Published in final edited form as:
J Comput Chem. 2014 February 5; 35(4): 335–341. doi:10.1002/jcc.23509.

LEAP: Highly accurate prediction of protein loop conformations by integrating coarse-grained sampling and optimized energy scores with all-atom refinement of backbone and side chains

Shide Liang[a],*, **Chi Zhang**[b], and **Yaoqi Zhou**[c],[d],*

[a]Systems Immunology Lab, Immunology Frontier Research Center, Osaka University, Suita, Osaka, 565-0871, Japan

[b]School of Biological Sciences, Center for Plant Science and Innovation, University of Nebraska, Lincoln, NE, 68588, USA

[c]School of Informatics, Indiana University Purdue University at Indianapolis, Indianapolis, IN 46202, Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, IN 46202, USA

[d]Institute for Glycomics and School of Informatics and Communication Technology, Griffith University, Parklands Drive, Southport Qld 4222, Australia

Abstract

Prediction of protein loop conformations without any prior knowledge (*ab initio* prediction) is an unsolved problem. Its solution will significantly impact protein homology and template-based modeling as well as *ab initio* protein-structure prediction. Here, we developed a coarse-grained, optimized scoring function for initial sampling and ranking of loop decoys. The resulting decoys are then further optimized in backbone and side-chain conformations and ranked by all-atom energy scoring functions. The final integrated technique called LEAP (Loop prediction by Energy-Assisted Protocol) achieved a median value of 2.1Å RMSD for 325 12-residue test loops and 2.0 Å RMSD for 45 12-residue loops of CASP 10 target proteins with native core structures (backbone and side chains). If all side-chain conformations in protein cores were predicted in the absence of the target loop, loop prediction accuracy only reduces slightly (0.2Å difference in RMSD for 12-residue loops in the CASP target proteins). The accuracy obtained is about 1Å RMSD or more improvement over other methods we tested. The executable file for a Linux system is freely available for academic users at http://sparks-lab.org.

Keywords

Loop modeling; coarse-grained energy function; energy minimization; Monte Carlo simulation; force field development

*To whom correspondence should be addressed. **Contact:**shideliang@ifrec.osaka-u.ac.jp and yaoqi.zhou@griffith.edu.au.

INTRODUCTION

The current best tools for protein structure prediction employ a template-based approach.[1] In this approach, a query sequence is aligned onto a structural template often with gaps in loop regions because unlike regions with secondary structures, loop structures are often not conserved. Thus, loop prediction (or modeling) is an essential component of protein structure modeling - an increasingly important task as the gap further expands between the number of proteins with experimentally determined structures (tens of thousands) and the number of proteins with known sequences (hundreds of millions and counting). Accurate modeling of loop structures is important because loops often play functionally important roles ranging from active sites of enzymes, binding sites of ions, to molecular recognition sites.[2,3] Moreover, permutation of loops is one way to generate new structural folds of proteins.[4-6]

Algorithms for loop prediction have been comprehensively reviewed.[7-9]They can be generally classified[9] into database-based loop selections,[10-13] energy-based sampling and selections[14,15] and their combinations.[16-18] In a database-based approach, a loop prediction is made by locating the best fit from a loop structural library. This would require a nearly complete library for loop structures that is true only for short loops.[19,20]

In this article, we will focus on energy-based methods that at minimum require a conformational sampling technique and an energy function. Recent work has significantly improved the accuracy in loop prediction.[21-27] In particular, the PLOP program[28] can achieve a median global backbone RMSD of <1Å for 104 loops of 11-13 residues[29] and <2Å for 89 loops of 14-17 residues with a molecular-mechanics OPLS all-atom force field.[26] These results were achieved in a crystal environment. The POS program[30] also achieved a subangstrom accuracy for 72.2% loops of 10-12 residues by employing molecular mechanics force field, knowledge-based, and empirical multiple scoring functions.[27] In a recent study, a new force field called VSGB 2.0 was developed for accurate loop prediction based on an optimized OPLS/SGB-NP force field[31] in which the coefficients of various physics-based terms were optimized by achieving high accuracy in single side-chain prediction as our early work.[32] However, it is not yet clear if these methods would achieve a similar level of accuracy in the absence of a crystal environment or for the loops in recently released protein structures that were never used for parameter optimization.

Most existing loop prediction techniques utilized either all-atom molecular mechanics force fields[28,33] or knowledge-based energy functions derived from protein structures[34,35] or their combinations.[27] Recently, we have developed an orientation-dependent energy function called OSCAR that is based on series expansion and the parameters were optimized by single side chain prediction using a large data set.[36] This energy function improves over several side-chain prediction techniques that are based on weight-optimized physical and/or knowledge-based energy functions. Direct application of OSCAR to loop decoys reveals that it is capable of selecting highly accurate near-native structures.[37] This success leads to development of a backbone-based OSCAR potential. The combination of this backbone potential with the side-chain based OSCAR potential yields a reasonably accurate method

for loop sampling and prediction (average RMSD of 2.08 Å for 528 ten-residue loops). However, its accuracy decreases quickly as the length of loop increases (from an average RMSD of 2.73Å for 392 eleven-residue loops to 3.58Å for 325 twelve-residue loops).[18]

In this paper, we develop a new loop prediction technique by combining coarse-grained sampling with refinement of both backbone and side-chains. We first establish an optimized reduced potential for initial coarse-grained sampling. This is followed by modeling side-chains and refining backbone with an all-atom OSCAR potential. The resulting top ranked loops are further refined with a mixed CHARMM bond energy[38] and the OSCAR potential. The new method called LEAP (Loop prediction by Energy Assisted Protocol) improves over the OSCAR-loop method for loops of all lengths tested (4-12). In particular, the average RMSD for 12 residue loops decreases by 1.04Å to 2.54Å. LEAP also makes consistent improvement over FREAD[39], Loop Builder[9] and PLOP[28] methods with default parameters for loops of lengths from 4 to 12 (4-17 for FREAD) with more than 1Å for the average or median RMSD for the longest 12-residue loops studied.

