JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

Protein Loop Modeling with Optimized Backbone Potential Functions

Shide Liang,*[,†] Chi Zhang,[‡] Jamica Sarmiento,[†] and Daron M. Standley[†]

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

[‡]School of Biological Sciences, Center for Plant Science and Innovation, University of Nebraska, Lincoln, Nebraska 68588, United States

ABSTRACT: We represented protein backbone potential as a Fourier series. The parameters of the backbone dihedral potential were initialized to random values and optimized by Monte Carlo simulations so that generated native-like loop decoys had a lower energy than non-native decoys. The low energy regions of the optimized backbone potential were consistent with observed Ramachandran plots derived from crystal structures. The backbone potential was then used for the prediction of loop conformations (OSCAR-loop) combining with the previously described OSCAR force field, which has been shown to be very accurate in side chain modeling. As a result, the accuracy of OSCAR-loop was improved by local energy minimization based on the complete force field. The average accuracies were 0.40, 0.70, 1.10, 2.08, and 3.58 Å for 4, 6, 8, 10, and 12-residue loops, respectively, with each size being represented by 325 to 2809 targets. The accuracy was better than that of other loop modeling algorithms for short loops (<10 residues). For longer loops, the prediction accuracy was improved by concurrently sampling with a fragment-based method, Spanner. OSCAR-loop is available for download at http://sysimm.ifrec.osaka-u.ac.jp/OSCAR/.

INTRODUCTION

Atomic force fields[1−3] are widely used in the prediction and design of protein structure and the simulation of protein dynamics. Widely used force fields such as AMBER[1] or CHARMM[2] typically involve van der Waals interactions, hydrogen bonds, or long-range electrostatics, the parameters of which are obtained from crystal structures of small molecules and quantum mechanics calculations. For example, the torsion parameters for the backbone are often fit to reproduce the ϕ,ψ energy maps obtained by quantum mechanical calculations.[4,5] However, in spite of their wide use, these backbone potentials are known to be biased toward α-helical[6] or β-strand[7] propensities. Inclusion of macromolecular experimental data as part of the target data in parameter optimization enhanced the performance of the CHARMM force field in reproducing crystallographically observed ϕ,ψ distributions in molecular dynamics simulations.[8] Another approach is the direct use of the Ramachandran plot,[9] which shows the empirical distribution of ϕ,ψ angles observed in protein structures as the backbone potential.[10−12] The conformational sampling efficiency was improved significantly by limiting the choices of dihedral angles to those that are known to be physically realizable.[13,14] Also, dihedral probability density functions of an amino acid residue, sometimes considering the identity and conformation of its nearest residues, were developed to produce smooth probability distributions and dihedral angle potentials.[15−18]

Although it is clear from these studies that statistics derived from observed protein structures can be used to improve atomic force field accuracies, such statistical backbone potential may overcount the interactions when combining with force fields. Following the approach used to analytically represent quantum chemical systems,[19] we recently derived orientation dependent energy functions (OSCAR-o) for protein side chain modeling using a series expansion approach.[20] We optimized the parameters by discriminating native side conformations from non-native conformations. The prediction accuracy of OSCAR-o was significantly better than that of several popular physics-based empirical force fields (AMBER and CHARMM) and statistical potential energy functions (OPUS-PSP[21]) for the prediction of individual protein side chains or modeling side chain conformations for the whole protein.

We next applied the OSCAR-o potential to the problem of protein loop selection for decoy sets generated and minimized with different methods.[22] As for side chain modeling, the loop selection accuracy of OSCAR-o was better than that of physics-based force fields (AMBER and OPLS[23]) or statistical potential energy functions (DFIRE[24]) for both the RAPPER decoy set[25,26] and the Jacobson decoy set.[27] Here, we extend OSCAR-o by including potentials that are functions of the backbone torsion degrees of freedom in order to produce a complete OSCAR potential for loop modeling. One motivation for obtaining continuous backbone potentials is the ability to use them for energy minimization or molecular dynamics (MD) simulations.[18]

Consistent with our earlier work, we chose to represent the backbone dihedral potential as a Fourier series. By artificially generating a pool of loop structures and ranking near-native ones as low as possible relative to other decoys, we optimized the parameters of the resulting dihedral basis functions. The side chain dihedral energy and atomic interactions beyond three consecutive covalent bonds were calculated by OSCAR energy functions developed previously.[20,28] Together, these form a complete empirical force field to capture all of the nonbonded interactions for the protein structure.

Received: February 14, 2012
Published: April 6, 2012

In this study, we applied the complete force field to the problem of predicting protein loop conformations (OSCAR-loop). Generally, loop prediction methods can be categorized into *ab initio* methods[22,25−27,29−43] and fragment-based methods.[44−50] The *ab initio* methods generate loop conformations by searching the angle degrees of freedom, while the fragment-based approaches extract loop structures from experimental structural databases such as the PDB.[51] The prediction accuracies of *ab initio* methods can often be improved by extensive energy minimization and accurate scoring for loop decoys.[25,52] Here, we show continued improvement by directly minimizing the OSCAR energy. The accuracy of OSCAR-loop is better than that of several popular *ab initio* methods (LOOPY[35] and Loop Builder[52]) and fragment-based methods (FREAD[48] and Spanner[50]) for loops with lengths less than 10 residues. The highest accuracy was achieved by combining the predictions of OSCAR-loop and Spanner and selecting the one with the lowest OSCAR energy regardless of the loop lengths.

METHODS

Training and Test Protein Sets to Derive Backbone Potential. The training and test proteins were chosen according to the following criteria: the sequence identity between any two proteins was less than 20%, the resolution was better than 2.0 Å, and the *R* factor was less than 0.25. A total of 3960 chains that met the above criteria were downloaded from the Dunbrack Lab Web site (http://dunbrack.fccc.edu/PISCES.php) in May, 2011. A protein was discarded if more than 2% of its residues had incomplete atomic coordinates. As a result, 3315 protein chains were accepted. We randomly selected 200 proteins to generate the test set, and the remaining 3115 proteins constituted the training set.

