	28	70
2ECE	A:180	A:192	WEIDRGDQYLAYD	2.40	−12	191	2.06	−33	55
2FB5	A:99	A:111	ERNETLEALIQTG	1.98	1	84	1.93	11	54
2IHT	A:320	A:332	PTVNPIPRVYRPD	3.56	−11	204	3.60	−10	115
2PLX	A:91	A:103	HPSYNSNTLNNDI	3.76	−2	83	0.60	−4	72
2VUN	A:320	A:332	LNTGVIAPGKEAD	0.16	−3	125	0.18	−3	119
3D3Z	A:36	A:48	WDYDPSDGPSDSW	2.62	−25	78	2.89	−21	71
3DO8	A:76	A:88	TNPYGKTLDVDFE	1.83	−11	89	1.00	−10	81
3DRF	A:550	A:562	KRVVGMTLDYGAM	0.63	10	153	0.23	1	128
3DTB	A:560	A:572	KEDALNLKGLGDV	3.43	−17	170	2.68	−20	140
3E9T	A:482	A:494	DGTASAGTDFVGR	0.34	−3	52	0.18	−2	36
3EA1	A:137	A:149	FVDPIFLKTEGNI	0.32	−4	146	0.31	1	132
3G2E	A:56	A:68	DKEILFPYAVEGE	2.04	−24	102	1.32	−17	62
Mean				2.02	−7	126	1.63	−5	95
Stardard Deviation				1.50	11	48	1.19	15	44

RMSD is "global" backbone RMSD, in Å. dE is the energy difference between predicted and native loop, in kcal/mol. The CPU time refers to the cumulative time counted on a single processor, in hours.

[a]All calculations were conducted on a much faster Linux cluster than the one used in super long loop predictions (e.g., data in Table VII). Each node on this cluster has an Intel Xeon E5405 CPU @ 2.0G Hz.

10970134, 2011, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/prot.23129 by Cas - Shanghaitech University, Wiley Online Library on [18/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

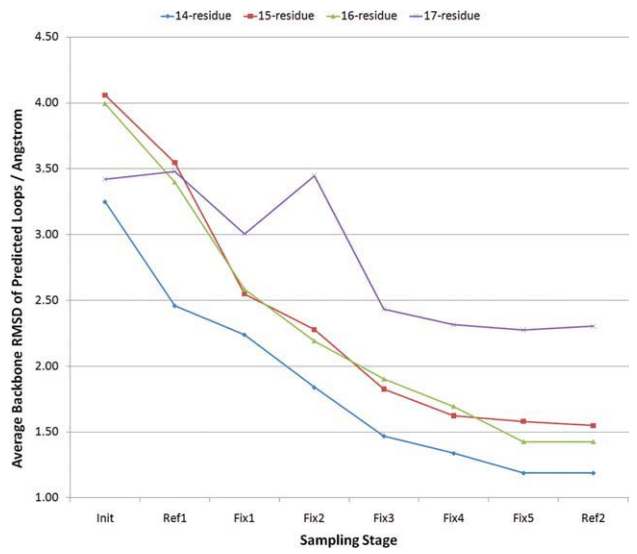

Figure 6

Average RMSD of the predicted results in each sampling stage for loops with 14, 15, 16, and 17 residues, using dipeptide sampling method.

conformations generated in the initial stages have rather large RMSDs (as shown in Fig. 6). A lowest energy conformation with a large RMSD most probably stays at or close to the bottom of a basin that corresponds to a local energy minimum, which is far from the deepest energy basin where the native stays. The extraordinarily rugged energy surface of super long loops, the limited ability of the energy function, combined with the constraint on Cα atoms, result in less accuracy of the predicted structures in the first refinement stage in some loops.

