PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Large loop conformation sampling using the activation relaxation technique, ART-nouveau method

Jean-François St-Pierre and Normand Mousseau*

Département de Physique and Regroupement Québécois sur les Matériaux de Pointe, Université de Montréal, C.P. 6128, succursale centre-ville, Montréal, Québec, Canada H3C 3J7

ABSTRACT

We present an adaptation of the ART-nouveau energy surface sampling method to the problem of loop structure prediction. This method, previously used to study protein folding pathways and peptide aggregation, is well suited to the problem of sampling the conformation space of large loops by targeting probable folding pathways instead of sampling exhaustively that space. The number of sampled conformations needed by ART nouveau to find the global energy minimum for a loop was found to scale linearly with the sequence length of the loop for loops between 8 and about 20 amino acids. Considering the linear scaling dependence of the computation cost on the loop sequence length for sampling new conformations, we estimate the total computational cost of sampling larger loops to scale quadratically compared to the exponential scaling of exhaustive search methods.

Proteins 2012; 80:1883–1894.

Key words: protein loop structure prediction; protein folding; Monte carlo sampling; OPEP potential; activation–relaxation technique; protein flexibility.

INTRODUCTION

Protein structure prediction has had much success in predicting the ordered alpha and beta secondary structure components as, often, sequence alone can determine their conformation, especially with the help of previously known homologous protein structures.[1] Loop regions, however, adopt conformations that are not as easily predicted because they lack strict arrangement rules. Prediction is also complicated by the fact that these regions often show intrinsic mobility and, therefore, are not as well resolved by X-ray diffraction crystallography and nuclear magnetic resonance. Over the last 10 years, considerable efforts have gone into developing and refining the prediction ability of three classes of loop-sampling algorithms. The fastest of these are knowledge-based methods that rely on structure databases and sequence homology to generate conformations[2–5] and sport an accuracy as low as 2 Å for sequences length of up to 20 amino acids (a.a.).[6] Ab initio methods, which build loop fragments from scratch and sample the conformation space in search of the lowest energy or best scoring conformations,[7–16] are more demanding computationally but they tend to lead to better results independent of the loop sequence. The last class of loop-sampling algorithms are hybrid methods that combine both algorithms for specific sequences.[17,18] Although many of the previous methods were tested on independent datasets, a quantitative comparison of these methods is presented in Table I of Arnautova *et al.*[19]

Here, we focus on ab initio approaches. These all share a limitation on the maximum loop size they can effectively sample due to the exponential increase in conformational space with loop length. Most studies, until now, have been done on loop datasets of 13 a.a. or less. Their cost, in terms of computational effort, tend to increase exponentially with loop length, following the growth in conformational space.[14,19] Sampling, moreover, is made more difficult by the constraints imposed by the fixed loop endings and is akin to protein folding in a confined environment, a problem which remains challenging.

Grant sponsor: Natural Sciences and Engineering Research Council of Canada; Grant sponsor: Canada Research Chair foundation.
*Correspondence to: Normand Mousseau, Departement de Physique and Regroupement Quebecois sur les Materiaux de Pointe, Universite de Montreal, C.P. 6128, succursale centre-ville, Montreal Quebec, Canada H3C 3J7.
E-mail: normand.mousseau@umontreal.ca
Received 29 November 2011; Revised 19 March 2011; Accepted 30 March 2012
Published online 10 April 2012 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/prot.24085

In this work, we investigate the loop structure prediction problem using the ART-nouveau[20] energy landscape sampling method, which has been used to study protein folding pathways and peptide aggregation of systems of up to 60 a.a.[21,22] We show that the method, although somewhat heavy for short loops of 8 and 12 a.a., can handle large loops of 20 a.a. or more in a very competitive manner, providing both extensive configurational sampling as well as low-energy structures.

METHODS

ART-nouveau potential energy landscape exploration method

We adapted the activation relaxation technique, ART nouveau,[20,21,23] to the exploration of the constrained potential energy landscape of loop segments covalently bound at both extremities to a fixed protein body. ART nouveau is an iterative process consisting of four steps through which the conformation of an atomic system is moved from one local minimum on the potential energy surface to another nearby minimum passing through an adjacent first-order saddle point. (1) Starting from a local-energy minimum, the conformation is first deformed in a random direction taken in the 3N-dimensional loop space. It is pushed along this direction until it leaves the harmonic basin and the lowest eigenvector of the Hessian matrix become negative. (2) The conformation is pushed along the direction of negative curvature, whereas its energy is minimized in the perpendicular hyperplane until the force falls below a small threshold, indicating that the system has converged onto a first-order saddle point. (3) The conformation is then pushed slightly over this saddle point and relaxed, using a damped molecular dynamics, into a new energy minimum. (4) The move from the initial to the final minimum is then accepted or rejected using a Metropolis criterion.[24]

To avoid N^3 operations, the lowest eigenvalue and corresponding eigenvector are computed using the Lanczós algorithm with typically less than 16 force evaluations per step. Implementation details of this very competitive algorithm[25] can be found in Refs. 21,26.

ART nouveau has been used to characterize the energy landscape of complex systems[27] as well as generate folding trajectories,[21] Here, we are interested in sampling the landscape of large loops and identify low-energy structures. To do so, we elected to use a Metropolis algorithm[24] with adaptative temperature. In the original algorithm,[20,25] the acceptance probability is given by

$$P_a = \min\left(1, \exp\left(\frac{-\Delta E_c}{k_B T}\right)\right), \qquad (1)$$

where ΔE is the energy difference between consecutive minima $c - 1$ and c, T is the Metropolis temperature, and k_B is the Boltzmann constant. To keep the probability P_a constant and avoid getting trapped into deep basins, the Metropolis temperature here is adjusted on the fly by applying a Berendsen bath[28] on the acceptance probability of conformation c:

$$P_{avg}(c) = P_{avg}(c-1) + \frac{P_a - P_{avg}(c-1)}{\tau}, \qquad (2)$$

where τ is the coupling parameter and P_{avg} is average effective acceptance rate over the previous w conformations defined as:

$$P_{avg}(c) = \frac{1}{w}\sum_{i=c-w}^{c} \min\left(1, \exp\left(\frac{-\Delta E_i}{k_B T}\right)\right), \qquad (3)$$

The Metropolis temperature for a given $P_{avg}(c)$ is solved iteratively. For our simulations, we selected a window size w of 15 conformations and a coupling τ of 20 with a target acceptance probability P_a of 50%. We also set a minimum metropolis temperature of 300 K to prevent the system from freezing in shallow basins, where the difference in energy between neighboring conformations is small.

For ART-nouveau's exploration, the protein is divided into a fixed protein body and the flexible loop regions. Atoms in the fixed region were assigned based on the experimentally derived native conformations. This procedure is similar to that used in previous studies of loop flexibility such as Refs. 7,29,30.