Loop Definition. We defined the loop regions following Chou and Fasman's method.[53] The α-helix and β-sheet (parallel or antiparalell) were defined as four or more consecutive residues having ϕ,ψ angles within 40° of (−60°, −50°) and three or more residues having ϕ,ψ angles within 40° of (−120°, 110°) or (−140°, 135°), respectively. The remaining regions were considered loop residues. A loop is comprised of consecutive loop residues and was searched after the end of the formerly identified loop so that the selected loops do not share common residues. Those loops were not accepted if they met any of the following conditions: more than half of the loop residues were core residues (solvent accessibility <20%); the distance between any loop residue and heterogen atom was less than 4.5 Å, and the main chain contained a *cis*-peptide bond (ω<90°). Here, we excluded loops with *cis*-peptide bonds for simplicity, and the actual φ,ψ dihedral potential is not dependent on the peptide bond conformation (*trans* or *cis*). As a result, 23 003 loops with a length of six residues were identified from 3115 training proteins (excluding two loops that failed to converge in following loop closure), and 1456 loops with a length of the same size were identified from 200 test proteins. We used six-residue loops for training and testing backbone potential functions because the backbone conformation of extremely short loops is frequently determined by the constraints from anchor residues, while the prediction for long loops is affected by incomplete sampling. The 200 test proteins and the program to identify loops in a protein structure are available for download at http://sysimm.ifrec.osaka-u.ac.jp/OSCAR/.

Loop Closure Algorithm. We adopted the cyclic coordinate descent (CCD) algorithm for loop closure.[36] For the six-residue loops in training and test sets, the ϕ,ψ angles of five randomly selected residues were consistent with the Ramachandran probability, and the angles of the remaining residue could be any value from the interval 0−360°. Here, the Ramachandran plot was derived from loop structures only. We included loop decoys with ϕ,ψ angles not observed in crystal structures in training so that the optimized energy function may correctly recognize non-native angles as high energy. Standard bond lengths and angles were used to build the main chain structure. The distance cutoff between three backbone atoms of the moving C-terminal anchor (Cα, C, and O atoms) of the generated loop and the corresponding atoms in the fixed C-terminal of the observed loop structure was set to 0.1 Å RMSD. If the loop failed to converge after 500 CCD iterative cycles, the loop conformation was initialized with new ϕ,ψ angles, and the cutoff value was increased by 0.001 Å. This procedure was repeated until the cutoff value increased to 0.2 Å. A total of 10 000 backbone decoys were generated for each of 23 003 training loops and 1456 test loops. The decoys often contained severe atomic clashes because the CCD algorithm did not account for the interaction between the loop backbone and the rest of the protein.

Side Chain Modeling and Deriving Backbone Potential Functions for OSCAR-star. The OSCAR-star force field[28] was used to select the top 1000 scored decoys from the 10 000 generated backbone conformations. Briefly, the orientation-dependent OSCAR-star was developed for side chain modeling with a rigid rotamer model. A total of 16 atom types were defined, and the atomic interaction energy was represented as series expansions. The parameters were optimized in order to correctly predict a side chain conformation at a given position, in which a limited number of rigid rotamers were exploited to find the rotamer that had the lowest energy. The optimized OSCAR-star is tolerant to small deviations from a specific rotamer state while recognizing different rotamer states. We have shown that the energy functions are effective in ranking near-native loop decoys as low energy, especially for the decoy set without extensive structural refinement.[28] The energy of a decoy is comprised of the loop interior energy and the interaction energy between the loop and the protein framework. Here, the absent side chain atoms (other than Cβ) of loop residues were not considered in energy calculation. Protein backbone energy was not considered either. Subsequently, OSCAR-star was used to build side chain conformations for the 1000 selected backbone decoys, and the resulting loop structures were used for parametrization of the backbone potential.

The backbone potential function is represented as series expansions:

$$E_{bb} = b_1 \times \cos\alpha + b_2 \times \sin\alpha + b_3 \times \cos 2\alpha + b_4 \times \sin 2\alpha + b_5 \times \cos 3\alpha + b_6 \times \sin 3\alpha \quad (1)$$

where α is a main chain dihedral angle (ϕ or ψ) and b_{1-6} are parameters to be optimized. Equation 1 is easy for parameter optimization and can be converted into the formula used by AMBER to calculate dihedral energy.[1] We consider three types of backbone potential functions for Gly, Pro, and other amino acids and employed a total 36 parameters. The parameters were initialized with random values, and Monte Carlo simulation

Table 1. Prediction Results for Six-Residue Training and Test Loops[a]

	23 003 training loops			1456 test loops		
	RMSD	backbone energy		RMSD	backbone energy	
loop decoy sets	(min∼average)	without	with	(min∼average)	without	with
10 000 backbones	0.49−2.93	1.15	0.99	0.49−2.94	1.19	1.01
1000 with side chains	0.50−2.10	0.94	0.87	0.50−2.10	0.98	0.90
20 minimized loops	0.44−1.43	0.72	0.66	0.46−1.44	0.74	0.68

[a]The initially generated 10 000 backbone decoys without side chains were ranked by the error-tolerant OSCAR-star, and the top 1000 decoys were selected for side chain modeling. The resulting loop decoys with predicted side chains were used to optimize the parameters of backbone potential functions for OSCAR-star. Then, OSCAR-star with the optimized backbone potential was used to rank the 1000 decoys. The top 10 and the other 10 diverse decoys were selected for energy minimization. The 20 minimized decoys were used to optimize the parameters of backbone potential functions for the accurate OSCAR-o and ranked by OSCAR-o methods.

