A Population-based Approach for Diversified Protein Loop Structure Sampling

Yaohang Li

Department of Computer Science
North Carolina A&T State University
Greensboro, NC 27411
yaohang@ncat.edu

Abstract. Protein loop structure modeling is regarded as a mini protein folding problem with significant scientific importance. Efficiently sampling the loop conformation space is a key step to computationally obtain accurate loop structure models. Due to the large size of the conformation space and the complication of the scoring functions describing protein energy, it is difficult to obtain broad, diverse coverage of the loop conformations with low energy (score). In this article, we present a new population-based approach to sample the backbone conformations of protein loops. The main advantage of the population-based approaches is that various selection schemes can be applied to enforce the conformations in a population to satisfy certain constraints. In our sampling approach, conformations are generated in the dihedral angles (ϕ,ψ)-space and the Differential Evolution (DE) method is employed to implement dihedral angle crossover for generating new conformations. A diversity selection scheme is applied to achieve diversified sampling. Using a narrowing gap selection scheme, decoys satisfying loop closure condition are obtained by gradually eliminating conformations with large terminal gaps in a population. Our computational results on modeling long loop targets have shown diverse and broad coverage of the loop conformation space, which leads to consistently reaching the native-like decoys in the sampling process.

1. Introduction

Protein loop structure modeling is important in structural biology for its wide applications, including determining the surface loop regions in homology modeling [1], defining segments in NMR spectroscopy experiments [2], designing antibody [3], and modeling ion channel [4]. The value of computer-generated protein loop models in biological research and practice relies critically on their accuracy. Protein loop structure modeling can be considered as a mini version of the *ab initio* protein folding problem. Despite their short length, protein loops exhibit greater structural flexibility than strands and helices and have few contacts with the remainder of the protein, which make it more difficult to predict than the geometrically regular β-strands and α-helices [5]. Currently, development of high-resolution computational approaches that can reliably produce accurate protein loop models, particularly in long loop targets, remains an unsolved problem. The main difficulties include the large protein loop

conformation space as well as the complicated landscape of the scoring functions describing the loop energy.

Similar to *ab initio* protein folding, the rationale of *ab initio* protein loop structure modeling is to optimize a protein loop energy function to discover the native-like conformations [7]. Typically, in protein modeling, the physics-based energy functions yield a rugged, funnel-like energy landscape, which can easily trap the optimization process and is extremely difficult to search. Several approaches, such as replacing the van der Waals potential with a soft-sphere potential [8], switching to a statistics-based term [9], etc., have been developed to produce scoring functions with "softened" energy landscape to facilitate the search process. However, such scoring functions also come with insensitivity and potentially inaccuracy, i.e., a conformation with the absolutely lowest score may not be a native-like conformation while a conformation with a relatively higher score may in fact be a more reasonable structure than the one with a lower score. Therefore, it is well-known that an optimization method seeking the very global minimum of a scoring function is usually not effective in finding the true native conformation. Instead, a sampling approach that can efficiently explore the low score regions in the scoring function landscape is more desirable [10].

For very short protein loop targets, one may be able to traverse the discretized dihedral angles (ϕ, ψ)-space to completely sample all possible conformations. However, for longer protein loops, the size of the conformation space grows exponentially where complete sampling becomes infeasible. Markov Chain Monte Carlo [6, 11] and genetic algorithms [12] have been applied to sample the loop conformation space to discover feasible structures with low scores (energy). The existing problems in these sampling methods include oversampling – the same conformations are repeatedly generated as well as undersampling – some conformations with low scores are not reachable during the sampling procedure. Oversampling will lead to wasted computational efforts while more seriously, undersampling may miss good, native-like conformations.

In this article, we present a population-based sampling algorithm to achieve broad exploration of protein loop backbone conformation space. We use the backbone dihedral angles (ϕ,ψ) array as a reduced representation of a loop conformation. A modified Differential Evolution (DE) scheme [13] is used to crossover dihedral angles of selected conformations in an old population to generate new conformations in (ϕ,ψ)-space. A diversity selection scheme is developed to filter conformations in a population similar to those already generated during the sampling procedure, which favors the sampling process to explore undiscovered conformations with low scores and thus reduces the chance of repeatedly generating decoys with similar structures. By gradually eliminating the conformations in a population not satisfying the loop closure condition, our narrowing gap selection scheme can also lead to decoys with loop closure satisfaction. We verify our sampling approach by applying it to the long targets provided in Jacobson's protein loop benchmark [14].

The remainder of the article is organized as follows. Sections 2 and 3 describe the general protein loop structure modeling procedure and our population-based sampling method, respectively. Section 4 shows our computational results on the 12-residue loop benchmark targets. Section 5 summarizes our conclusions and future research directions.

2. General Protein Loop Structure Modeling Paradigm

The *ab initio* protein loop structure modeling procedure [15, 16, 17] typically involves the phases of sampling, filtering, clustering, and refining, although some additional steps may be employed in different programs. Figure 1 shows a conceptual illustration of these phases.