Dataset

In this study, we used two previously published datasets for the 8 and 12 a.a. loops, respectively. The first set, from Olson *et al.*,[16] is a subset of a large database[31] and is composed of 25 eight-amino acid loops from 22 proteins. The second set is a subset of 38 loops of length 12 a.a. from the Fiser *et al.*[7] dataset. This later subset was used in a number of publications, either in part or as a whole.[11,13,32–34]

Initial loop structures for the ART-nouveau were generated by stretching the loop into an arc of length 3.25 Å times the number of loop amino acids using a harmonic potential applied onto the a.a. center of mass. Five stretched loop conformations were generated per protein with an angle between arc supporting planes of 30°. Between 1 and 5 initial conformations were selected for the loop regions of size 8 a.a. and between 2 and 3 for loops of size 12 a.a., based on the potential energy and rejecting loops segments clashing with the protein body. Using stretched structures as initial conformation, we ensure that simulations are starting far away from the global energy minimum, decreasing possible biases of the initial state.

Table I
Simulation Details for the 8 a.a. Loops of the Olson et al. Dataset[16]

Protein	Nb runs	RMSD initial	Best RMSD	Energy rank (%)	TOP RMSD OPEP	Nb runs finding energy minimum	Average acceptance rate
1a62	3	3.21–5.59	0.72	96.6	2.73	3	0.40
1a62 2	4	4.31–6.76	0.15	55.0	3.18	2	0.31
1aac	3	3.42–5.06	2.01	71.8	2.89	2	0.35
1aba	4	3.75–6.67	0.36	85.1	3.05	1	0.37
1awd	2	4.02–6.06	0.11	32.0	3.94	2	0.34
1c52	3	2.68–4.44	2.65	92.1	3.75	3	0.21
1cbn	5	3.82–5.77	0.33	58.4	2.44	5	0.43
1hfc	3	2.84–6.18	0.43	97.4	2.89	3	0.33
1ig5	5	5.00–7.33	0.00	0.1	3.68	2	0.37
1lit	4	4.13–5.54	0.03	87.3	3.42	4	0.49
1msi	2	5.23–7.22	0.01	91.8	3.59	2	0.32
1nls	1	4.92	4.87	99.5	6.24	1	0.54
1nox	4	3.31–6.95	2.27	96.8	2.96	4	0.36
1opd	3	3.16–5.25	1.17	89.8	3.57	3	0.38
1plc	3	4.70–7.22	1.20	28.8	3.29	3	0.38
1plc 2	1	3.84	1.98	33.8	2.31	1	0.41
1ppn	3	2.15–5.79	0.52	32.8	1.20	3	0.29
1ppn 2	3	4.43–6.58	2.01	99.9	4.43	3	0.42
1ra9	4	3.86–5.42	2.29	70.9	4.36	1	0.32
1rat	3	3.71–5.46	2.99	89.9	4.02	1	0.26
1rro	3	3.18–7.64	0.01	57.1	4.87	3	0.23
1vwj	4	2.75–5.92	1.67	99.6	6.44	1	0.32
3nul	4	3.22–5.86	0.06	88.9	2.12	4	0.23
3seb	2	3.42–4.40	2.14	72.1	4.00	2	0.28
5pal	3	3.73–5.66	0.85	46.4	2.12	1	0.44
Average	3.2	4.71	1.23	71.0	3.50	2.4	0.35
Median			1.01	85.1	3.46		
St. Dev.			1.20	28.3	1.17		

All RMSD are calculated with respect to the native loop structure and are presented in Å. RMSD initial is the distance between the initial stretched structures and the native conformation. Best RMSD corresponds to the structure of lowest RMSD and the energy rank is the percentage of conformation that have lower energy than this structure. "TOP RMSD OPEP" is the RMSD of the structure of lowest energy with the OPEP potential. The acceptance rate of a new conformation is averaged over all runs.

For the 8 a.a. loops, the standard Metropolis criterion, with a Metropolis temperature of 700 K, was used to accept or reject new local minima. For the 12 a.a. loops, we found that using the constant probability rate of Eq. (3) yielded a wider sampling of the conformation space and, therefore, was used for these loops.

For both loop sets, the selected conformations were given 5–10 days of simulation time on single-core Intel Xeon 2.8 GHz microprocessors. Simulation details are presented in Tables I and II. In addition, five batches of preliminary simulations were executed to optimize the ART parameters on the 12 a.a. loop set. Although the results of these preliminary simulations are not included in the analysis presented in the next section, the search for a global energy minimum was done on all generated conformations including preliminary and test simulations.

For longer loop evaluation, we constructed a dataset of 10 proteins using the PISCES server[37] among all proteins with an X-ray structure of resolution lower than 2.0 Å, a sequence identity lower a 25% and a sequence length between 140 and 600 a.a. Regions with no defined secondary structure elements were identified using DSSP,[38] When the loop was found in a multimeric protein, the first chain containing the loop was used and it was verified that the loops did not interact with the removed meres. Because of the difficulty in finding long-loop regions completely devoid of secondary structure, the 19 and 20 a.a. loops presented in Table III have up to 3 a.a. in bend or hydrogen bounded turn conformation, with the exception of 1ofl, which also has 2 a.a. in α-helix conformation. Simulations were executed for 20 days on the same machines as above.

For analysis, we use the global definition of root mean square deviation (RMSD) in which the fixed portions of the proteins are superimposed before calculating the RMSD of the flexible loop region alone without further translations or rotations of the protein. Only the backbone atoms of the loop are included in the RMSD calculations.

OPEP force field

We have modified the optimized potential for efficient peptide-structure prediction (OPEP)[39] coupled to

Table II
Simulation Details for the 12 a.a. Loops of the Fiser et al. Dataset.[7]