annealing was used to determine the optimal parameters by minimizing the following objective function:

$$\sum_{k=1}^{M} \frac{\sum_{i=1}^{N} [e^{-E(i)} \times \mathrm{RMSD}(i)]}{M \times \sum_{i=1}^{N} e^{-E(i)}} \quad (2)$$

where N is the number of loop decoys with predicted side chains, i.e., 1000 for this work. $E(i)$ is the energy of decoy i including backbone energy and the energy calculated by OSCAR-star, $\mathrm{RMSD}(i)$ is the backbone RMSD to the native conformation of decoy i, and M is the total number of calculated loops, i.e., 23 003. The optimized backbone potential functions were combined with OSCAR-star and called OSCAR-bstar:

$$E_{\text{OSCAR-bstar}} = E_{\text{OSCAR-star}} + E_{\text{bb}} \quad (3)$$

where the backbone potential function to calculate E_{bb} was trained with loop decoys containing discrete errors.

Energy Minimization and Deriving Backbone Potential Functions for OSCAR-o. The parameters of backbone potential functions for the accurate OSCAR-o force field,[20] which was previously developed for side chain modeling with a flexible rotamer model, have to be optimized in a different way from that for OSCAR-star. OSCAR-o contains two parts, the atomic interaction energy and the side chain dihedral angle energy, represented by power and Fourier series, respectively. The parameters of OSCAR-o were optimized by discriminating the native side-chain conformation from non-native conformations at each modeled position. OSCAR-o is very accurate and sensitive to small coordinate errors. The continuous energy functions are especially appropriate for energy calculations for minimized structures. Correspondingly, minimized decoys should be used to optimize the parameters for the backbone potential functions. We selected the top 10 loop decoys ranked by OSCAR-bstar plus 10 low-energy decoys, which had an RMSD more than 1 Å from the top 10 decoys and from each other, for energy minimization. Bond lengths and bond angles were identically given constraints of $1000 \times (l - l_0)^2$ and $1000 \times (a - a_0)^2$. Here, l, l_0, a, and a_0 were the bond length, standard bond length, bond angle, and standard bond angle, respectively. Nonbond interactions were calculated with OSCAR-o excluding the backbone dihedral energy. We applied a maximum of 200 steps of the Powell method to minimize the loop energy,[54] or the minimizer exited when the loop energy change from step to step was less than 0.0001. After energy minimization, the bond energy usually constituted less than 1% of the total energy and was not considered in loop selection. We used resilient constraints for the bonds in order to minimize the loop structure while maintaining the bond geometry (RMSD to the standard bond length ≪0.1 Å after minimization).

The backbone potential for OSCAR-o was also calculated with eq 1. The parameters were optimized by minimizing eq 2 with 20 minimized decoys for each training loop. The optimized backbone potential functions were combined with OSCAR-o and called OSCAR-bo:

$$E_{\text{OSCAR-bo}} = E_{\text{OSCAR-o}} + E_{\text{bb}}^{*} \quad (4)$$

where the backbone potential function to calculate E_{bb}^{*} was trained with minimized loop decoys. That is, E_{bb}^{*} and E_{bb} are calculated with the same formula but different parameters.

Loop Prediction. The prediction procedure (OSCAR-loop) is similar to that for deriving backbone potential functions with some modifications. The optimized backbone potential is used if necessary. Briefly, 10 000, 100 000, and 200 000 backbone conformations are generated for loops with lengths of 4−6 residues, 7−9 residues, and 10−12 residues, respectively. The values of backbone dihedral angles are limited to those observed in the training proteins for all loop residues. OSCAR-bstar is used to rank the initially generated backbone decoys and select the top 1000 decoys, for which side chains are added with OSCAR-star. Then OSCAR-bstar is used one more time to rank the 1000 decoys with predicted side chains. For the top 10 decoys, energy minimization is performed with OSCAR-bo. The minimized decoy calculated as the lowest energy by OSCAR-bo constitutes the prediction.

For fragment-based loop predictions using Spanner, the target loop was treated as an insertion in an alignment to a template lacking the loop residues. Fragments from protein structures with a global sequence identity of 95% or higher were excluded. In addition, minimization was performed on the loop residues only, instead of on the full structure, as is the default behavior. Finally, in order to produce as many native-like loops as possible, candidate fragments were not clustered. Ten predictions per target loop were made by Spanner, but predictions with severe clashes following minimization were discarded.

Evaluation Methods. The RMSD value of the decoy with the lowest energy was calculated excluding the native loop structure or minimized native conformers unless specifically indicated. We used global RMSD for evaluation; the backbone heavy atoms (N, CA, C, and O) were used to calculate the RMSD between the decoy and observed loop structure after aligning the protein framework.

RESULTS

Deriving a Backbone Potential for OSCAR-star and OSCAR-o. We derived backbone potential functions for the accurate OSCAR-o and the relatively coarse-grained OSCAR-star force fields, respectively. With the optimized backbone potential, the prediction accuracy was improved in the protein loop selection for both force fields (Table 1). We obtained remarkably high accuracy with the OSCAR-o method for energy minimized loop decoys. The average RMSD was 0.66 Å for 23 003 training loops with a length of six residues and 0.68 Å for 1456 test loops of the same size. Without the backbone energy, the prediction accuracy of the training loops (0.72 Å) was also slightly better than that of the test loops (0.74 Å). The improvement resulting from using the backbone energy was the same for the two sets of loops, which indicates we did not overfit the parameters. Moreover, the reduction in RMSD was 8% for 1456 test loops, a result that is statistically significant according to a paired t student test (p value < 0.0001). In addition, the RMSD of some decoys decreased during OSCAR-o energy minimization, which enhanced the overall performance for the minimized decoys. For the 1456 test loops, for example, the lowest RMSD of the 20 selected decoys decreased from 0.65 Å to 0.46 Å by energy minimization.