METHODS

The summary of loop prediction protocol

As shown in Figure 1, the loop prediction protocol comprises of three steps. First, the backbone conformations of a given target loop are generated by the cyclic coordinate descent (CCD) algorithm.[40] These conformations are selected and optimized by a to-be-described reduced energy function. Second, side chains for these selected backbone structures were built by the OSCAR-based side-chain prediction tool with a rigid rotamer model. The resulting all-atom loop models were optimized and selected by the same all-atom OSCAR plus a backbone potential. Third, the top selected models are further minimized and selected by a combined energy of the OSCAR potential for flexible side-chain rotamers and CHARMM bond energies. There are four optimized energies involved in the above protocol: the original OSCAR potential for side-chain prediction, labeled as E_{SCP}, the OSCAR potential optimized for loop prediction, E_{LP}, by adding an additional backbone term to E_{SCP}, the side-chain reduced potential, E_R^{rigid}, and the mixing potential E_{CLP}^{flex}. We will introduce them separately below. Especially, there are two versions of E_{SCP}: one optimized for flexible rotamers E_{SCP}^{flex} [36] and one optimized for rigid rotamers E_{SCP}^{rigid}.[41] The soft sphere potential E_{SCP}^{rigid} is used at the initial stages to account for errors due to discrete approximations and more accurate E_{CLP}^{flex} that incorporates E_{SCP}^{flex} is used for energy minimization and selection at the final stage.

The OSCAR potential for side-chain prediction (E_{SCP})

The orientation-dependent OSCAR potential for side-chain prediction was described by the following equation.

$$E_{SCP} = \sum E(r_{ij})\, E(\theta_{ij}, \varphi_{ij}, \psi_{ij}) + \sum E^{Rotamer}(\chi),$$

where the distance-dependent

$$E(r_{ij})=\frac{a_1}{r_{ij}^2}+\frac{a_2}{r_{ij}^4}+\frac{a_3}{r_{ij}^6}++\frac{a_4}{r_{ij}^8},$$

the angle dependent

$$E(\theta_{ij},\varphi_{ij},\psi_{ij})=b_1cos^2\theta_{ij}+b_2cos^2\varphi_{ij}+b_3cos^2\psi_{ij}+b_4cos\theta_{ij}cos\varphi_{ij}+ b_5cos\theta_{ij}cos\psi_{ij}+b_6cos\varphi_{ij}cos\psi_{ij}+b_7cos\theta_{ij}+b_8cos\varphi_{ij}+b_9cos\psi_{ij}+C,$$

and the side-chain rotamer-torsion dependent

$$E^{Rotamer}(\chi)=t_1cos\chi+t_2sin\chi+t_3cos2\chi+t_4sin2\chi+t_5cos3\chi+t_6sin3\chi.$$

The above parameters *a*, *b*, *C* and *t* were all optimized by maximizing the energy gap between the native rotamer conformation from other conformations. As mentioned above, there are two versions of E_{SCP}: one optimized for flexible rotamers E_{SCP}^{flex} and one for rigid rotamers E_{SCP}^{rigid}.

The OSCAR potential for loop prediction (E_{LP})

To improve the usefulness of the above side-chain prediction potential for loop prediction, an additional backbone torsion angle was introduced.[18]

$$E_{LP}=E_{SCP}+d_1cos\phi+d_2sin\phi+d_3cos2\phi+d_4sin2\phi+d_5cos3\phi+d_6sin3\phi+ f_1cos\psi+f_2sin\psi+f_3cos2\psi+f_4sin2\psi+f_5cos3\psi+f_6sin3\psi,$$

where φ and ψ are backbone torsion angles, *d* and *f* are parameters optimized so that near-native loop decoys have lower energies than those loop conformations far from native ones.[18] Similarly, there are two versions of E_{LP}: one optimized for flexible rotamers E_{LP}^{flex} and one for rigid rotamers E_{LP}^{rigid}.

The OSCAR reduced potential E_R^{rigid}

In this paper, we introduce a reduced side-chain OSCAR potential E_R^{rigid} to improve the initial sampling of loop conformations.

$$E_R^{rigid}= E_{LP}^{rigid}(backbone)+\sum_{i=1}^{4}\alpha_ie^{-r_i/3}+\sum_{i=1}^{4}\sum_{j=i+1}^{4}\beta_i\gamma_je^{-(r_i+r_j)/6}+\sum_{i=1}^{4}\delta_ie^{-r_i/6}+\varepsilon,$$

where E_{LP}^{rigid} (backbone) is the portion of the interaction energy between loop backbones including Cβ atoms and between loop backbone atoms and the rest of proteins from the OSCAR energy optimized for loop prediction with rigid rotamers, r_1 is the distance between the Cα atoms of two residues, r_2 is the distance between the geometric centroids of side

chain atoms of two residues based on all rotamer conformations,[42] r_3 and r_4 are the distances between the Cα atom of one residue and the side-chain centroid of the other residue, respectively, α, β, γ, δ, and ε are to-be-optimized parameters. The cutoff distance for r_1 is 15Å.

Parameter optimization for the OSCAR reduced potential

There are a total of 3,070 α, β, γ, δ, and ε parameters for 210 amino acid pairs (i.e., 210×15-80 reduced parameters for interaction between the same residue types). They were obtained by minimizing $\sum_{k=1}^{M}\sum_{i=1}^{N} RMSD(i,k)e^{-E_R^{rigid}(i,k)}/M\sum_{i=1}^{N} e^{-E_R^{rigid}(i,k)}$ where M is the total number of training loops, N is the number of decoys per training loop, $E_R^{rigid}(i,k)$ is the reduced energy for the decoy i of the k^{th} training loop and RMSD(i,k) is the backbone RMSD to the native conformation.