Protein	Nb runs	RMSD initial	Best RMSD	Energy rank (%)	TOP RMSD OPEP	Nb runs finding energy minimum	TOP RMSD dFIRE	Average acceptance rate
154L	3	8.61–11.00	2.00	83.0	14.59	3	3.87	0.49
1ARP	3	4.93–7.71	2.40	99.0	5.77	1	5.61	0.49
1CTM	4	6.36–8.41	2.41	65.1	7.13	1	4.59	0.48
1DTS	3	5.11–7.33	2.70	85.9	5.44	1	3.68	0.50
1ECO	3	6.61–8.38	1.15	54.9	4.38	3	3.45	0.50
1EDE	3	5.73–6.65	2.35	76.3	6.32	0	3.20	0.49
1EZM	3	3.91–5.22	0.36	73.0	5.31	3	1.74	0.49
1HFC	3	8.11–9.09	3.01	66.2	11.31	3	7.90	0.50
1MSC	3	6.97–9.27	2.06	95.7	7.82	3	8.17	0.49
1ONC	4	7.20–8.43	1.51	77.1	4.80	1	3.54	0.49
1PBE	3	5.97–7.02	0.94	68.5	4.09	3	2.09	0.48
1PMY	3	4.74–5.85	2.21	71.5	4.78	0	3.43	0.48
1PRN	3	5.24–7.34	1.62	83.7	6.38	2	7.41	0.48
1RCF	3	6.24–9.59	2.22	83.0	4.09	3	4.06	0.48
1RRO	3	3.73–4.73	1.29	89.9	4.42	3	3.85	0.50
1SCS	2	5.80–10.09	0.34	49.5	3.32	2	2.9	0.49
1SRP	3	3.21–5.98	1.14	97.8	3.05	3	2.16	0.50
1TCA	2	6.20–8.42	3.04	6.9	5.11	0	5.21	0.48
1THG	2	5.70–6.41	1.73	24.4	2.58	1	2.92	0.49
1THW	2	5.76–8.14	3.63	99.9	9.61	0	9.45	0.49
1TML	3	7.98–8.70	1.18	17.2	3.85	3	2.93	0.49
1XIF	3	5.72–6.22	0.14	13.2	1.62	1	1.55	0.49
2CPL	4	7.16–9.14	2.84	72.4	6.59	1	5.34	0.49
2CYP	3	4.63–9.03	2.61	86.7	4.20	1	3.84	0.49
2EBN	3	5.68–9.97	2.52	88.1	7.98	1	4.70	0.50
2EXO	3	4.18–7.80	3.39	27.3	5.89	0	3.07	0.48
2PGD	3	5.74–7.88	1.39	95.4	7.36	1	3.09	0.48
2RN2	3	5.22–6.08	1.73	40.3	3.59	0	6.29	0.48
2SIL	3	7.59–9.10	0.00	30.9	3.61	2	1.87	0.49
2SNS	3	6.96–11.74	0.37	55.8	3.93	1	3.91	0.48
2TGI	3	6.75–7.32	1.70	76.1	3.23	3	3.17	0.50
3B5C	3	4.05–6.09	0.30	41.7	2.77	3	2.97	0.49
3CLA	3	4.53–8.99	2.84	30.2	5.46	1	5.80	0.48
3COX	3	5.04–5.67	2.02	89.8	5.61	0	4.84	0.48
3HSC	3	8.68–10.28	1.81	82.4	4.96	3	5.40	0.48
451C	2	7.41–7.65	2.92	80.2	5.93	2	6.11	0.48
4ENL	3	4.53–4.89	0.90	83.8	5.95	2	1.96	0.48
4I1B	3	7.42–8.23	0.01	75.3	10.2	2	6.25	0.49
Average	2.8	6.93	1.75	66.8	5.60	1.7	4.27	0.49
Median			1.77	75.7	5.21		3.85	
St. Dev.			0.98	26.4	2.53		1.87	

All RMSD are calculated with respect to the native loop structure and are presented in Å. RMSD initial is the distance between the initial stretched structures and the native conformation. Best RMSD corresponds to the structure of lowest RMSD and the energy rank is the percentage of conformation that have lower energy than this structure. Two scoring methods were compared to RMSD of the minimum energy conformation, first the OPEP simulation potential (TOP RMSD OPEP), then the dFIRE scoring method[35] (TOP RMSD dFIRE) after conversion of the coarse-grained model to an all-atom representation using SCWRL4.[36] The acceptance rate of a new conformation is averaged over all runs.

ART-nouveau to allow faster sampling of loop regions. OPEP is a coarse-grained potential for which all amino acids side chains are represented by a unified bead except for glycine and proline. All backbone heavy atoms and the hydrogen atom bound to the backbone nitrogen are also represented. To increase the efficiency of the energy computation, all interactions involving two atoms outside of the loop region were removed from the force field. These fixed protein body atoms formed a constant background potential for the docking of the free loop atoms. The forces and energy between the loop's atoms and the rest of the protein are calculated as usual, but the protein's body atoms are not allowed to change conformation. This potential was successfully used to study protein folding[21,23,40] and peptide aggregation.[22,41–43] OPEP was recently compared to the AMBER99SB and OPLS-AA all-atom force field on two small peptides by parallel tempering metadynamics and was found to be in agreement with the two detailed potentials and could reproduce the features of the free-energy landscape at a much lower computational cost.[44]

Table III
Simulation Details for the 19–20 a.a. Loops Dataset

Protein	Loop length	Loop	Secondary structure a.a.	RMSD initial	Nb runs		Nb runs finding energy minimum		Best RMSD	Energy rank (%)	TOP RMSD OPEP	Avg % accepted conformations	
					SS	LS	SS	LL				SS	LL
1gwe	20	G406-D425	2	6.58–13.58	5	10	1	0	4.11	94.5	8.39	0.52	0.38
1ofl	20	Y434-N453	4	11.41–18.86	5	10	0	1	9.45	97.3	12.53	0.53	0.39
1q6z	20	V329-Q348	2	2.75–10.33	5	10	1	0	1.69	97.0	9.30	0.58	0.38
2ess	20	C139-P158	2	5.35–9.85	5	10	1	0	3.80	89.5	7.13	0.54	0.39
2gag	20	F445-P464	0	5.20–21.96	5	10	1	0	2.17	29.0	10.32	0.60	0.37
2i9i	20	H59-H78	0	2.20–12.69	5	10	2	0	0.83	2.8	1.10	0.56	0.41
2vk8	20	G343-S362	3	3.41–8.31	5	10	1	0	1.85	78.4	5.40	0.59	0.39
3cx5	20	N227-G246	3	3.93–8.96	5	10	1	0	1.80	84.0	6.98	0.57	0.39
3d3y	19	L218-I236	2	3.19–7.52	4	10	1	0	1.62	72.4	2.89	0.52	0.39
3igx	20	E261-I280	2	3.29–12.25	5	10	1	0	1.93	90.9	7.73	0.58	0.40
Average				8.26					2.98	73.6	7.18	0.56	0.39
Median									1.89	86.8	7.43		
St. Dev.									2.37	32.0	3.21		

Secondary structure a.a. is the number of a.a. in turn and bend conformation and, in the case of 1ofl, in α-helical conformation as annotated by DSSP.[36] RMSD initial is the distance between the initial stretched structures and the native conformation. SS and LS refer to the short step and long step parameterization, respectively. Best RMSD corresponds to the structure of lowest RMSD and the energy rank is the percentage of conformation that have lower energy than this structure. "TOP RMSD OPEP" is the RMSD of the structure of lowest energy with the OPEP potential. The acceptance rate of a new conformation is averaged over the number of runs.

RESULTS

The following analysis of the ART-nouveau method is divided into three sections. First, we evaluate the ability of the method and the OPEP potential to sample conformation of low energy in the vicinity of the native structure providing a proper score. Then, the ability of the method to sample the conformational space and to find the global energy minima regardless of the native structure is presented for the short loops of 8 and 12 a.a. and the long loops of 19 and 20 a.a. Finally, we evaluate the scaling performance of the method as a function of loop-length.

Conformation scoring

The ability of the OPEP potential to find low-energy loop conformations compatible with the crystallographic structure is presented in Tables I and II. We find that for the 8 a.a. dataset, our results show a comparable accuracy to low-energy structures to that of Olson *et al.*[16]. More precisely, the lowest energy conformations for our simulations are, on average, 3.50 Å (St. Dev. 1.17 Å) away from the native structure, as compared with 3.89 Å for their lattice-based work and 3.14 Å for their all-atom MD simulations.[16]

Even though the trajectories sample conformations within 1.75 Å of the native state for the longer 12 a.a loops, the lowest energy structures show an average RMSD with respect to the native structure of 5.60 Å (St. Dev. 2.53 Å). This discrepancy is due to the coarse-grained nature of the OPEP potential, which does not discriminate sufficiently between various steric packing, as well as to the rigid spatial representation of the non-loop protein regions which prevents structures from adopting the optimal conformations.