We then investigated the correspondence between the calculated backbone dihedral energy and the φ,ψ distributions obtained from crystal structures. Since the peptide conformation is determined by the total energy of backbone dihedral energy, side chain dihedral energy, and atomic interaction energy, we should not expect that the sum of φ,ψ dihedral energy exactly matches the φ,ψ distributions in crystal structures. Nevertheless, the right handed and left handed α-helix, polyproline II structure, and extended β-strand conformations in the Ramachandran plot derived from crystal structure statistics were consistent with the calculated low-energy regions in Figure 1a.

Loop Selection for Jacobson Decoy Set. We used six-residue loop decoys for optimizing the parameters of the backbone potential because the simple CCD loop closure algorithm frequently failed to generate near native decoys for longer loops. To test the accuracy on loops with different sizes, we combined the derived backbone potential with the OSCAR-o force field, which had shown the highest loop selection accuracy compared to other energy functions in our previous study.[22] For this test, we used the Jacobson decoy set (downloaded from http://www.jacobsonlab.org/decoy.htm in April 2010) containing from 4- to 12-residue loops, which were generated by a systematic backbone dihedral angle search, followed by interactive cycles of clustering, side chain optimization, and energy minimization on selected loop structures. There were 17−158 target proteins of a given loop length and 200−1500 decoys for each loop target. The RMSD to the native structure was calculated for the decoy with the lowest energy, and RMSD values were averaged over all proteins of a given loop length.

The prediction accuracy of OSCAR-bo (the optimized backbone potential in combination with OSCAR-o) was higher than that reported by Jacobson et al. for loops of all sizes in spite of their use of crystal packing constraints to improve the accuracy (Table 2). The performance of OSCAR-bo was also better than that of OSCAR-o, which does not consider backbone energy, in cases where minimized native structures were included in the evaluation to avoid the effects of incomplete sampling. If minimized native structures were excluded, the performance of OSCAR-bo was better than that of OSCAR-o for short loops (4−9 residues) but worse than that of OSCAR-o for long loops (10−12 residues). The decrease in performance of OSCAR-bo for long loops was believed to result from the significantly lower accuracy for some targets, of which none of the generated decoys are near to the native conformation. In a former study,[22] we showed that

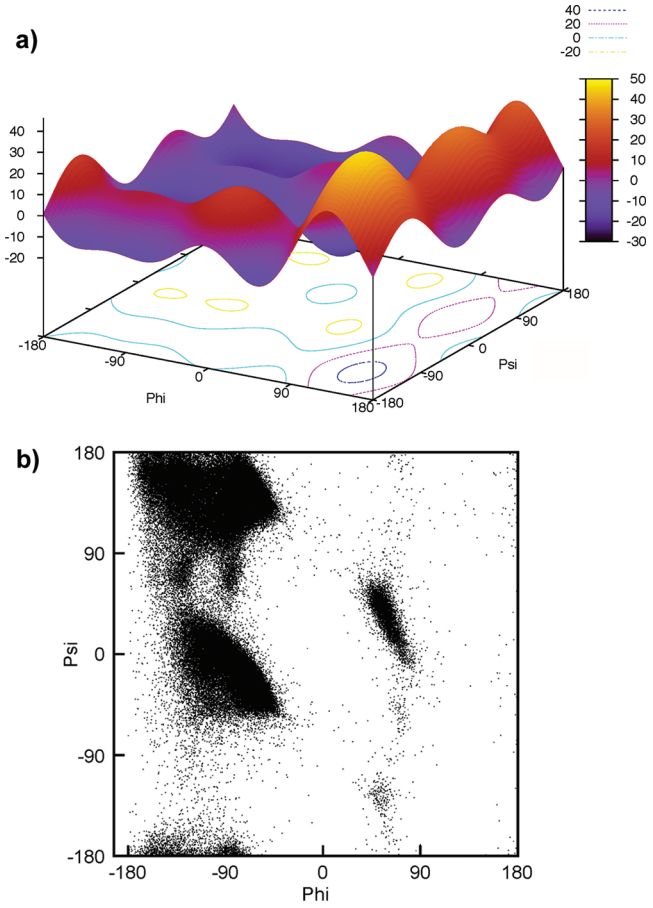

Figure 1. Ramachandran plot for 18 amino acids excluding Gly and Pro. (a) Calculated backbone potential energy surface with parameters optimized to be compatible with the accurate OSCAR-o force field. (b) Distribution of backbone dihedral angles in the training loops.

Table 2. Protein Loop Selection for Jacobson Decoy Set

	mean RMSD (Å)			
loop length	OPLS/SGB-NP (N,CA,C)[a,b]	OSCAR-bo (N,CA,C)[b]	OSCAR-o[c]	OSCAR-bo[c]
4	0.24	0.23	0.30	0.27
5	0.44	0.35	0.45	0.39
6	0.59	0.40	0.48	0.45
7	0.77	0.65	0.54	0.48
8	0.98	0.75	0.79	0.64
9	1.37	0.91	0.65	0.62
10	1.70	1.12	0.80	0.76
11	2.66	1.95	0.72	0.67
12	2.67	1.79	0.64	0.54

[a]Results obtained from Jacobson and co-workers.[27] [b]Backbone RMSD calculated based on N, CA, and C atoms excluding O. [c]Results were calculated based on N, CA, C, and O. The minimized native structures were included in evaluation to avoid the effect of incomplete sampling.