About 10% of the 90 cases in our test set have either energy error or sampling error, as shown Table III. After carefully examining the outliers, we found that π–π stacking interactions are missing in several predicted loops while exist in their native conformations. π–π stacking is generally thought of as being of significance when two aromatic rings interact in a planar geometry, but in fact, we see a wider variety of analogous interactions when a careful analysis of protein structures is performed, including for example interactions between a phenyl ring and the plane of the guanidium group in arginine. As a concrete illustration, consider the 14 residue loop in the protein 2BWR (B:158-B:171, as shown in Fig. 7), the π–π interaction between Tyr162 and Arg167 exists in the native loop (in cyan) are missing in the predicted structure (in magenta). It should be noted that the traditional understanding of the enhanced energetics attributable to "π–π" interactions, which is related to high level quantum mechanical electron correlation effects, may not be the most important factor in explaining our observations and results. Empirically, what we have found is that flat functional groups in proteins which contain double bonds (of which benzene rings and the guanidinium side chain are examples) have a high propensity to pack together in a stacking geometry, and that our older energy model (OPLS-AA plus the SGB continuum solvation model) was unable to reproduce the frequency with which such structures occur. By adding a term to the energy which rewards the stacking of these chemical groups, we may in fact be correcting problems with the continuum solvation model (which may not properly capture the high efficiency with which stacked geometries exclude water from regions which cannot make hydrogen bonds effectively to water molecules).

Recently, a new energy function has been developed in our lab,[29] in which a number of important changes, including an explicit π–π stacking interaction, have been incorporated. We have carried out a new set of calculations on the problematic loops in our test set, using identical sampling protocol, but replacing our old energy model with the new model. The details of this scoring function will be described in detail in Reference 29; here, we examine its performance for our long loop data set. Importantly, no parameters have been adjusted to improve this performance (the parametrization was carried out using a completely independent set of individual side chain rotamer data) so the results shown here can be considered a fair test as to whether the predictive power of the new model in high resolution structural refinement is superior to the model we employed previously.

Many of the outliers are substantially improved in RMSD when the new energy model is used, as shown in Table V. The new energy model eliminates most energy errors (only three cases, 1JP4, 3BB7, and 3CSS, now exhibits RMSD greater than 2 Å when the energy gaps are negative), and even improves some of the sampling errors (presumably by stabilizing intermediate structures that appear in the various stages of the loop prediction process). This dramatic reduction in energy errors with-

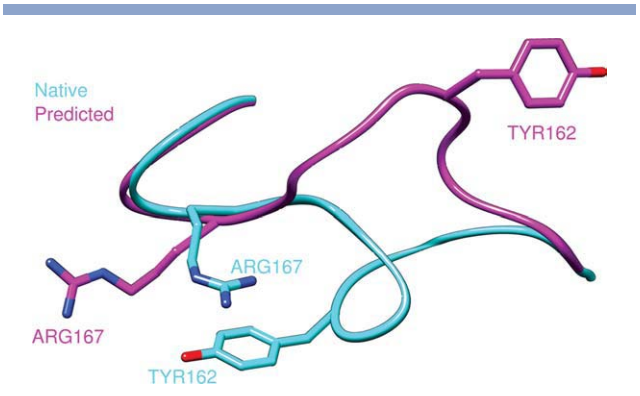

Figure 7

The native (in cyan) and predicted loop (in magenta) in 2BWR, B chain, residues 158–171. The π–π interaction between Tyr162 and Arg167 is missing in the predicted structure.

10970134, 2011, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/prot.23129 by Cas - Shanghaitech University, Wiley Online Library on [18/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table V
Summary of the Results for the Outliers (RMSD >2 Å) in Table IV (New Energy Model is Used)

				Old energy model		New energy model	
PDB	Starts	Ends	Length	RMSD (Å)	dE (kcal/mol)	RMSD (Å)	dE (kcal/mol)
1JP4	A:153	A:166	14	7.26	−43	4.89	−20
2BWR	B:158	B:171	14	4.56	35	0.33	−5
2C0H	A:40	A:53	14	5.96	108	3.54	26
1RA0	A:283	A:297	15	2.78	−39	1.14	−2
1ZHX	A:392	A:406	15	7.10	67	4.03	18
2AEB	B:156	B:170	15	2.55	25	0.87	4
2PKF	A:26	A:40	15	2.34	−8	0.45	6
3A64	A:350	A:364	15	2.55	3	0.21	−4
3BB7	A:231	A:245	15	6.26	−19	3.89	−5
3BF7	A:49	A:63	15	5.66	72	0.24	2
3CSS	A:95	A:109	15	2.36	−24	2.57	−12
1DJ0	B:19	B:34	16	7.08	72	6.20	51
2BG1	A:708	A:723	16	2.15	−4	0.64	13
2HDW	A:131	A:147	17	2.22	−126	2.43	25
2PYW	A:321	A:336	16	2.57	−37	1.18	8
3H2G	A:124	A:140	17	4.07	56	3.16	58
3HUH	A:71	A:87	17	8.56	−39	6.52	4
Mean				4.47	6	2.49	10
Median				4.07	−4	2.43	4
Standard Deviation				2.22	58	2.10	21