To test the impact of these two limitations, we induce flexibility by reconstructing the coarse-grained side-chains of the whole proteins using the SCWRL4 automated tool,[36] then rescored the all-atom representations using dFIRE.[35] This analysis show that lower RMSD low-energy structures were sampled with ART nouveau but improperly scored by the modified OPEP potential. The resulting average RMSD of the best scored conformations to the native structure is improved to 4.27 Å (St. Dev. 1.87 Å), essentially identical to the average 4.32 Å obtained by Zhang *et al.* with a similar protocol[45] and slightly higher than other ab initio methods including FALCm4 that scores using dFIRE potential with 3.84 Å RMSD[34] and LOOPER with 4.08 Å RMSD,[32] Methods making use of predefined structures, such as ROSETTA do, of course, better: 3.62 Å RMSD for ROSETTA[13] and 2.3 Å RMSD for ROSETTA with a kinetic closure algorithm.[33]

The efficiency of the reconstruction and rescoring of the ART nouveau-generated datasets suggests that even though OPEP could not fully discriminate between the various energy minima on the energy landscape, ART-nouveau samples the configurational space rather efficiently. This is confirmed by the fact that the smallest RMSD between trajectories and the experimentally derived native structure is only, on average, 1.23 and 1.75 Å, for the 8 a.a. (Table I) and 12 a.a. loops (Table II), respectively.

We observe similar results for the dataset composed of loops of 19 and 20 a.a, the RMSD of the conformations of lowest energy to the native conformations averages to 7.17 Å (St. Dev. 3.21 Å) (Table III), which is comparable to, or better than previously published results on loops of the same size, that is, ~7 Å for CABS, ~9 Å for Rosetta, and ~12 Å for MODELLER on the Jamroz and Kolinski dataset,[46] and 10.49 Å for MODELLER, 10.64 Å for RAPPER, 11.14 Å for PLOP, and 7.64 Å for Original FREAD on the Choi and Deane dataset.[6] However, the lowest RMSD to the native structure observed in these 20 a.a. loop studies is of 2.98 Å (St. Dev. 2.37 Å), consistently lower then the lowest RMSD for the above studies, a value that ranges between ~4–5 Å[46] and 5.20–8.43 Å average RMSD.[6] This suggests that our prediction precision for large loops is competitive with previously published methods and that the ART method can sample conformations closer to the native structure even when this structure is not the global energy minimum of the used potential.

The focus of this project is to evaluate the ability of the ART method to sample a wide range of loop structures and identify low-energy conformations on a potential energy surface. We, therefore, leave aside the issue of proper scoring and prediction capacities, which are entirely dependent on the chosen energy potential, to analyze sampling capacities of ART-nouveau which is potential-independent. In the following analysis, RMSD values are calculated with respect to the global low energy conformations of the energy potential instead of the native conformation. For this purpose, the low computational cost of the OPEP potential is well suited as it allows for longer simulation times and wider sampling of the conformational space to identify the global low energy structures.

Exploration of the conformation space for the 8 a.a. and 12 a.a. loop dataset

ART-nouveau's sampling ability can be evaluated by characterizing the volume of the conformation space sampled by the method and its ability to find the conformations of global lowest energy on the OPEP potential energy surface. In particular, proper care must be taken to insure that conformations are not trapped in local energy minima, away from the native state.

For the 8 a.a. loops, we see that in the 23 loops for which more than one simulation was executed, it was possible to recover he same lowest energy minimum, as defined with OPEP, at least two times or more in 18 of them (Table I), suggesting that the exploration of the conformational space is sufficiently thorough to reach regularly the global-energy minimum. For the 12 a.a. loops, analysis of our preliminary simulations shows that in six cases, conformations of lowest energy were also sampled in the preliminary simulation, but not in the production simulations. The lowest energy conformations of these six preliminary simulations were used for further analysis. When considering all calibration and production runs, the global energy minimum was found two times or more for 31 of the 38 loops, 19 of which were found twice or more in the production runs of Table II.

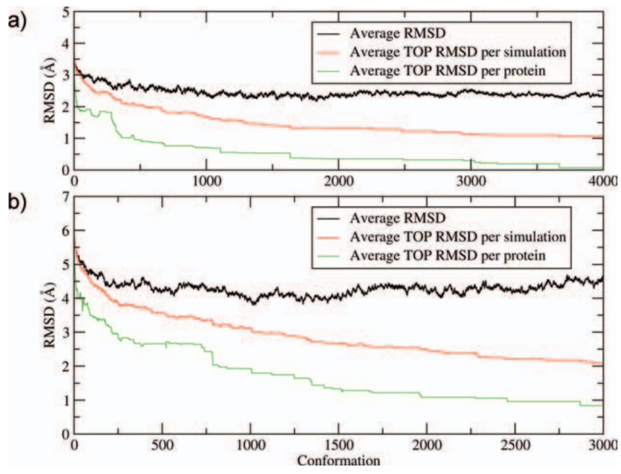

Figure 1

RMSD evolution for the (**a**) 8 a.a. and (**b**) 12 a.a. loops. In black, the current conformation's RMSD of each simulation is calculated to the global energy minimum conformation of the sampled protein. The average TOP RMSD is calculated between the global energy minimum of a system and the lowest energy conformation found so far per simulation (red) or per protein (green). Curves are presented from top to bottom in the same order as in the legend. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

To understand how sampling occurs, we plot the evolution of the RMSD, averaged over three different subsets as a function of event number for both sets of loops in Figure 1. The black curve shows the averaged RMSD of the current ART event conformation, computed with respect to its respective global energy configuration. We see that this measure reaches a plateau after roughly 500 steps and remains around 2.4 Å for the 8 a.a. loops and 4.4 Å for the 12 a.a. loops. These distances are relatively near the maximum deformation distance achieved by stretching the loop, 4.63 and 7.49 Å, respectively, which indicates that each trajectory samples widely the energy landscape.

The red curve shows the evolution of the RMSD of the lowest energy conformation identified, or TOP RMSD, for each trajectory launched and averaged over all simulations. This quantity shows how a group of trajectory can be used to identify the lowest energy basin. When simulations are examined individually, we first see that not all runs for a given loop sequence sample the lowest energy conformation but that, overall, the probability of passing nearby this conformation increases with the number of steps, albeit at a constantly slower pace. After 4000 and 3000 steps, respectively, the 8 a.a. and 12 a.a. loop simulation sets reach an average per simulation RMSD value of 1.05 Å and 2.05 Å.