OSCAR-o yielded reliable predictions for high quality decoys when the lowest RMSD was below 0.4 Å. Similarly, the backbone potential is only effective for a high quality decoy set. In the Jacobson decoy set, there were 49 targets containing 10-residue loops. Twenty-nine of them had a lowest RMSD of less than 0.4 Å; for this subset of targets, predicted mean RMSD decreased to 0.46 Å from 0.50 Å with the addition of the backbone energy. For the remaining 20 targets with a minimum RMSD above 0.4 Å, the predicted RMSD of OSCAR-bo was 2.31 Å compared with 2.10 Å for OSCAR-o. We did not use long loops in training to derive a backbone potential appropriate for low quality decoys because our goal is to derive an accurate and continuous OSCAR energy function that can be combined with efficient gradient-based sampling methods, such as MD,[55] to solve the sampling problem in future studies.

Loop Modeling for the 200 Test Proteins. We next implemented the complete OSCAR force field for loop modeling (OSCAR-loop). Side chains were added to the generated loop backbones, and these low-energy decoys with built-in side chains were selected for energy minimization. The RMSD of the loop decoys with the lowest energy were calculated following side chain modeling and energy minimization, respectively. Loop targets with sizes varying from 4 to 12 residues were selected from among the 200 test proteins. The prediction accuracy was improved by energy minimization for loops of all sizes (Table 3). The predicted RMSDs were 0.40 Å, 0.70 Å, 1.10 Å, 2.08 Å, and 3.58 Å for loops with lengths of 4, 6, 8, 10, and 12 residues, respectively.

Table 3. Predicted Mean (Median) RMSD (Å) for the Loops Selected from 200 Test Proteins

		OSCAR-loop		
loop length	loop targets[a]	side chain modeling	energy minimization	LOOPY
4	2809	0.53(0.42)	0.40 (0.29)	0.63(0.46)
5	1863	0.67(0.51)	0.52(0.36)	0.86(0.61)
6	1456	0.89(0.67)	0.70(0.46)	1.18(0.85)
7	1053	1.05(0.79)	0.83(0.54)	1.50(1.06)
8	862	1.39(1.03)	1.10(0.74)	1.73(1.35)
9	634	1.91(1.52)	1.60(1.10)	2.24(1.86)
10	528	2.49(2.12)	2.08(1.65)	2.57(2.25)
11	392	3.25(3.02)	2.73(2.36)	3.15(2.71)
12	325	4.03(3.44)	3.58(3.05)	3.46(2.93)

[a]LOOPY could not process 2x6w and 2qe8. The prediction accuracy of OSCAR-loop made little difference without the two test proteins. In addition, LOOPY failed in loop closure for 31 and 4 loops with a length of four and five residues, respectively. The OSCAR-loop failed for one loop with a length of four residues. The failed predictions were excluded in evaluation.

The accuracy was remarkable for short loops but low for 11- and 12-residue loops due to incomplete sampling. As was discussed regarding loop selection on the Jacobson decoy set, the prediction results of OSCAR energy functions are particularly reliable if the lowest RMSD of the decoy set is below 0.4 Å. For 11- and 12-residue loops, the prediction accuracy was unfavorably affected by the fact that the lowest RMSDs of the generated 200 000 initial conformations were well above the preferred value (1.19 Å for the 11-residue loops and 1.42 Å for the 12-residue loops).

Comparison with Other Loop Modeling Programs. We predicted loop conformations for loop targets selected from the 200 test proteins using the loop modeling program LOOPY.[35] LOOPY accounts for interactions between the loop and the rest of protein as part of the loop closure procedure, which is slow but efficient in generating native-like loop conformers compared with CCD. Here, 5000 initial conformations were generated for loops with different sizes. Extensive energy minimization was not performed by LOOPY, and the prediction results should be compared with those of OSCAR-loop calculated immediately after side chain modeling. The accuracy of OSCAR-loop was better than that of LOOPY for loops varying from 4 to 10 residues but worse than that of LOOPY for 11- and 12-residue loops (Table 3).

Next, we compared OSCAR-loop with Loop Builder developed by Soto et al.[52] using the Loop Builder test set (Table 4). Loop Builder includes extensive sampling of backbone conformations and side chain remodeling of loop residues with LOOPY. The statistical potential DFIRE[24] is used to select a subset of these conformations for energy minimization and ranking with a commercial all-atom force field OPLS/SGB-NP.[27] Loop Builder was designed for the prediction of long loops and was more accurate than other reported methods.[52] OSCAR-loop shows higher accuracy than Loop Builder for eight- and nine-residue loops whether the energy minimization was conducted or not (Table 4). However, the accuracy was lower than that of Loop Builder for 11- and 12- residue loops. The two programs showed similar accuracies for 10-residue loops.

Table 4. Comparison with Loop Builder for Long Test Loops[a]

		Loop Builder		OSCAR-loop	
loop length	no. of loops	side chain modeling	energy minimization	side chain modeling	energy minimization
8	63	1.89(1.59)	1.31(0.97)	1.41(1.04)	1.18(0.77)
9	56	2.71(2.04)	1.88(1.17)	2.26(1.75)	1.68(1.20)
10	40	2.42(2.18)	1.93(1.64)	2.43(2.09)	1.86(1.86)
11	54	3.02(2.48)	2.50(1.95)	3.41(2.98)	2.95(2.51)
12	40	3.15(2.71)	2.65(2.41)	4.13(3.66)	3.41(2.89)

[a]The prediction accuracy was calculated after side chain modeling of the loop residues and further energy minimization. The test proteins and results of Loop Builder were obtained from Soto et al.[52]

We then compared OSCAR-loop with fragment-based loop modeling methods such as FREAD[48] and Spanner.[50] The comparison was based on the standard benchmark set recently used to test FREAD excluding extremely long loops (>12 residues). The accuracy of OSCAR-loop was better than those of FREAD and Spanner for loops with lengths of 4 to 11 residues but lower than both fragment-based methods for 12-residue loops. Interestingly, the accuracy of Spanner, which was recently developed in our group, was also higher than that of FREAD except for the 12-residue loops. Unlike *ab initio* methods, fragment-based methods depend on local sequence similarities and anchor geometries, suggesting similar local structures. Their accuracy was less sensitive to loop length than was the accuracy of OSCAR-loop, which decreased significantly as a function of loop length. To make maximal use of the fragment-based sampling and the OSCAR force field accuracy, we energy minimized the top 10 predictions of Spanner by OSCAR potential functions and combined these with the top

10 predictions of OSCAR-loop. The accuracy of the decoy with the lowest OSCAR energy in the combined predictions was higher than each of three methods, especially for long loops (Table 5).