The parameters were optimized based on 13378 8-residue target loops (i.e. M=13,378) collected using our previous method.[18] First, 100,000 decoys per loop were generated by CCD algorithm.[40] Then, we picked 20 decoys with the lowest RMSD from the native loop conformation. The next 120 decoys were selected sequentially according to $E_R^{rigid}(i,k)$ in the remaining decoys whose RMSD is >1Å from all previously selected decoys. The next 60 decoys were selected sequentially according to $E_R^{rigid}(i,k)$ in the remaining decoys whose RMSD is >2Å, 3Å and 4Å, respectively, from all previously selected decoys. A total of 200 decoys were selected per loop (i.e. N=200). In some cases N can be less than 200 if not enough decoys satisfy above conditions from 100,000 generated decoys. All of the parameters were initialized with a random value and then optimized by Monte Carlo simulations with the objective function shown above. A total of 40 cycles of simulated annealing were repeated. Each cycle makes either successful 30,700 parameter changes or a total of number of 307,000 changes whichever comes first.

The mixing potential E_{CLP}^{flex}

The mixing potential for the final selection of loop conformations is obtained from linear combination of CHARMM bonded interactions and the OSCAR energy for loop prediction with flexible rotamers. That is, $E_{CLP}^{flex}=E_{LP}^{flex}+\eta E_{Bond}^{CHARMM}$ where η is a to-be-optimized mixing coefficient. Here, we used CHARMM 19 parameters of bond lengths, bond angles, and improper dihedral angles for energy calculation. A simple grid search at η = 2, 4, 6 and 8 was made for locating the single value for the final selection of loop decoys in the training loops. More specifically, 1,000 loops with a length of 8 residues were randomly selected from the above-mentioned 13,378 training loops. Top 10 decoys with built side chains were selected for each target with the loop prediction protocol described in the next section. E_{CLP}^{Flex} with a pre-defined mixing coefficient was used for minimization and selection. The final mixing coefficient is 4 for achieving the highest accuracy of 0.88 Å for 1,000 8-residue loops. The overall accuracy was only slightly lower for other coefficients (0.89-0.92 Å).

Implementation of loop prediction protocol

Here are the actual steps implemented for LEAP (Figure 1). First, a fixed number of backbone decoy conformations are generated by the CCD algorithm (10,000, 100,000 and 1,000,000 backbone conformations for loops with lengths of 4-6 residues, 7-9 residues, and 10 or more residues, respectively). Top 200 decoys are selected by the reduced side-chain OSCAR potential E_R^{rigid}. Additional 800 decoys are selected sequentially based on E_R^{rigid} and the RMSD >1Å, 2Å, 3Å, and 4Å from previously selected decoys. That is, a total of 1,000 decoys are selected at the most. The energies of selected decoys (E_R^{rigid}) are optimized by slightly changing backbone ϕ and ψ dihedral angles in the range of ± 0.5° with 2,000 steps of Monte Carlo (MC) simulated annealing. Second, side chains for these decoys were added and optimized by E_{SCP}^{rigid} and then the backbone conformation was further refined by E_{LP}^{rigid} for 2,000 MC steps with fixed loop side chains. Third, the top 10 decoys ranked by E_{LP}^{rigid} are minimized for 200 Powell steps by the all-atom mixing potential E_{CLP}^{flex} (or less than 200 Powell steps if the stepwise energy change is less than 0.0001). The final predicted loop is ranked based on minimized E_{CLP}^{flex} values.

Evaluation of Predicted Loops

We employed global RMSD for evaluation. The backbone heavy atoms (N, Cα, C, and O) were utilized to calculate the RMSD between the loop decoy with the lowest energy and the observed loop structure after aligning the protein framework.

Training and Test Loop Sets

Training and test loop sets are collected the same way as our previous work.[18] Briefly, a total of 3,315 protein chains were obtained with a sequence identity cutoff of 20%, resolution of <2Å, R factor<0.25, and more than 98% residues with complete coordinates. A randomly selected 200 proteins constitutes the test set and the remaining 3,115 proteins are the training sets. In this work, only 8-residue loops in 3115 proteins were utilized as the training sets for E_R^{rigid} and the mixing coefficient η. We chose 8-residue loops because their intermediate length allows efficient training.

In addition to 200 chains as an independent test set, we further employed target proteins from CASP 10 (Critical Assessment of Structure Prediction techniques, 2012). The structures of these proteins were released in 2012 and were not considered in developing LEAP or other loop modeling methods compared in this study. We downloaded the list of CASP 10 target proteins from http://www.predictioncenter.org/casp10/targetlist.cgi. Loops in 21 targets of monomeric proteins with available structures in PDB were identified according to the definition employed earlier.[18] More specifically, helical and sheet regions were excluded according to torsion angles, a loop region is selected if it has more than 50% residues exposed (>20%solvent accessibility), does not interact with a ligand (>4.5Å in distance) and does not contain a cis-peptide bond in the main chain. All loops satisfying the above criteria are included. Only the single protein chain was used in the prediction. PDB

IDs for the 21 proteins are 4f67, 4fmw, 4hqf, 2ymv, 2luz, 4ftd, 4gl6, 4hg2, 4gpv, 4epz, 4fgm, 4f54, 4fd0, 4fr9, 4fs7, 4g2a, 4gt6, 4h09, 4e6f, 4fdy, and 4h0a.

Other Methods

The PLOP (version 25.1) program was downloaded from https://plop.jacobsonlab.org/plop_releases/. Default parameters were utilized for loop modeling in the absence of crystal packing constraints.

Loop Builder[9] is the extension of a loop modeling program called LOOPY.[43] A statistical potential DFIRE[44] is used to select 50 loop structures predicted by LOOPY. The selected decoys are then minimized and ranked with an all-atom force field OPLS/SGB-NP implemented in the PLOP.[28] The program was downloaded from http://bhapp.c2b2.columbia.edu/software/Loopy. The same parameters described by Soto *et al*.[9]were used for long loops with a length of 8-12 residues. For short loops (4-7 residues), 1,000 initial conformations were generated and other parameters were the same as for long loops.

RESULTS

Loop modeling for 200 independent test proteins

Figure 2 displays the results from each step of loop prediction in the LEAP algorithm for loops at different lengths of 200 test proteins. This is a fairly large test set with the number of loops ranging from 325 loops for 12-residue loops to 2809 for 4-residue loops. For all loop lengths tested, there is a steady reduction of RMSD (either average or median value, only median is shown) from coarse-grained sampling, side-chain modeling and backbone refinement to further flexible all-atom minimization. The improvement at each step is more significant (~0.5Å) for longer loops.