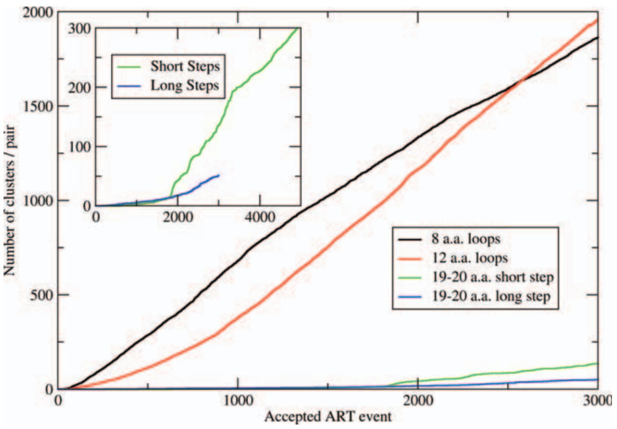

Figure 2

Average number of clusters common between two simulations run for each protein defined by a RMSD distance of less than 1.0 Å between clusters central conformation. Inset is for short-steps and long-steps parameterization of the 19–20 a.a. loops simulations. As the probability of two simulations overlapping is proportional to the square of the number of simulations, the plots are normalized by the number of pairs of simulations per protein.

It is useful to follow convergence of the full set of simulations. In the same figure, we combine all runs for each loop and plot the overall average TOP RMSD for each sequence (green curve). As expected, we see a faster convergence in the first ART steps, leading to an average RMSD to the global energy structure of 0.1 Å after 3700–4000 conformations for the 8 a.a. loops and 1.25 Å after 1600 conformations, and 0.84 Å after 3000 conformations, for the 12 a.a. loops.

For the 12 a.a. loops, we see a lowest average RMSD of 1.0 Å because not all sequences manage to find their global energy-minimum structure in the production runs here presented. In some cases, these structures were only identified in the preliminary simulations.

Although not all runs for one protein converge to the global energy minimum, all individual runs do overlap, suggesting that longer runs would allow all trajectories to find the global energy minimum. To see this, conformations were divided into clusters of maximum RMSD of 0.6 Å between each member using a hierarchical clustering algorithm and an average linkage clustering criteria.[47] The center of each cluster for a given simulation was compared with that of all other runs for the same loop and a new clustering is performed on this dataset. Figure 2 presents the average number of clusters with a maximum RMSD of 1.0 Å that are sampled in at least two runs for the three datasets studied here. For 8 a.a and 12 a.a. loops, we observe a linear increase as a function of increasing number of visited conformations, which indicates that, on average, trajectories continue to sample the configurational space without being trapped as simulations progress.

The size of the sampled conformational space can also be estimated by measuring the number of clusters within a fixed minimum RMSD between each other. The evolution of this RMSD rank is presented in Figure 3 for the 8 a.a. and 12 a.a. loops. In both cases, we see a rapid increase in the number of clusters meeting the minimum RMSD cutoff for the 400 first conformations sampled followed by a slower linear stage to persists until the end of the runs.

Two different behaviors can be identified depending on the size of the RMSD cutoff. With higher minimum RMSD cutoffs, we measure the diameter of the hypervolume accessible to the loops. The rapid convergence of this quantity indicates that the initial configurations are chosen properly as they rapidly bring the various simulations in very different parts of configurational space.

With a minimum RMSD of 2 or 3 Å, the average number of clusters with minimum RMSD between each other gives us a sense of the finer sampling the configurational space. The continuous growth of the curves even after 300–5000 steps indicates that the various trajectories are still sampling the conformational space at a finer level.

Exploration of the conformation space for novel 19–20 a.a. loop dataset

The 19–20 a.a. loop dataset was constructed to test the efficiency and scaling of our method on larger model loops. As described below, because of the increase in configurational space, the parameters used in ART nouveau for sampling smaller loops is not optimal for this dataset. Therefore, two different parameterizations are used on the 19–20 a.a. loops. The first one, dubbed "short step," produces conformations with an average interminimum RMSD of about 0.5 Å. It is the set used in our study of both 8 a.a. and 12 a.a. loop datasets. The second one, which we call "long step," generates conformations with an average of 1.1 Å RMSD displacement between adjacent minima. To obtain this increased travel distance between minima, we modified two parameters in the ART method. The first one is the number of iteration of the Lanczos routine used to find the eigenvector of lowest eigenvalue of the Hessian matrix.[26] By reducing from 12 to 4 iterations, the weight of the previous eigenvector, which is used as the seed direction in Lanczos, is more important, stabilizing the trajectory and reducing the impact of local fluctuations in Hessian curvature. The second modified parameter is the force threshold used in relaxing the forces in the perpendicular hyperplane to the activation relaxation. By increasing this threshold from 1.9 to 2.5 kcal/(mol Å), the probability of loosing a negative eigenvalue is further reduced. These modifications decrease the reliability of the saddle point and the physical basis for the initial minimum—saddle—final minimum pathway. However, here we are interested in

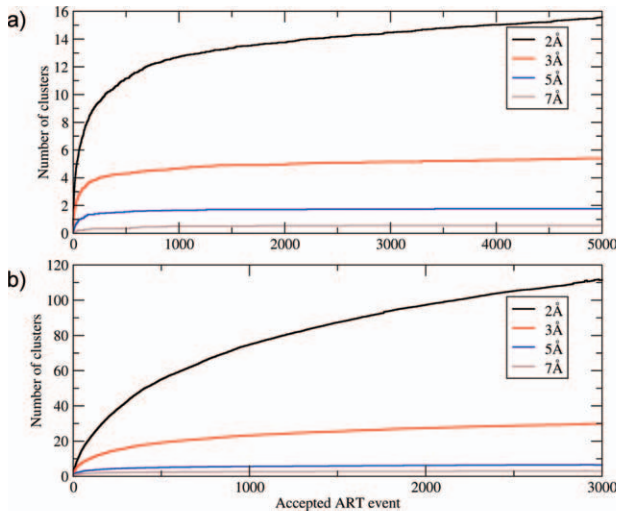

Figure 3

Size of the largest group of clusters per simulation for the (**a**) 8 a.a. and (**b**) 12 a.a. loops with minimum RMSD between each member of the group greater then 2 Å (black) to 7 Å (brown). Curves are presented from top to bottom in the same order as in the legend. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

moving through the landscape and mostly making sure that the visited minima are acceptable thermodynamically. Although fairly aggressive, this set of parameters allows us to keep a reasonable acceptance rate.

The details of both simulation sets are presented in Table III. Comparing the average potential energy of the sampled minima for the two different parameterization sets, we see that the use of short steps leads to an average potential energy ~7 kcal/mol lower and a continued convergence toward low energy structures compared to the long steps (data not shown). Nine of the 10 sequence-dependent global energy minima were found in the short step simulations.

We first characterize the sampling of the configurational space for the 19–20 a.a. loops by following the evolution of the conformation RMSD as a function of the number of generated conformations. As with the loops of 8 and 12 a.a., we observe that the average RMSD, measured from the native state, remains very high at 6.3 Å, close to the value of the initial conformation, 8 Å [Fig. 4(a) black].