Table 5. Comparison with Fragment-Based Methods

loop length[a]	FREAD[b]	Spanner[c]	OSCAR-loop	Spanner and OSCAR-loop[d]
4	1.29	1.02	0.51	0.50
5	2.19	1.58	0.46	0.49
6	1.79	1.76	0.70	0.63
7	2.53	2.20	0.95	0.89
8	2.88	2.64	1.20	1.08
9	3.08	2.24	1.41	1.21
10	4.25	3.61	2.87	2.02
11	4.55	3.91	3.68	2.09
12	3.99	4.96	5.46	3.88

[a]Test proteins were obtained from Choi and Deane.[48] A total of 30 targets were calculated given a loop length. [b]Results obtained from Choi and Deane.[48] [c]Similar to FREAD, proteins used to create the fragment database have a sequence identity less than 95% with the query. [d]RMSD of the decoy with the lowest OSCAR energy in the combined predictions of OSCAR-loop and Spanner.

DISCUSSION

Accurate loop modeling is plagued by the chicken-and-egg problem of sampling and scoring. Nevertheless, efficient sampling methods would seem to depend on accurate scoring functions more than the converse, so we sought to first invest our efforts in the development of a highly accurate scoring function. However, as the results for longer (>10 residue) loops show, it becomes difficult to judge the accuracy of the scoring function without effective sampling. But the results for loops of less than 10 residues are consistent: the latest OSCAR force field is an improvement over our previous best-performing force field OSCAR-o. Here, an unsophisticated energy minimization program was developed with the optimized backbone potential, and we found that the RMSDs of loop backbones decreased during energy minimization for most loop decoys, especially those near the native conformation. For example, using the standard OSCAR-loop approach, the decoy ranked number 1 following side chain modeling had an RMSD less than 1 Å in 30 out of the 63 Jacobson eight-residue filtered targets (Figure 2). Furthermore, the mean RMSD of the 30 loops decreased from 0.74 Å to 0.55 Å using energy minimization. The RMSD increased for only one loop structure. For the other 33 loops, the mean RMSD decreased from 2.02 Å to 1.90 Å. Indeed, the improved loop modeling accuracy resulted not only from improved selection of near native decoys compared with non-native decoys but also from the uniformly lower backbone RMSDs of all decoys during energy minimization.

The average run time of OSCAR-loop is 2 h for eight-residue loops and 5 h for 12-residue loops, based on one AMD Opteron 2.7 GHz processor. For the eight-residue loops, Loop Builder[52] is 4 times faster than OSCAR-loop. It takes about half an hour for OSCAR-loop to generate 100 000 initial backbone conformations, select top ranked 1000 backbone conformations, add side chains, and select the top 10 loop decoys for energy minimization. The rate-limiting step was energy

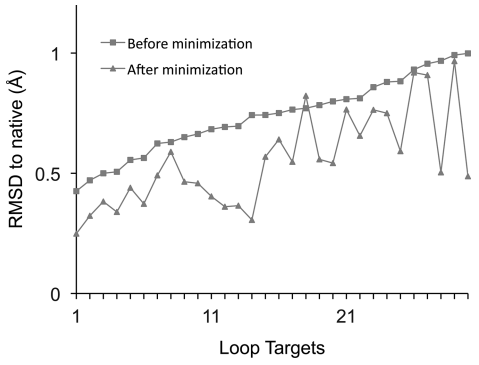

Figure 2. Decreased backbone RMSD by energy minimization for near native decoys. Thirty loops with a length of eight residues and an initial backbone RMSD less than 1 Å were subject to energy minimization.

minimization (about 1−2 h) because we did not use gradient-based methods. In future studies, we will adopt efficient gradient-based minimization methods to speed up the calculations.

We generated a huge number of initial backbone conformations with the simple CCD algorithm to address the sampling problem. In most cases, the prediction accuracy improved as a function of the number of initial backbone conformations (Figure 3). Nevertheless, 100 000 backbone

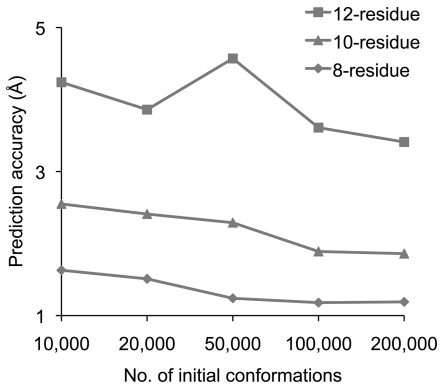

Figure 3. Effect of the number of initial loop conformations on prediction accuracy. The test proteins were obtained from Soto et al.[52] and predicted by OSCAR-loop.