Comparison to PLOP, Loop Builder and OSCAR-loop

Figure 3 compares LEAP with PLOP, Loop Builder and OSCAR-loop for the same test dataset. All RMSD values reported here are based on N, Cα C, and O atoms. Default RMSD values based on N, Cα and C atoms from PLOP were converted. Not all loops were predicted by PLOP and Loop Builder. The results of LEAP for those slightly reduced sets are indistinguishable from each other. Thus, only the LEAP results for the whole set is shown in this figure. The performance of LEAP is consistently better than PLOP, Loop Builder, or OSCAR-loop at each loop length. The longer the loop length is, the more significant the improvement is. For example, the median RMSD for 12-residue loops is 4.07Å by PLOP, 3.18Å by Loop Builder, 3.05Å by OSCAR-loop but only 2.06Å by LEAP.

Comparison to FREAD

The test set of the homology-based FREAD contains 30 targets for each loop length and has shown to be difficult for previously developed *ab initio* loop modeling methods.[39] Here, we found that LEAP significantly improves over FREAD (Table 1) except for loops beyond 18 residues where both are not accurate (>5.5Å RMSD). For the loops between 4 and 17 residues, LEAP typically makes more than 1Å improvement based on the average of 10

predictions. The accuracy can be further improved if the loop with the lowest energy in 10 separate predictions is considered.

Comparison using experimental structures of CASP 10 target proteins

To further confirm the performance of the LEAP, we apply it to the CASP 10 target proteins, which were released recently and served as an additional independent test set for all the methods employed here. Results along with those by PLOP and Loop Builder are shown in Figure 4. This dataset has 413, 276, 225, 146, 126, 83, 74, 51, and 45 loops for loop lengths of 4 to 12, respectively. The magnitude of the improvement is similar to those in Figure 3. The longer the chain length is, the larger the improvement is. It should be noted that some of the 21 CASP target proteins are homologous to one of the 3315 training and test proteins. The maximal sequence identity is more than 60% for 10 out of the 21 targets in local alignment between the two groups. Nevertheless, the median accuracy for 12-residue loops of these 10 targets (1.84Å) is only slightly better than that of the other 11 targets with lower maximal sequence identity (2.12Å). The independence of the median value on the maximal sequence identity supports the robustness of our training set. We also tested the use of our all-atom mixing scoring function E_{CLP}^{flex} for minimizing and re-ranking the top 50 loops predicted by Loop Builder (Open triangle in Figure 4). This brings the accuracy of predicted loops comparable to the average of 10 LEAP predictions. More importantly, loop prediction by LEAP based on the lowest energy in 10 independent predictions further significantly improves the accuracy for loop targets with various lengths (closed circles). There was no such improvement from multiple predictions by PLOP or Loop Builder.

DISCUSSION

In this paper, we have developed a new loop-prediction technique that integrates coarse-grained sampling and scoring with all-atom (backbone and side chain) refinement in a single automatic software package called LEAP. The method achieves the median value of 2Å RMSD for 12-residue loops that is 1Å or more than other methods tested using default parameters.

The improvement in performance of our methods over previous techniques is mainly due to the accuracy of the optimized E_{CLP}^{flex}. This is illustrated by the results shown in Figure 4. When the decoys generated by Loop Builder are minimized by our mixing all-atom scoring function, the resulting accuracy of predicted loops is similar to that of LEAP. Our all-atom energy function was combined from CHARMM bond energy and OSCAR all-atom optimized potential. Loop Builder employed the physical-based energy, OPLS/SGB-NP, a general-purpose molecular mechanics force field with an approximate generalized Born solvation model. It has been thought that physical-based energy functions are more suitable for all-atom models while knowledge-based (statistical or optimized) potentials are appropriate only for coarse-grained models. The usefulness of all-atom knowledge-based potentials, however, is demonstrated by more and more studies[45] ranging from protein structure refinement[46] to partial refolding[47]. This study offers another example that an energy function extracted from a large database of protein structures can better serve for a specific purpose than a physical based energy function.

In addition to the all-atom potential, the reduced energy also plays an important role in the overall accuracy of the LEAP program. Removing this energy function will decrease the median accuracy from 2.71 to 2.98 Å at the initial sampling stage for 325 12-residue test loops.

The performance of the LEAP program, however, is limited by insufficient, initial conformational sampling for long loops, in particular. Insufficient sampling is demonstrated by the significant improvement when the loop having the lowest energy in 10 predicted loops is employed for prediction (Table 1 and Figure 4). Figure 5 confirms insufficient sampling by examining the dependence of the loop-prediction accuracy on the number of initial conformations sampled. For 8-residue and 10-residue loops, the number of initial conformations employed in this study is adequate because more initial conformations lead to essentially the same accuracy. However, for 12 residue loops, sampling of 1,000,000 conformations does not yet lead to a converged result, indicating that improving sampling techniques is needed for further increasing the accuracy of LEAP. In our previous study,[37] the OSCAR force field is very effective for loop selections if there is a conformation less than 0.4Å RMSD from the native conformation. However, such a conformation is difficult to generate for long loops, even with 1,000,000 initial conformations. We examined the distribution of RMSDs in the initial conformations. We found that there are an average of 69/100,000 for 8 residue loops, 16/1,000,000 for 10-residue loops, and 0.2/1,000,000 for 12-residue loops with RMSD <1Å from the native loop conformations in 10 independent runs. Thus, improving the current method for loop sampling (CCD algorithm) and for global minimum search will likely lead to a more accurate loop-prediction method.