RMSD is "global" backbone RMSD, dE is the energy difference between predicted and native loop.

out any explicit parameter adjustment is a clear demonstration of the superiority of the new energy model. Further work on the sampling algorithm is clearly indicated, as a number of significant sampling errors remain, and we plan to pursue this in future development efforts.

Different the sampling from many fragment assembly methods (e.g., the 3-residue and 9-residue fragments in Rosetta), our method is unique based on these aspects below. First of all, our dipeptide rotamer library was built on a large variety of high resolution protein crystallographic structures, which not only ensures the reliability of the library but also allows the application of our method for modeling of all different kinds of proteins, whereas some fragment assembly methods build the fragments based on small proteins. Second, our dipeptide rotamers only contain explicit information of backbone conformations while fragments in other methods such as Rosetta contain side chain or reduced side chain information.[27] Furthermore, while fragments in many fragment assembly methods are applied as a sliding window along the protein chain, we don't use the dipeptide rotamers in an overlapping fashion in a sampling at one stage. Overall, our method of dipeptide sampling was built on a large reliable data set and could be applied for highly efficient and accurate loop modeling.

Crystal packing

The impact of crystal environment to polar groups on the surfaces of proteins has been extensively discussed in previous works.[17,28] In this work, crystal packing was explicitly included in our calculation in order to provide a fair comparison to the crystal structures. Although the additional information of crystal environment could reduce the sampling space to some extent, the calculation of loop prediction should be carried out under the same condition as the crystallographic experiment, for the purpose to test our methodology. In order to further understand the effect of crystal packing in our loop predictions, five 14-residue loop cases with little or some crystal contacts were tested with the same method we describe above except that crystal neighbors is removed. As shown in Table VI, the results obtained without crystal packing applied are highly consistent with those with crystal packing: the difference in predicted RMSDs ranges from 0.07 to 0.88 Å. This is a powerful evidence to show the effectiveness of our methodology regardless of crystal packing.

Computational costs

When using dipeptide sampling method, the average number of generated loops in each sampling stage is less than 17% of that when using single residue sampling method (Shown in Fig. 5). With many fewer generated loop candidates, the computational cost of the new sampling method is less expensive. Table VII compares the various statistics of computational time for loop sets with 11–13 residues in our previous work,[12] and for loop sets with 14–17 residues in this work. All

Table VI
Comparison of Five 14-Residue Loop Cases With and Without Crystal Packing Information

				With crystal packing		Without crystal packing	
PDB	Starts	Ends	No. of loop residues with crystal contacts[a]	RMSD (Å)	dE (kcal/mol)	RMSD (Å)	dE (kcal/mol)
1R6X	A:72	A:85	2	0.30	1	0.47	−117
1VYR	A:235	A:248	1	1.17	6	0.96	−18
3B40	A:389	A:402	0	1.28	48	0.40	19
3BY9	A:205	A:218	5	0.28	−11	0.35	−104
3EHR	A:95	A:108	5	0.94	7	0.56	−1

[a]The cutoff is 4 Å.