However, the reduction of the average RMSD of the lowest energy conformation found so far in each run to the global minimum is much slower than with the smaller loops, reaching 5.0 Å in the first 5000 steps. This can be explained by the fact that out of 10 loop models, only the trajectories of protein 2i9i sampled the global minimum more then once (see Table III). The larger conformation space of the longer proteins means that there is a smaller probability that two independent trajectories overlap, finding the same folding energy funnel to the global minima. Indeed, the simulation runs that do find the global energy minima evolve at rate roughly two times slower than the 12 a.a. loops. Not surprisingly, the number of clusters shared between runs of the same 19–20 a.a. loop also grows at a slower rate that for shorter sequences. (Fig. 2).

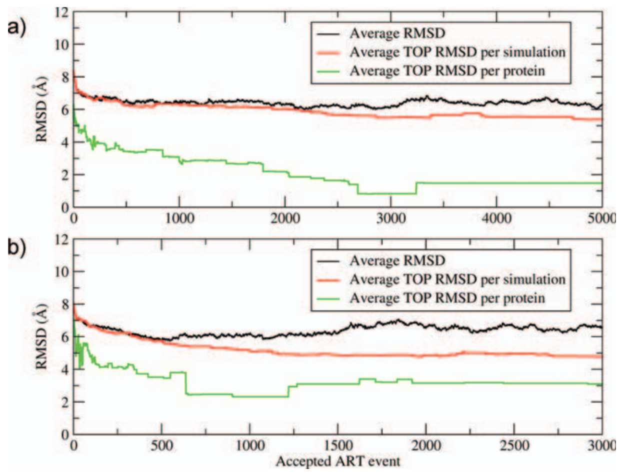

Figure 4

RMSD evolution for the 19–20 a.a. loops using (**a**) short steps and (**b**) long steps parameterization. In black, the current conformation's RMSD of each simulation is calculated to the global energy minimum conformation of the sampled protein. The average TOP RMSD is calculated between the global energy minimum of a system and the lowest energy conformation found so far per simulation (red) or per protein (green). Curves are presented from top to bottom in the same order as in the legend. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

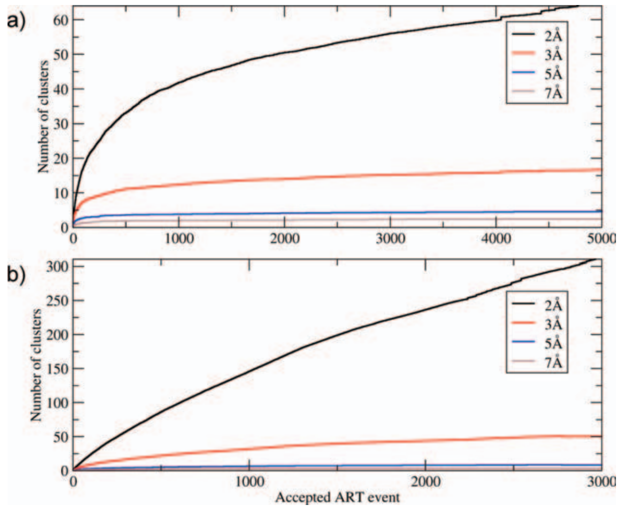

Figure 5

Size of the largest group of clusters per simulation with minimum RMSD between each member of the group greater than 2 Å (black) to 7 Å (brown). Curves are presented from top to bottom in the same order as in the legend. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Table IV
Scaling Parameters of the Sampling of One New Conformation Through ART-nouveau Method

Loop size	Protein size scaling factor (μs)	Protein size scaling correl.	Nb. force. per new conformation
8 a.a.	4.76	0.97	27,123
12 a.a.	7.38	0.98	31,572
20 a.a.	8.59	0.98	31,596 (ss) 28,030 (ls)

The protein size scaling factor represents the slope of the time needed for one force field evaluation in relation to the protein's size for three loop size obtained through linear regression. Also presented is the scaling factor correlation coefficient and the average total number of force field evaluations needed to sample one new local minimum. Abbreviations "ss" and "ls" refer to the short step and long step parameterization of the 19–20 a.a. loop simulations.

The evolution of the cluster rank metric for the short step simulations presented in Figure 5(a) also points to difficulties in sampling the conformation space in each simulation run using the short step parameterization. With a larger conformation space and the possibility of larger RMSD between two loop conformations, the cluster rank metric or the 19–20 a.a. loops should increase at a higher rate than the 12 a.a. loops. However, the shorter average interminimum RMSD for the short step simulations (0.52 Å) compared to the 12 a.a. loops simulations (0.72 Å) may explain the lower cluster rank observed.

As mentioned earlier, to correct for the limited sampling of the short-step ART nouveau 19–20 a.a simulations, we also launched a number of runs with longer ART nouveau step. These newly generated trajectories sample saddle points of higher energy, and yet, sampling speed is greatly enhanced. Indeed, with the new parameters, we see that the number of clusters with minimum RMSD of 2 Å increases four times more rapidly in the first 1000 events [black in Fig. 5 (b)] as compared to the short steps simulations [Fig. 5 (a)]. Moreover, the rate of increase does not slow down after the first 1000 events. On the other side, the median lowest RMSD structure to the global minimum is found at a distance of 1.92 Å (average 2.44 Å) compared to 3.71 Å (average 4.14 Å) for the short steps simulations even though the long step simulations are much less likely to find the global energy minimum of the various loops. This means that the long steps parameterization is better suited for wide sampling of the energy surface, whereas the small steps parameterization are better for in-depth sampling and structure refinement. This is also demonstrated by the speed at which the long step simulations will find a low energy minimum of low RMSD [green line in Fig. 4 (b)] compared to previous parameterization [Fig. 4 (a)].

Scaling

ART manages to avoid the exponential increase in complexity of the conformational space as a function of the number of amino acids by not attempting to sample the whole configurational space but rather sampling low-energy structures only through the generation of connected physical trajectories (at least, when using small steps). The time needed for the ART method to pass from one local minimum to neighboring minimum is proportional to the number of integration steps required to generate a new conformation, that is, activate to a nearby saddle point and relax into a new minimum. The cost of each of these step is, of course, dependant on computational efforts required by the force field. The modifications to the OPEP potential treating the protein's body as a background potential lead to a theoretical scaling of the force field computation time that is linear with the size of the loop (n) and the size of the protein (N), leading to an order of $O(n \times N)$. Experimental scaling results are presented in Table IV, where we see that, as expected, force field evaluation times scale linearly with the protein's size with an average correlation coefficient of 0.98 ± 0.01 and scales linearly between loop size 8 a.a. and 12 a.a and sublinearly between loops of 12 a.a. and 20 a.a (which can be explained by the presence of cutoffs for some parts of the potential). The average number of force field evaluation per even is not influenced greatly by the size of the loop with an average of 30,000 ± 2000 evaluations [see Table IV and Ref. 25], and the total empirical scaling factor for the sampling of a new conformation is linear with the loop size.