We would like to emphasize that the comparison between our method and other methods is not exact because it is difficult to set the same parameters at each stage of sampling and scoring for different methods. For example, the default option of PLOP employs only 2^N initial conformations (4096 for 12 residue loops). We attempted to increase the number of conformations sampled by PLOP. However, the program often fails for unknown reasons. We also tested 10 independent runs and found that combining 10 independent runs did not improve the accuracy of PLOP or Loop Builder. Furthermore, the released version of PLOP does not contain the hierarchical refinement strategy employed by Jacobson et al.[28] In the study, they also included crystal packing which makes direct comparison with our study impossible. Here we attempt to build a method that does not rely on crystal packing for prediction because in a real-world situation, crystal packing information is often not available.

One limitation of LEAP is its computational requirement. Dependent on loop length, it takes 1 to 10 hours to complete a loop prediction on a single Intel Xeon processor operating at 3.5 GHz. For example, the calculation time for a 12-residue loop is about 4-7 hours. By comparison, it is 1-3 hours for Loop Builder with optimized parameters[9] and 5-10 minutes for PLOP with default parameters. The accuracy of PLOP could be adversely affected by insufficient sampling as discussed above. Similarly, more efficient sampling and minimization techniques will be also useful for speeding up the calculation of LEAP.

The ultimate purpose of LEAP is to improve the accuracy of modelling of the gap regions in template-based structure prediction. In such a real-world situation of homolog modelling, core backbone structures and side-chain conformations are all approximate and missing loops are often more than one per structure. To make an initial assessment for usefulness of LEAP in homology modelling, we maintain the native backbone conformations of protein cores but all side chains are rebuilt by our side-chain prediction program based on the OSCAR orientation dependent energy function.[36] The core side-chains can be rebuilt with or without the presence of native loops. For 12-residue loops of CASP 10 target proteins, the median loop RMSD changes from 1.28Å with native side chains, 1.25Å with core side chains built in the presence of native loop backbone conformations, to 1.44Å with core side chains built in the absence of loops. Repacking side chains without removing native backbone conformations of target loops does not change the accuracy of our prediction. By comparison, an increase of 0.4Å RMSD was observed for a similar study with Loop Builder by Soto et al.[9] The accuracy of LEAP decreases slightly from 1.28 to 1.44Å if core side chains are repacked in the absence of loops. The minor reduction in accuracy with approximate side chain conformations is very encouraging for applying LEAP in a more realistic situation of homology modeling.

Summary

In this paper, we have developed a new loop prediction algorithm called LEAP that combines coarse-grained sampling and scoring with backbone refinement and all-atom minimization. In the absence of a crystal environment, our method can achieve a median value of 2Å RMSD for 325 12-residue test loops. A similar value is obtained for 45 12-residues in CASP targets. This is about 1Å RMSD or more improvement over other methods we tested. Further test of the method for homology models is in progress.

Acknowledgments

This work was partly supported by a kakenhi grant 24570184: Grant-in-Aid for Scientific Research (C) from the Japan Society for the Promotion of Science (JSPS). Y. Z. was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number R01GM085003. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Mariani V, Kiefer F, Schmidt T, Haas J, Schwede T. Proteins. 2011; 79(Suppl 10):37. [PubMed: 22002823]
2. Fetrow JS. FASEB J. 1995; 9:708. [PubMed: 7601335]
3. Tendulkar AV, Joshi AA, Sohoni MA, Wangikar PP. J Mol Biol. 2004; 338:611. [PubMed: 15081817]
4. Dai L, Zhou Y. J Mol Biol. 2011; 408:585. [PubMed: 21376059]
5. Cunningham BA, Hemperly JJ, Hopp TP, Edelman GM. Proc Natl Acad Sci U S A. 1979; 76:3218. [PubMed: 16592676]
6. Lindqvist Y, Schneider G. Curr Opin Struct Biol. 1997; 7:422. [PubMed: 9204286]
7. Shehu A, Kavraki LE. Entropy. 2012; 14:252.
8. Fiser A, Do RKG, Sali A. Protein Sci. 2000; 9:1753. [PubMed: 11045621]
9. Soto CS, Fasnacht M, Zhu J, Forrest L, Honig B. Proteins. 2008; 70:834. [PubMed: 17729286]