10970134, 2011, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/prot.23129 by Cas - Shanghaitech University, Wiley Online Library on [18/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

calculations were conducted on the same cluster of Intel or AMD processors in the range of 1.4 GHz or 900 MHz. The computational jobs were randomly distributed into the processors of the cluster. Despite this, there is some difference in the performance of the various processors, but this does not affect the qualitative analysis of our methodology. The average CPU times for 14-, 15-, 16-, and 17-residue loops are about 9, 13, 12, and 17 days, respectively. Unsurprisingly, the dipeptide sampling method is significantly faster than the single residue sampling method. For example, the statistical computational times of 17-residue loops, when using the dipeptide sampling method, are less than that of 12-residue loops with the single residue sampling method. Previously, when effective sampling resolution decreases, sometimes PLOP gets stuck by millions of generated conformations. The maximum time cost of the 13-residue data set (shown in Table VII) shows how long it will take when such scenario happens: nearly four months was used, about four times of the average. The new sampling method eliminates this problem. The most time-consuming case took one and a half months, only 27% more than the average.

The jobs using single residue sampling algorithm in Table V were run under a time-saving mode. When turning the time-saving mode on, RMSDs of the generated loop conformations in the buildup stages are not calculated. When large numbers of conformations are generated, the RMSD calculation can be very time-consuming. For example, turning the time-saving mode on usually can save about 20% of total computational time for 13-residue long loops, when single residue sampling is used. In this work, the jobs were run with the time-saving mode off. If we turn it on, another 5–10% time could be saved, considering the averages of generated loop candidates in the buildup stages are close to half of that of the single residue sampling method.

Most of the calculations are highly parallel [5 parallel jobs in the initial stage, 15–30 jobs in the first refinement stage, (5–10 times $(n + 1)$)] jobs in the nth fixed stage, five jobs in the second refinement stage). In all, there are about 150–200 jobs in a full loop prediction. We have a cluster with hundreds of processors, in average about 10 or more jobs can be run at the same time. As a result, most of our jobs can be finished in 1–3 days. Another 5–10% time can be saved if turning the time-saving mode on. In the future, we will continue to seek a variety of methods to speed up our algorithm and make it more efficient.

Table VII
Comparison of the Computational Costs of Two Sampling Algorithms: Single Residue Sampling and Dipeptide Sampling

Reference	Loop length (residue)	CPU Time (day)[a]			
		Mean	Median	Min	Max
Zhu et al.	11	12.3	11.7	5.0	25.4
	13	19.1	15.6	8.4	43.9
	13	31.4	22.5	12.3	114.6
This work	14	9.0	8.7	4.8	15.7
	15	12.9	11.0	1.9	34.2
	16	11.6	9.9	4.2	24.0
	17	17.0	16.5	10.1	23.3

The CPU time refers to the cumulative time counted on a single processor. Since most of the calculations are highly parallel, most of cases can be finished in 1–3 days in actual time.
[a]All calculations were conducted on the same cluster of Intel or AMD processors in the range of 1.4 GHz or 900 MHZ.

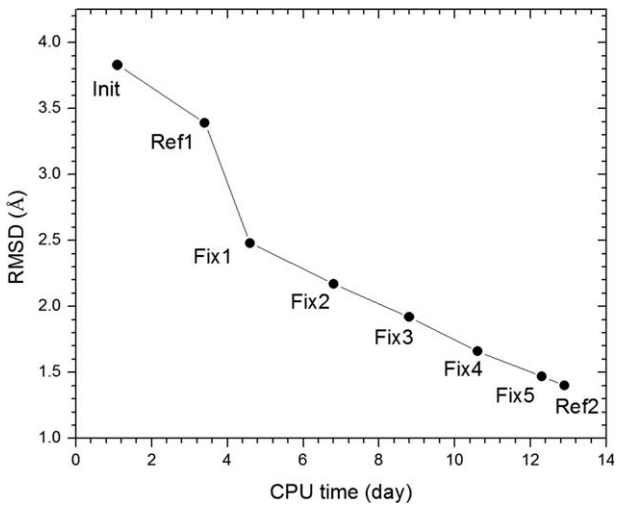

Figure 8
Average CPU times and RMSDs for 15 residue loops as a function of the sampling stages.