Scaling is also measured by the number of sampled conformations needed to reach a given conformation of interest as a function of loop length. In both cases presented in Figure 1(a,b), the RMSD measured with respected to the global energy minimum shows a fast collapse within the first 1500 sampled events, followed by a slow optimization. For the 19–20 a.a. loops, this collapse is evident in the 3000 first conformations sampled. What

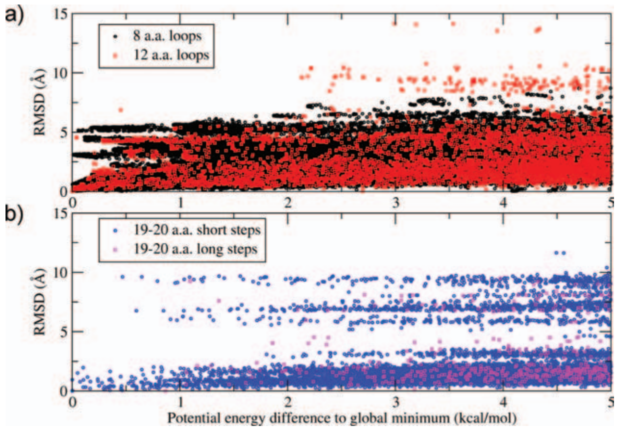

Figure 6
Distribution of the RMSD to the global energy minimum structures for (**a**) small and (**b**) large loops for structures of potential energy 5 kcal/mol or less over the global minimum. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

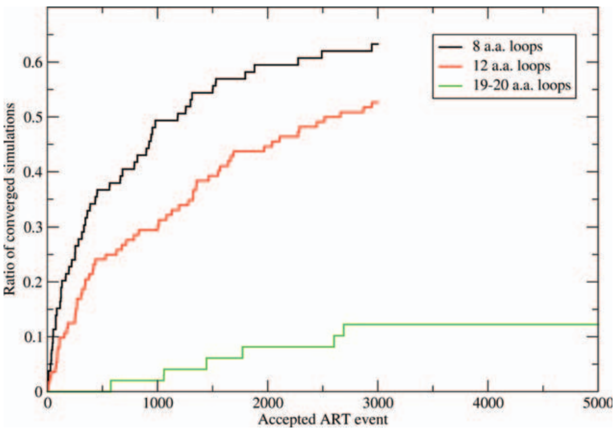

Figure 7

Proportion of the number of simulation that have found their protein's global minimum loop structured as a function of the number of accepted conformations based on a 0.1 Å RMSD cutoff to the global minimum. Curves are presented from top to bottom in the same order as in the legend. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

differs is the minimum RMSD to which the sampling converges after the initial fast collapse. As shown in Figure 7, after 500 sampled conformation, 29 of the 79 8 a.a. loop simulations have visited the energy global minimum for their respective sequence (36.7%) with 27 of the 108 12 a.a. loop simulations doing the same (25%). After 1500 conformations have been sampled, these ratio are 54.4 and 40.7%, respectively. Therefore, to maintain the same probability, that is, requiring that at least one simulation per protein finds its global minimum conformation in 1500 sampled conformations, we need 34–47% more generated conformations per 12 a.a. loop then per 8 a.a. loop, which is in line with the 50% increase in loop length. Combining this linear increase in the number of conformations needed to the previous linear increase in simulation time required to generate a conformation, we estimate the total computational efforts are quadratic with loop length.

DISCUSSION AND CONCLUSION

Small-loop structure prediction methods have seen significant improvements in terms of required efforts and achieved precision in the last decade. Loop predictions at the level of 1.25 Å RMSD are now available for 12 à. a loops using methods that scale exponentially with system size.[48] Recent advances even boast lower than 1 Å RMSD precision on loops of up to 12 a.a.[33] or as low as 2 Å for loops of up to 20 a.a. that are identifiable by sequence homology and other similarity criterions.[6]

In this article, we have shown that the ART nouveau method can be used to sample efficiently the conformation space of loops of 20 a.a. or more. In particular, ART nouveau is very competitive as compared with previously published methods on these large loops,[6,46] demonstrating an efficient sampling of a wide range of conformations and is also able to sample conformations of lower RMSD to the native structure. This advantage is likely due to the fact that events represent a physical trajectory with local minima connected through a common saddle point. Given that the conformation space increases exponentially with the loop length, large random moves are very likely to end up in unphysical parts of this space, something that is avoided with ART nouveau even with the relatively long steps used on the long loops. The trajectory we generate during the event, which attempts to follow a direction of negative curvature with all other 3N-1 directions near their minimum, ensures that.

By extensively sampling low-energy structures, ART nouveau can also provide useful information beyond the best score. Although the proteins that were chosen for this study have well-defined structures, it is interesting to note that our simulations sampled conformations of low energy and high RMSD to the global energy minimum for both the small and large loops as displayed in Figure 6(a, b), respectively. For the 8 a.a. loops, multiple conformations with RMSD up to 6 Å to the global energy minimum structures were found within less then 1 kcal/mol above the global energy minimum. For the 12 a.a. and 19–20 a.a. loops, this value reaches 7 and 9 Å, respectively, on a few occasions. On more flexible loop targets, these distant conformations may well be populated at the equilibrium, playing biological role for these structures.

As ART-nouveau is a method that can be used with any underlying energy potential, its ability to find global energy minimum would not be altered using a more faithful protein representation in which the global minima corresponds to the native structures. From the collected data on loops of size 8–20 a.a., we estimate the computational time requirements to scale roughly quadratically with the sequence length of the simulated loop.

The current adaptation of the ART-nouveau method is a promising tool to tackle the problem of long loop sampling. All the metrics presented show that the first 1500 sampled conformations are the most rewarding in their ability to minimize the RMSD to the global minimum and that it is preferable to launch multiple short simulation runs than a few long ones. Results from a few test cases with the 19–20 a.a. loops demonstrate that, by alternating large and smaller moves, ART nouveau can avoid being trapped into the numerous basins associated with the longer loops complex energy landscape, sampling the configurational space efficiently at rough and fine levels, leading to the identification of a number of competing states, slightly above the minimum-energy conformation, that could plan an important biological role.

ACKNOWLEDGMENTS

The authors are grateful to Calcul Québec for access to its computer infrastructure.