10. Greer J. Proc Natl Acad Sci U S A. 1980; 77:3393. [PubMed: 6932026]
11. Chothia C, Lesk AM. J Mol Biol. 1987; 196:901. [PubMed: 3681981]
12. Claessens M, Vancutsem E, Lasters I, Wodak S. Protein Eng. 1989; 2:335. [PubMed: 2928296]
13. Lis M, Kim T, Sarmiento JJ, Kuroda D, Dinh HV, Kinjo AR, Amada K, Devadas S, Nakamura H, Standley DM. Immunome Res. 2011; 7:1.
14. Moult J, James MN. Proteins. 1986; 1:146. [PubMed: 3130622]
15. Fine RM, Wang H, Shenkin PS, Yarmush DL, Levinthal C. Proteins. 1986; 1:342. [PubMed: 3449860]
16. Chothia C, Lesk AM, Levitt M, Amit AG, Mariuzza RA, Phillips SE, Poljak RJ. Science. 1986; 233:755. [PubMed: 3090684]
17. van Vlijmen HW, Karplus M. J Mol Biol. 1997; 267:975. [PubMed: 9135125]
18. Liang SD, Zhang C, Sarmiento J, Standley DM. J Chem Theory Comput. 2012; 8:1820.
19. Du PC, Andrec M, Levy RM. Protein Eng. 2003; 16:407. [PubMed: 12874373]
20. Baeten L, Reumers J, Tur V, Stricher F, Lenaerts T, Serrano L, Rousseau F, Schymkowitz J. PLoS Comp Biol. 2008; 4:E1000083.
21. Subramani A, Floudas CA. J Phys Chem B. 2012; 116:6670. [PubMed: 22352982]
22. Danielson ML, Lill MA. Proteins. 2010; 78:1748. [PubMed: 20186974]
23. Miller EB, Murrett CS, Zhu K, Zhao SW, Goldfeld DA, Bylund JH, Friesner RA. J Chem Theory Comput. 2013; 9:1846. [PubMed: 23814507]
24. Fiser A, Sali A. Bioinformatics. 2003; 19:2500. [PubMed: 14668246]
25. Simons KT, Bonneau R, Ruczinski I, Baker D. Proteins. 1999:171. [PubMed: 10526365]
26. Zhao S, Zhu K, Li J, Friesner RA. Proteins. 2011; 79:2920. [PubMed: 21905115]
27. Li Y, Rata I, Chiu SW, Jakobsson E. BMC Struct Biol. 2010; 10:22. [PubMed: 20642859]
28. Jacobson MP, Pincus DL, Rapp CS, Day TJ, Honig B, Shaw DE, Friesner RA. Proteins. 2004; 55:351. [PubMed: 15048827]
29. Zhu K, Pincus DL, Zhao S, Friesner RA. Proteins. 2006; 65:438. [PubMed: 16927380]
30. Li YH, Rata I, Jakobsson E. J Chem Inform Mod. 2011; 51:1656.
31. Li J, Abel R, Zhu K, Cao Y, Zhao S, Friesner RA. Proteins. 2011; 79:2794. [PubMed: 21905107]
32. Liang S, Grishin NV. Protein Sci. 2002; 11:322. [PubMed: 11790842]
33. Spassov VZ, Flook PK, Yan L. Protein Eng Des Sel. 2008; 21:91. [PubMed: 18194981]
34. Zhang C, Liu S, Zhou Y. Protein Sci. 2004; 13:391. [PubMed: 14739324]
35. de Bakker PI, DePristo MA, Burke DF, Blundell TL. Proteins. 2003; 51:21. [PubMed: 12596261]
36. Liang S, Zhou Y, Grishin N, Standley DM. J Comput Chem. 2011; 32:1680. [PubMed: 21374632]
37. Liang S, Zhang C, Standley DM. Proteins. 2011; 79:2260. [PubMed: 21574188]
38. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. J Comput Chem. 1983; 4:187.
39. Choi Y, Deane CM. Proteins. 2010; 78:1431. [PubMed: 20034110]
40. Canutescu AA, Dunbrack RL Jr. Protein Sci. 2003; 12:963. [PubMed: 12717019]
41. Liang S, Zheng D, Zhang C, Standley DM. Bioinformatics. 2011; 27:2913. [PubMed: 21873640]
42. Dunbrack RL Jr, Cohen FE. Protein Sci. 1997; 6:1661. [PubMed: 9260279]
43. Xiang Z, Soto CS, Honig B. Proc Natl Acad Sci U S A. 2002; 99:7432. [PubMed: 12032300]
44. Zhou H, Zhou Y. Protein Sci. 2002; 11:2714. [PubMed: 12381853]
45. Zhou Y, Duan Y, Yang Y, Faraggi E, Lei H. Theor Chem Acc. 2010; 127:3. [PubMed: 21423322]
46. Summa CM, Levitt M. Proc Natl Acad Sci U S A. 2007; 104:3177. [PubMed: 17360625]
47. Zhu J, Xie L, Honig B. Proteins. 2006; 65:463. [PubMed: 16927337]

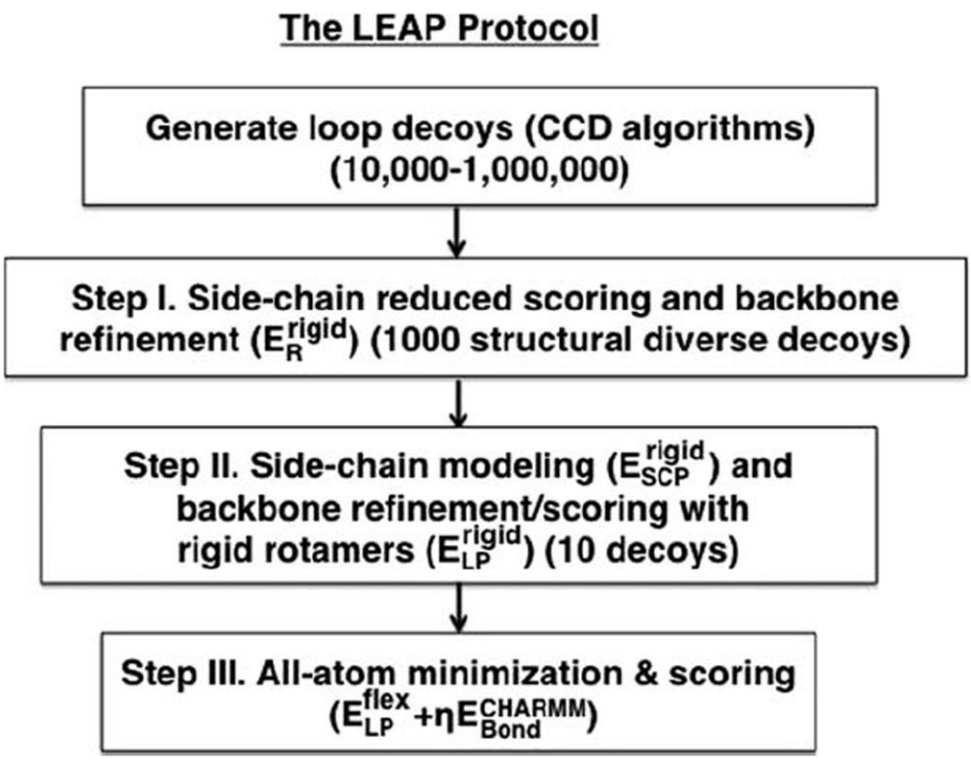

Figure 1.
The protocol of the LEAP algorithm for protein loop prediction.