conformations closed by the CCD algorithm were sufficient to achieve the best prediction accuracy for eight-residue loops. More initial backbone conformations did not result in better prediction accuracy, partly due to the subsequent steps of scoring and selection. We selected 1000 decoys for side chain modeling and 10 decoys for energy minimization as a compromise between accuracy and run time. For some loop targets, near native decoys (RMSD < 1 Å) might be discarded at early stages regardless of the large number of generated backbone conformations, which resulted in lower accuracy in the final prediction. For 1hbq in the Jacobson eight-residue test set, for example, the lowest RMSD within the top 1000 decoys was 1.75 Å and 2.06 Å, respectively, when using 100 000 and 200 000 initial backbone conformations. It is unrealistic to perform side chain modeling and energy minimization for all of these decoys. For long loops, we may not solve the problem by simply generating more and more backbone conformations. Efficient sampling algorithms[27,36−38,40−42,50,56] and accurate

energy functions[22,35,39,43,57−60] are both essential to achieving accurate loop predictions, and as mentioned above, the sampling and scoring problems are coupled. With continuous OSCAR energy functions, our long-term goal is to accurately predict conformations for loops of all sizes by molecular dynamics simulation.[55,61] Meanwhile, the impressive results of recently reported fragment-based sampling methods[49,50] can be combined with OSCAR-loop to improve the accuracy for long loops.

CONCLUSIONS

We represented protein backbone force fields as Fourier series expansions. We optimized the parameters by discriminating near-native decoys from non-native ones. The resulting backbone potential was used for loop selection in combination with OSCAR-o energy functions developed for side chain modeling. The accuracy was improved by using the optimized backbone potential for high quality decoys with the lowest RMSD below 0.4 Å. We then used the force fields for *ab initio* loop prediction and obtained high accuracy for short loops (<10 residues). The accuracy for long loops was limited by incomplete sampling but improved by incorporating fragment-based sampling methods.

AUTHOR INFORMATION

Corresponding Author

*Tel.: +81-6-6879-9490. Fax: +81-6-6879-4272. E-mail: shideliang@IFReC.osaka-u.ac.jp.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

S.L. thanks Prof. Yaoqi Zhou (Indiana University−Purdue University, Indianapolis) for his valuable comments and suggestions. This work was supported by a kakenhi grant 24570184: Grant-in-Aid for Scientific Research (C) from the Japan Society for the Promotion of Science (JSPS).

REFERENCES

(1) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1996**, *118*, 2309−2309.
(2) MacKerell, A. D.; Bashford, D.; Bellott; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, *102*, 3586−3616.
(3) Scott, W. R. P.; Hunenberger, P. H.; Tironi, I. G.; Mark, A. E.; Billeter, S. R.; Fennen, J.; Torda, A. E.; Huber, T.; Kruger, P.; van Gunsteren, W. F. *J. Phys. Chem. A* **1999**, *103*, 3596−3607.
(4) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins: Struct., Funct., Bioinf.* **2006**, *65*, 712−725.
(5) Arnautova, Y. A.; Abagyan, R. A.; Totrov, M. *Proteins: Struct., Funct., Bioinf.* **2011**, *79*, 477−498.
(6) Garcia, A. E.; Sanbonmatsu, K. Y. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 2782−2787.
(7) Ono, S.; Nakajima, N.; Higo, J.; Nakamura, H. *J. Comput. Chem.* **2000**, *21*, 748−762.
(8) Mackerell, A. D., Jr.; Feig, M.; Brooks, C. L., 3rd. *J. Comput. Chem.* **2004**, *25*, 1400−1415.
(9) Ramachandran, G. N.; Sasisekharan, V. *Adv. Protein Chem.* **1968**, *23*, 283−438.
(10) Shortle, D. *Protein Sci.* **2002**, *11*, 18−26.
(11) Ormeci, L.; Gursoy, A.; Tunca, G.; Erman, B. *Proteins: Struct., Funct., Bioinf.* **2007**, *66*, 29−40.
(12) Tosatto, S. C.; Battistutta, R. *BMC Bioinf.* **2007**, *8*, 155.
(13) Abagyan, R.; Totrov, M. *J. Mol. Biol.* **1994**, *235*, 983−1002.
(14) Kuszewski, J.; Gronenborn, A. M.; Clore, G. M. *J. Magn. Reson.* **1997**, *125*, 171−177.
(15) Betancourt, M. R.; Skolnick, J. *J. Mol. Biol.* **2004**, *342*, 635−649.
(16) Betancourt, M. R. *J. Phys. Chem. B* **2008**, *112*, 5058−5069.
(17) Rata, I. A.; Li, Y.; Jakobsson, E. *J. Phys. Chem. B* **2010**, *114*, 1859−1869.
(18) Amir, E. D.; Kalisman, N.; Keasar, C. *Proteins: Struct., Funct., Bioinf.* **2008**, *72*, 62−73.
(19) Kryachko, E. S.; Koga, T. Modern Aspects of Diatomic Interaction Theory. In *Advances in Quantum Chemistry*; Löwdin, P.-O., Ed.; Academic Press: Orlano, FL, 1985; Vol. *17*, pp 104−114.
(20) Liang, S.; Zhou, Y.; Grishin, N.; Standley, D. M. *J. Comput. Chem.* **2011**, *32*, 1680−1686.
(21) Lu, M.; Dousis, A. D.; Ma, J. *J. Mol. Biol.* **2008**, *376*, 288−301.
(22) Liang, S.; Zhang, C.; Standley, D. M. *Proteins: Struct., Funct., Bioinf.* **2011**, *79*, 2260−2267.
(23) Jorgensen, W. L.; Maxwell, D. S.; TiradoRives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225−11236.
(24) Zhou, H. Y.; Zhou, Y. Q. *Protein Sci.* **2002**, *11*, 2714−2726.
(25) de Bakker, P. I.; DePristo, M. A.; Burke, D. F.; Blundell, T. L. *Proteins: Struct., Funct., Bioinf.* **2003**, *51*, 21−40.
(26) DePristo, M. A.; de Bakker, P. I.; Lovell, S. C.; Blundell, T. L. *Proteins: Struct., Funct., Bioinf.* **2003**, *51*, 41−55.
(27) Jacobson, M. P.; Pincus, D. L.; Rapp, C. S.; Day, T. J.; Honig, B.; Shaw, D. E.; Friesner, R. A. *Proteins: Struct., Funct., Bioinf.* **2004**, *55*, 351−367.
(28) Liang, S.; Zheng, D.; Zhang, C.; Standley, D. M. *Bioinformatics* **2011**, *27*, 2913−2914.
(29) Fine, R. M.; Wang, H.; Shenkin, P. S.; Yarmush, D. L.; Levinthal, C. *Proteins: Struct., Funct., Bioinf.* **1986**, *1*, 342−362.
(30) Moult, J.; James, M. N. *Proteins: Struct., Funct., Bioinf.* **1986**, *1*, 146−163.
(31) Bruccoleri, R. E.; Karplus, M. *Biopolymers* **1987**, *26*, 137−168.
(32) Shenkin, P. S.; Yarmush, D. L.; Fine, R. M.; Wang, H. J.; Levinthal, C. *Biopolymers* **1987**, *26*, 2053−2085.
(33) Collura, V.; Higo, J.; Garnier, J. *Protein Sci.* **1993**, *2*, 1502−1510.
(34) Zhang, H. Y.; Lai, L. H.; Wang, L. Y.; Han, Y. Z.; Tang, Y. Q. *Biopolymers* **1997**, *41*, 61−72.
(35) Xiang, Z.; Soto, C. S.; Honig, B. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 7432−7437.
(36) Canutescu, A. A.; Dunbrack, R. L., Jr. *Protein Sci.* **2003**, *12*, 963−972.
(37) Coutsias, E. A.; Seok, C.; Jacobson, M. P.; Dill, K. A. *J. Comput. Chem.* **2004**, *25*, 510−528.
(38) Cui, M.; Mezei, M.; Osman, R. *Protein Eng. Des. Sel.* **2008**, *21*, 729−735.
(39) Olson, M. A.; Feig, M.; Brooks, C. L., 3rd. *J. Comput. Chem.* **2008**, *29*, 820−831.
(40) Spassov, V. Z.; Flook, P. K.; Yan, L. *Protein Eng. Des. Sel.* **2008**, *21*, 91−100.
(41) Mandell, D. J.; Coutsias, E. A.; Kortemme, T. *Nat. Methods* **2009**, *6*, 551−552.
(42) Liu, P.; Zhu, F.; Rassokhin, D. N.; Agrafiotis, D. K. *PLoS Comput. Biol.* **2009**, *5*, e1000478.
(43) Li, J.; Abel, R.; Zhu, K.; Cao, Y.; Zhao, S.; Friesner, R. A. *Proteins: Struct., Funct., Bioinf.* **2011**, *79*, 2794−2812.
(44) Jones, T. A.; Thirup, S. *EMBO J.* **1986**, *5*, 819−822.
(45) Wojcik, J.; Mornon, J. P.; Chomilier, J. *J. Mol. Biol.* **1999**, *289*, 1469−1490.
(46) Peng, H. P.; Yang, A. S. *Bioinformatics* **2007**, *23*, 2836−2842.
(47) Hildebrand, P. W.; Goede, A.; Bauer, R. A.; Gruening, B.; Ismer, J.; Michalsky, E.; Preissner, R. *Nucleic Acids Res.* **2009**, *37*, W571−574.
(48) Choi, Y.; Deane, C. M. *Proteins: Struct., Funct., Bioinf.* **2010**, *78*, 1431−1440.