REFERENCES

1. Kryshtafovych A, Fidelis K. Protein structure prediction and model quality assessment. Drug Discov Today 2009;14:386–393.
2. Fernandez-Fuentes N, Querol E, Aviles FX, Sternberg MJE, Oliva B. Prediction of the conformation and geometry of loops in globular proteins: testing ArchDB, a structural classification of loops. Proteins-Struct Funct Bioinform 2005;60:746–757.
3. Levefelt C, Lundh D. A fold-recognition approach to loop modeling. J Mol Model 2006;12:125–139.
4. Peng HP, Yang AS. Modeling protein loops with knowledge-based prediction of sequence-structure alignment. Bioinformatics 2007; 23:2836–2842.
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. Nucleic Acids Res 2009;37:W571–W574.
6. Choi Y, Deane CM. FREAD revisited: accurate loop structure prediction using a database search algorithm. Proteins-Struct Funct Bioinform 2010;78:1431–1440.
7. Fiser A, Do RKG, Sali A. Modeling of loops in protein structures. Protein Sci 2000;9:1753–1773.
8. Xiang ZX, Soto CS, Honig B. Evaluating conformational free energies: the colony energy and its application to the problem of loop prediction. Proc Nat Acad Sci USA 2002;99:7432–7437.
9. Canutescu AA, Dunbrack RL. Cyclic coordinate descent: a robotics algorithm for protein loop closure. Protein Sci 2003;12: 963–972.
10. DePristo MA, de Bakker PIW, Lovell SC, Blundell TL. Ab initio construction of polypeptide fragments: efficient generation of accurate, representative ensembles. Proteins-Struct Funct Genet 2003; 51:41–55.
11. 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-Struct Funct Genet 2003;51:21–40.
12. Coutsias EA, Seok C, Jacobson MP, Dill KA. A kinematic view of loop closure. J Comput Chem 2004;25:510–528.
13. Rohl CA, Strauss CEM, Chivian D, Baker D. Modeling structurally variable regions in homologous proteins with rosetta. Proteins-Struct Funct Bioinform 2004;55:656–677.
14. Zhu K, Shirts MR, Friesner RA. Improved methods for side chain and loop predictions via the protein local optimization program: variable dielectric model for implicitly improving the treatment of polarization effects. J Chem Theory Comput 2007;3, 2108–2119.
15. Sellers BD, Zhu K, Zhao S, Friesner RA, Jacobson MP. Toward better refinement of comparative models: predicting loops in inexact environments. Proteins-Struct Funct Bioinform 2008;72, 959–971.
16. Olson MA, Feig M, Brooks CL. Prediction of protein loop conformations using multiscale Modeling methods with physical energy scoring functions. J Comput Chem 2008;29:820–831.
17. Deane CM, Blundell TL. CODA: A combined algorithm for predicting the structurally variable regions of protein models. Protein Sci 2001;10:599–612.
18. Tosatto SCE, Bindewald E, Hesser J, Manner R. A divide and conquer approach to fast loop modeling. Protein Eng 2002;15:279–286.
19. Arnautova YA, Abagyan RA, Totrov M. Development of a new physics-based internal coordinate mechanics force field and its application to protein loop modeling. Proteins-Struct Funct Bioinform 2011;79:477–498.
20. Malek R, Mousseau N. Dynamics of lennard-jones clusters: A characterization of the activation-relaxation technique. Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics 2000;62:7723–7728.
21. St-Pierre JF, Mousseau N, Derreumaux P. The complex folding pathways of protein A suggest a multiple-funnelled energy landscape. J Chem Phys 2008;128:045101–045108.
22. Dong X, Chen W, Mousseau N, Derreumaux P. Energy landscapes of the monomer and dimer of the Alzheimer's peptide A beta(1–28). J Chem Phys 2008,128–137.
23. Wei GH, Mousseau N, Derreumaux P. Exploring the energy landscape of proteins: a characterization of the activation-relaxation technique. J Chem Phys 2002;117:11379–11387.
24. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equation of state calculations by fast computing machines. J Chem Phys 1953;21:1087–1092.
25. Machado-Charry E, Beland LK, Caliste D, Genovese L, Deutsch T, Mousseau N, Pochet P. Optimized energy landscape exploration using the ab initio based activation-relaxation technique. J Chem Phys 2011;135:034102–034112.
26. Marinica MC, Willaime F, Mousseau N. Energy landscape of small clusters of self-interstitial dumbbells in iron. Phys Rev B 2011;83:094119–094132.
27. Kallel H, Mousseau N, Schiettekatte F. Evolution of the potential-energy surface of amorphous silicon. Phys Rev Lett 2010;105: 045503–045506.
28. Berendsen HJC, Postma JPM, Gunsteren WFv, DiNola A, Haak JR. Molecular dynamics with coupling to an external bath. J Chem Phys 1984;81:3684–3690.
29. Jacobson MP, Pincus DL, Rapp CS, Day TJF, Honig B, Shaw DE, Friesner RA. A hierarchical approach to all-atom protein loop prediction. Proteins-Struct Funct Bioinform 2004;55:351–367.
30. Lee DS, Seok C, Lee J. Protein loop modeling using fragment assembly. J Korean Phys Soc 2008;52:1137–1142.
31. Canutescu AA, Shelenkov AA, Dunbrack RL. A graph-theory algorithm for rapid protein side-chain prediction. Protein Sci 2003; 12:2001–2014.
32. Spassov VZ, Flook PK, Yan L. LOOPER: a molecular mechanics-based algorithm for protein loop prediction. Protein Eng Des Select 2008;21:91–100.
33. Mandell DJ, Coutsias EA, Kortemme T. Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. Nat Methods 2009;6:551–552.
34. Lee J, Lee D, Park H, Coutsias EA, Seok C. Protein loop modeling by using fragment assembly and analytical loop closure. Proteins-Struct Funct Bioinform 2010;78:3428–3436.
35. Zhou H, Zhou Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci 2002;11:2714–26.
36. Krivov GG, Shapovalov MV, Dunbrack, JRL. Improved prediction of protein side-chain conformations with SCWRL4. Proteins 2009; 77:778–95.
37. Wang GL, Dunbrack RL. PISCES: a protein sequence culling server. Bioinformatics 2003;19:1589–1591.
38. Kabsch W, Sander C. Dictionary of protein secondary structure – pattern–recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
39. Derreumaux P. From polypeptide sequences to structures using Monte Carlo simulations and an optimized potential. J Chem Phys 1999;111:2301–2310.
40. Cotè S, Derreumaux P, Mousseau N. Distinct morphologies for amyloid beta protein monomer: A beta(1–40), A beta(1–42), and A beta(1–40)(D23N). J Chem Theory Comput 2011;7:2584–2592.
41. Santini S, Wei G, Mousseau N, Derreumaux P. Pathway complexity of Alzheimer's beta-amyloid Abeta16–22 peptide assembly. Structure 2004;12:1245–1255.

42. Chen W, Mousseau N, Derreumaux P. The conformations of the amyloid-beta (21–30) fragment can be described by three families in solution. J Chem Phys 2006;125:084911.
43. Nasica-Labouze J, Meli M, Derreumaux P, Colombo G, Mousseau N. A multiscale approach to characterize the early aggregation steps of the amyloid-forming peptide GNNQQNY from the yeast prion Sup-35. PLOS Comput Biol 2011;7:e1002051.
44. Barducci A, Bonomi M, Derreumaux P. Assessing the quality of the OPEP coarse-grained force field. J Chem Theory Comput 2011;7, 1928–1934.
45. Zhang C, Liu S, Zhou YQ. Accurate and efficient loop selections by the DFIRE-based all-atom statistical potential. Protein Sci 2004;13: 391–399.
46. Jamroz M, Kolinski A. Modeling of loops in proteins: a multimethod approach. BMC Struct Biol 2010;10:5–13.
47. Gronau I, Moran S. Optimal implementations of UPGMA and other common clustering algorithms. Inform Process Lett 2007;104:205–210.
48. Zhu K, Pincus DL, Zhao SW, Friesner RA. Long loop prediction using the protein local optimization program. Proteins-Struct Funct Bioinform 2006;65:438–452.