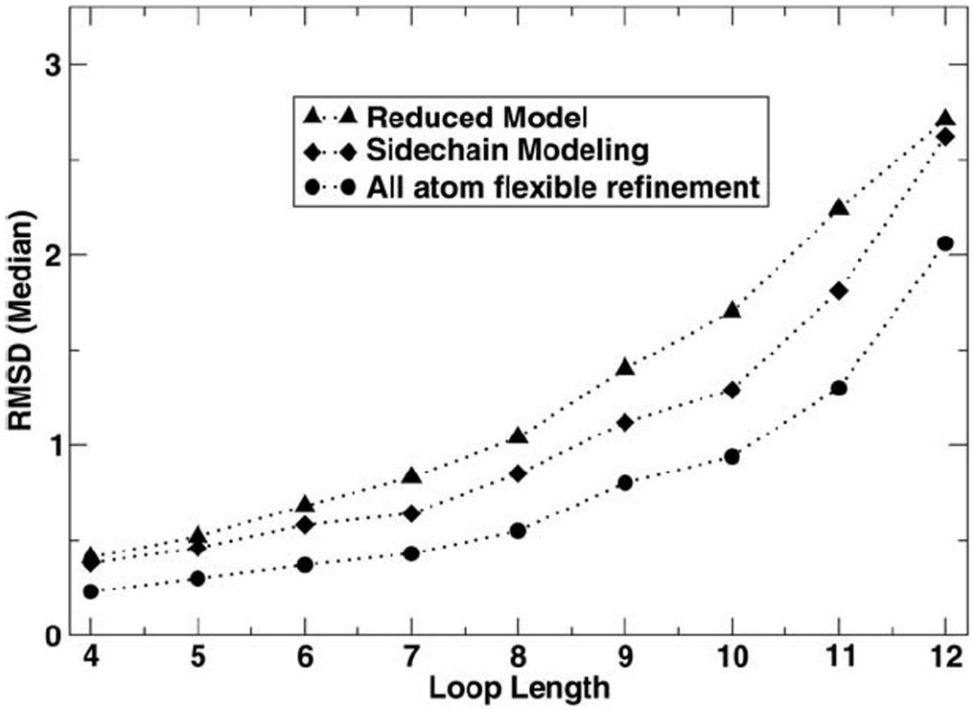

Figure 2.

Median RMSD values of predicted loops as a function of loop length given by three steps of the LEAP algorithm as labeled for loops in 200 test proteins.

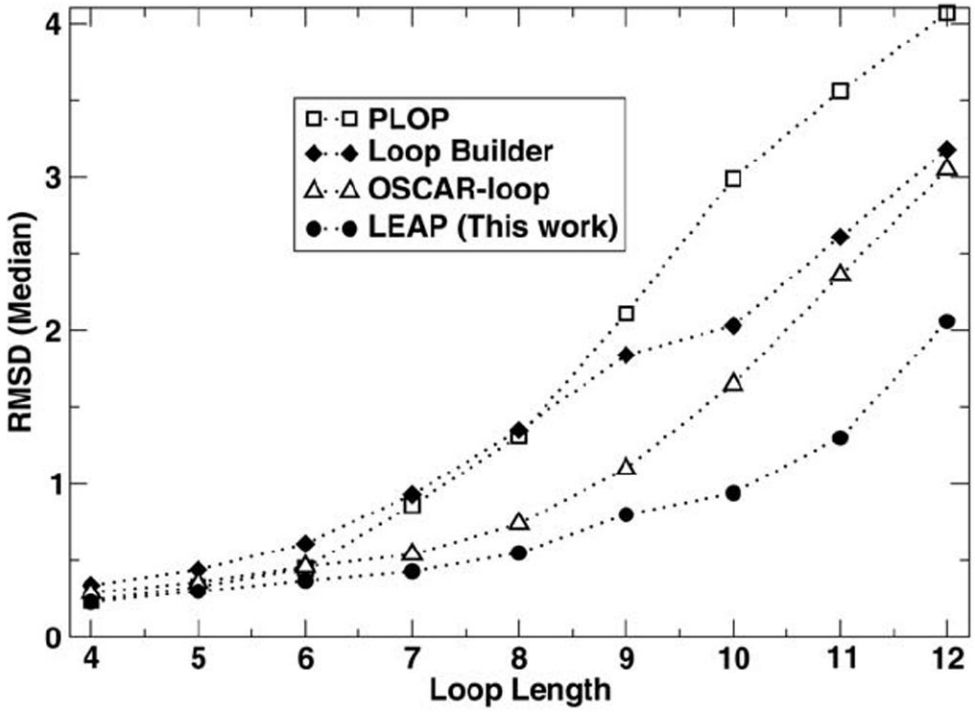

Figure 3.
Median RMSD values of predicted loops as a function of loop length given by several methods as labeled for loops in 200 test proteins.