(49) Joo, H.; Chavan, A. G.; Day, R.; Lennox, K. P.; Sukhanov, P.; Dahl, D. B.; Vannucci, M.; Tsai, J. *PLoS Comput. Biol.* **2011**, *7*, e1002234.
(50) Lis, M.; Kim, T.; Sarmiento, J. J.; Kuroda, D.; Dinh, H. V.; Kinjo, A. R.; Amada, K.; Devadas, S.; Nakamura, H.; Standley, D. M. *Immunome Res.* **2011**, *7*, 1−8.
(51) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(52) Soto, C. S.; Fasnacht, M.; Zhu, J.; Forrest, L.; Honig, B. *Proteins: Struct., Funct., Bioinf.* **2008**, *70*, 834−843.
(53) Chou, P. Y.; Fasman, G. D. *Annu. Rev. Biochem.* **1978**, *47*, 251−276.
(54) Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical Recipes in C++: The Art of Scientific Computing*; 2nd ed.; Cambridge University Press: Cambridge, U. K., 2002.
(55) Fiser, A.; Do, R. K.; Sali, A. *Protein Sci.* **2000**, *9*, 1753−1773.
(56) Lee, J.; Lee, D.; Park, H.; Coutsias, E. A.; Seok, C. *Proteins: Struct., Funct., Bioinf.* **2010**, *78*, 3428−3436.
(57) Zhang, C.; Liu, S.; Zhou, Y. *Protein Sci.* **2004**, *13*, 391−399.
(58) Das, B.; Meirovitch, H. *Proteins: Struct., Funct., Bioinf.* **2003**, *51*, 470−483.
(59) Gallicchio, E.; Levy, R. M. *J. Comput. Chem.* **2004**, *25*, 479−499.
(60) Fogolari, F.; Tosatto, S. C. *Protein Sci.* **2005**, *14*, 889−901.
(61) Kannan, S.; Zacharias, M. *Proteins: Struct., Funct., Bioinf.* **2010**, *78*, 2809−2819.

■ NOTE ADDED AFTER ASAP PUBLICATION

This article was published ASAP on April 18, 2012. Figures 2 and 3 have been replaced. The correct version was published on April 19, 2012.