Finally, one can ask the question as to how the RMSD in the calculations is reduced as a function of the computation time that is utilized. The methods in PLOP are intended as high resolution methods (many much faster algorithms exist, and should be used if feasible, for example in cases with very high sequence identity in an available homologue and in the target loop in that homology), and we therefore have not explored performance if very low levels of CPU time are employed. Figure 8 below presents average CPU times and RMSDs for 15 residue loops as a function of the stage of the algorithm that has been reached (described in the text). It can be seen that there is a straightforward correlation between the expended CPU time and the quality of the result. Achieving high accuracy results requires relatively large amounts of CPU time as noted above. If low resolution predictions are desired, other methods are likely more cost effective.

CONCLUSION

We have developed an improved sampling algorithm, which is a modification of our algorithm described in Reference 12. By using dipeptide sampling instead of

10970134, 2011, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/prot.23129 by Cas - Shanghaitech University, Wiley Online Library on [18/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

single residue sampling, very good results for loops up to 17 residues have been achieved. When single residue sampling method is used, there are a lot of redundancies in the generated loop candidates, especially for long loops. A number of combinations of backbone dihedrals can lead to very similar loop conformations, due to the nonlinear relation between dihedrals and the backbone coordinates. The dipeptide libraries contain more information of two adjacent residues, thus the redundancies generated by the single residue sampling method due to ignoring the correlation between the two adjacent residues are removed. As a result, using dipeptide sampling method can generate better results in a much more efficient way.

There is no impediment to using polypeptide (such as tripeptide, tetrapeptide, and pentapeptide etc.) libraries in addition to the buildup algorithm we are using now. With libraries of longer polypeptide segments implemented, our algorithm will have a capability to deal with more challenging sampling problems that appear in homology modeling.

Program availability

The PLOP program can be downloaded at no cost by academic users from the Web site of Prof. Matthew Jacobson, our long time collaborator in this project. It can also be purchased from Schrodinger, Inc. as a component of the Prime program, which contains additionally a graphical user interface and other features added by Schrodinger. Incorporation of the latest algorithms into the released version takes some time but there is a systematic process in place for doing this and we expect that at some point, a version will be available for general use containing the capabilities described in this manuscript.

ACKNOWLEDGMENTS

This work was supported in part by a grant to RAF from the NIH (GM-52018). RAF is a stockholder in Schrodinger, Inc., which distributes a commercialized version of the PLOP program, Prime, and is on the Board of Directors and Scientific Advisory Board of Schrodinger, Inc.