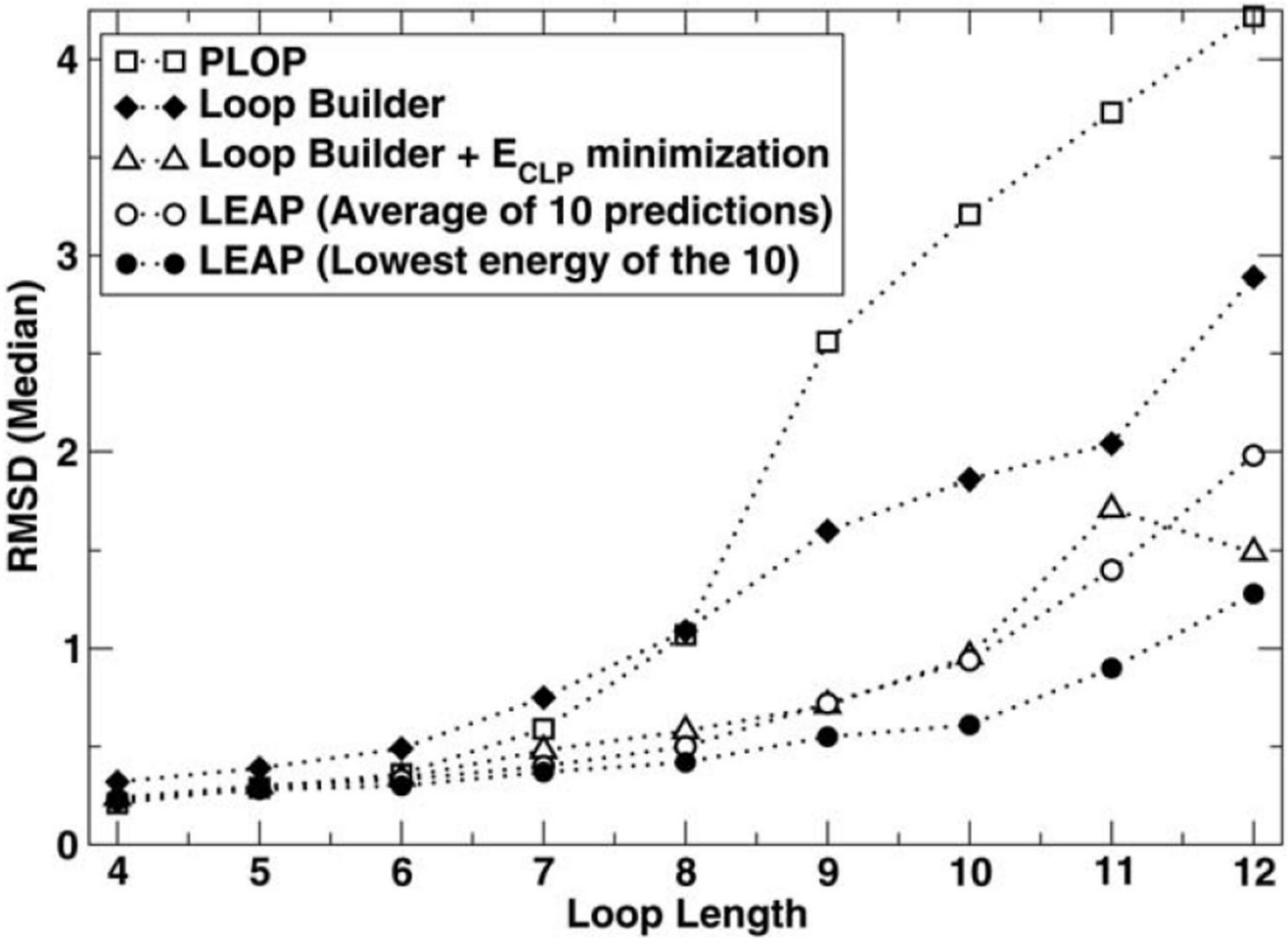

Figure 4.

Median RMSD values of predicted loops as a function of loop length given by several methods as labeled for loops in CASP target proteins.

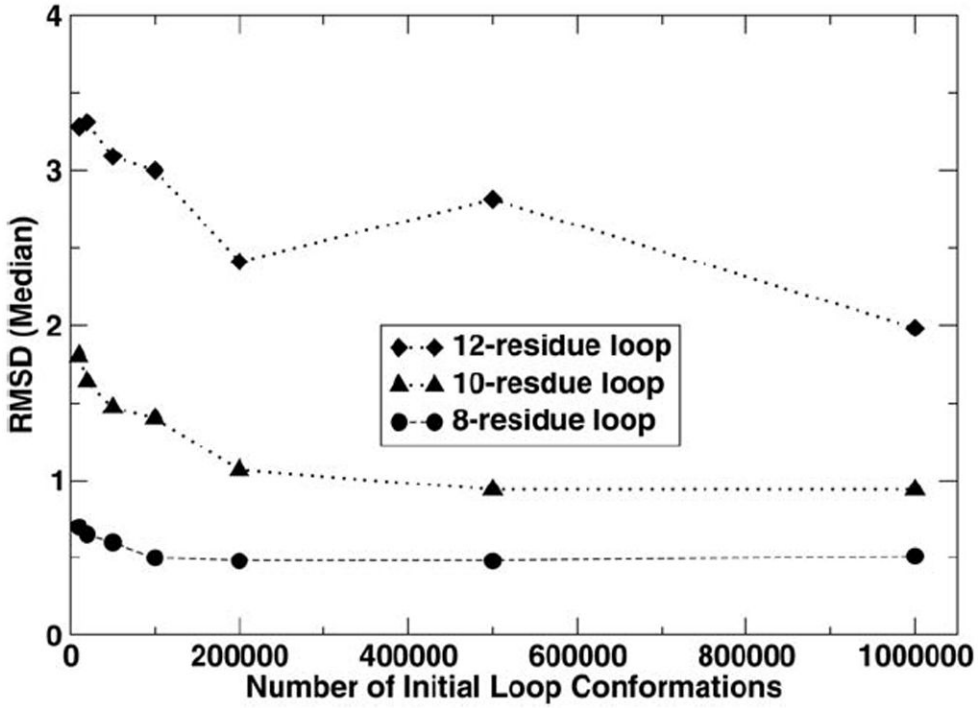

Figure 5.
Median RMSD values of predicted loops as a function of the number of initial loop conformations as labeled for 8-, 10- and 12-residue loops in CASP target proteins.

Table 1

Comparison between LEAP and FREAD for the FREAD dataset.

Loop Length	The average (median) of 30 loops		
	FREAD[a]	The average of 10 predictions by LEAP	The lowest-energy prediction of 10 by LEAP
4	1.29	0.40 (0.23)	0.39 (0.23)
5	2.19	0.43 (0.30)	0.40 (0.27)
6	1.79	0.55 (0.37)	0.49 (0.33)
7	2.53	0.79 (0.45)	0.69 (0.38)
8	2.88	0.98 (0.74)	0.68 (0.56)
9	3.08	1.20 (0.89)	0.93 (0.69)
10	4.25	1.76 (1.07)	1.44 (0.84)
11	4.55	2.56 (1.42)	2.24 (1.08)
12	3.99	3.68 (2.92)	3.14 (2.52)
13	5.54	3.37 (3.12)	2.91 (2.62)
14	6.07	5.10 (4.31)	4.44 (3.70)
15	6.41	5.16 (4.30)	4.58 (4.16)
16	7.50	5.33 (4.68)	4.90 (4.43)
17	7.84	6.84 (6.27)	5.66 (5.50)
18	5.48	7.60 (6.86)	6.53 (6.30)
19	7.67	7.04 (6.52)	5.87 (4.64)
20	7.64	9.01 (8.53)	8.21 (7.82)

[a]The results were obtained from ref.[39]