REFERENCES

1. Baker NA, Sept D, Joseph S, Holst MJ, McCammon JA. Electrostatics of nanosystems: application to microtubules and the ribosome. Proc Natl Acad Sci USA 2001;98:10037–10041.
2. de Bakker PIW, DePristo MA, Burke DF, Blundell TL. *Ab initio* construction of polypeptide fragments: accuracy of loop decoy discrimination by an all-atom statistical potential and the AMBER force field with the Generalized Born solvation model. Proteins 2003;51:21–40.
3. DePristo MA, de Bakker PIW, Lovell SC, Blundell TL. *Ab initio* construction of polypeptide fragments: efficient generation of accurate, representative ensembles. Protein Struct Funct Bioinformat 2003;51:41–55.
4. Fiser A, Do RK, Sali A. Modeling of loops in protein structures. Protein Sci 2000;9:1753–1773.
5. Hildebrand PW, Goede A, Bauer RA, Gruening B, Ismer J, Michalsky E, Preissner R. SuperLooper—a prediction server for the modeling of loops in globular and membrane proteins. Nucl Acids Res 2009;37:W571–W574.
6. Jacobson MP, Pincus DL, Rapp CS, Day TJ, Honig B, Shaw DE, Friesner RA. A hierarchical approach to all-atom protein loop prediction. Proteins 2004;55:351–367.
7. Michalsky E, Goede A, Preissner R. Loops in proteins (LIP)—a comprehensive loop database for homology modeling. Protein Eng 2003;16:979–985.
8. Rohl CA, Strauss CE, Chivian D, Baker D. Modeling structurally variable regions in homologous proteins with rosetta. Proteins 2004;55:656–677.
9. Soto CS, Fasnacht M, Zhu J, Forrest L, Honig B. Loop modeling: sampling, filtering, and scoring. Proteins 2007;70:834–843.
10. Spassov VZ, Flook PK, Yan L. LOOPER: a molecular mechanics-based algorithm for protein loop prediction. Protein Eng Design Selection 2008;21:91–100.
11. Xiang ZX, Soto CS, Honig B. Evaluating conformational free energies: the colony energy and its application to the problem of loop prediction. Proc Natl Acad Sci USA 2002;99:7432–7437.
12. Zhu K, Pincus D, Zhao S, Friesner RA. Long loop prediction using the protein local optimization program. Proteins 2006;65:438–452.
13. Sellers BD, Zhu K, Zhao S, Friesner RA, Jacobson MP. Toward better refinement of comparative models: predicting loops in inexact environments. Proteins 2008;72:959–971.
14. Peng HP, Yang AS. Modeling protein loops with knowledge-based prediction of sequence-structure alignment. Bioinformatics 2007;23:2836–2842.
15. Fernandez-Fuentes N, Zhai J, Fiser A. ArchPRED: a template based loop structure prediction server. Nucl Acids Res 2006;34:W173–W176.
16. Felts AK, Gallicchio E, Chekmarev D, Paris KA, Friesner RA, Levy RM. Prediction of protein loop conformations using the AGBNP implicit solvent model and torsion angle sampling. J Chem Theory Comput 2008;4:855–868.
17. Jacobson MP, Friesner RA, Xiang Z, Honig B. On the role of the crystal environment in determining protein side-chain conformations. J Mol Biol 2002;320:597–608.
18. Jorgensen WL, Maxwell DS, Tirado-Rives J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. J Am Chem Soc 1996;118:11225–11236.
19. Kaminski GA, Friesner RA, Tirado-Rives J, Jorgensen WL. Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. J Phys Chem B 2001;105:6474–6487.
20. Jacobson MP, Kaminski GA, Friesner RA, Rapp CS. Force field validation using protein side chain prediction. J Phys Chem B 2002;105:11673–11680.
21. Ghosh A, Rapp CS, Friesner RA. Generalized Born model based on a surface integral formulation. J Phys Chem B 1998;102:10983–10990.
22. Gallicchio E, Zhang LY, Levy RM. The SGB/NP hydration free energy model based on the surface generalized born solvent reaction field and novel nonpolar hydration free energy estimators. J Comput Chem 2002;23:517–529.
23. Eldridge MD, Murray CW, Auton TR, Paolini GV, Mee RP. Empirical scoring functions. I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. J Comput Aid Mol Des 1997;11:425–445.

10970134, 2011, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/prot.23129 by Cas - Shanghaitech University, Wiley Online Library on [18/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

24. Hartigan JA. Clustering alogrithms. New York: Wiley; 1975.
25. Hartigan JA, Wong MA. A K-means clustering algorithm. Appl Stat 1979;136:100–108.
26. Li X, Jacobson MP, Zhu K, Zhao S, Friesner RA. Assignment of polar states for protein amino acid residues using an interaction cluster decomposition algorithm and its application to high resolution protein structure modeling. Proteins 2007;66: 824–837.
27. Bujnicki JM. Prediction of protein structures, functions, and interactions. New York: John Wiley & Sons, Ltd.; 2009.
28. Xiang Z, Steinbach PJ, Jacobson MP, Friesner RA, Honig B. Prediction of side-chain conformations on protein surfaces. Proteins 2007;66:814–823.
29. Li J, Abel R, Zhu K, Cao Y, Zhao S, Friesner RA. A next generation energy model for high resolution protein structure modeling. DOI: 10.1002/prot.23106.
30. Zhu K, Shirts MR, Friesner RA. Multiscale optimization of a truncated Newton minimizer and application to proteins and protein-ligand complex. J Chem Theory Comp 2007;3:2108–2119.
31. Xiang Z, Honig B. Extending the accuracy limits of prediction for side-chain conformations. J Mol Biol 2001;311:421–430.

10970134, 2011, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/prot.23129 by Cas - Shanghaitech University, Wiley Online Library on [18/11/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License