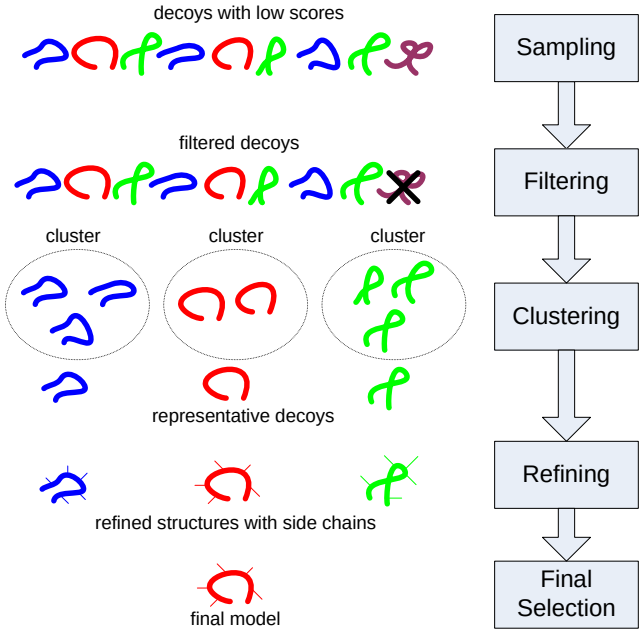

Fig. 1. Typical steps in high resolution *ab initio* protein loop structure modeling

In the sampling phase, the loop conformation space is explored and decoys with low scores are produced. In order to reduce the degree of freedom, usually only loop backbone with reduced representation are used in this phase with simplified, smooth scoring functions. Afterward, the infeasible, bad decoys will be eliminated in the filtering phase. Then, in the clustering phase, decoys with similar structures will be grouped into a cluster and representative decoys for each cluster will be selected. Next, in the refining phase, side chains are added and complicated all-atom energy functions are used to locally optimize the representative decoys. Finally, the refined representative decoy with the lowest energy will be selected as the predicted model.

Broadly sampling the loop conformation space to generate low-score decoys with diverse structures in the sampling phase is critical in successfully predicting high-resolution protein loop models. This is due to the fact that if a native-like decoy is not reachable in loop conformation sampling, it is unlikely to obtain a high-resolution model close to the native structure in the refining phase. For the population-based sampling approach described in this article, we only consider the modeling computation in the sampling phase.

3. Population-based Diversified Sampling Approach

We develop a population-based sampling approach which intends to diversely sample the loop conformation space. Initially, a population with N conformations, C_1, …, C_N, is randomly generated. Each loop structure conformation C_i with n residues is represented by a vector (θ_1, …, θ_{2n}), which represents the backbone dihedral angles of (ϕ_1, ψ_1, …, ϕ_n, ψ_n). The dihedral angles of ω_i are kept constants at their average value of 180°. A statistical distance-based atom pair-wise scoring function is used as the sampling scoring function [18]. When scoring function evaluation or structure comparison is needed, the dihedral angles representation of C_i is converted to the backbone atom representation. We adopt the Differential Evolution (DE) [13] approach to produce new conformations for the next population. A diversity selection scheme is designed to achieve diversified sampling and a narrowing gap selection scheme is used to guarantee loop closure.

3.1 Differential Evolution for Conformation Crossover

DE [13] is a powerful computational method for continuous function optimization, which has demonstrated its effectiveness on several hard optimization problems with complicated objective functions [22]. In our loop sampling approach, DE is used to crossover old conformations to produce new ones in continuous dihedral angles space. For each loop conformation C_i, a mutant vector V_i is formed by

$$V_i = C_{r1} + F(C_{r2} - C_{r3}), \quad \textbf{(1)}$$

where *r1*, *r2*, and *r3* are mutually distinct, uniformly distributed integer random numbers in the interval [1, N] and $F > 0$ is a tunable amplification control constant as described in [4]. Then, a new conformation $C_i'(\theta_1', \ldots, \theta_{2n}')$ is generated by the crossover operation on V_i and C_i:

$$\theta_j' = \begin{cases} v_j & j = \langle s \rangle_{2n}, \langle s+1 \rangle_{2n}, \ldots, \langle s+L-1 \rangle_{2n} \\ \theta_j & otherwise \end{cases} \quad \textbf{(2)}$$

where $\langle \cdot \rangle_{2n}$ denotes the modulo operation with modulus $2n$, s is a randomly generated integer from the interval [0, $2n$-1], L is an integer drawn from [0, $2n$-1] with probability $\Pr(L = k) = (\mathrm{CR})^L$, and $\mathrm{CR} \in [0, 1)$ is the crossover probability. Practical advice suggests that CR = 0.9 and F = 0.8 are favorable choices in the DE scheme [13], which is also adopted in our program. Our slight modification to the DE scheme is to always keep θ_i in the ranged of $[-\pi,\pi]$.

3.2 Diversity Selection Scheme

The diversity selection scheme encodes the capability of enforcing the conformations in a population to satisfy the diversified sampling requirement. Our

diversity selection scheme is based on the similarity of a conformation s to a given set of generated decoys D = $\{d_1, \ldots, d_k\}$, which is measured by

$$S(s) = \min_i dist(s, d_i) \quad (3)$$

where *dist*(.) is a distance function measuring the structural difference.

In our implementation, we keep track of the already generated decoys $d_1, \ldots, d_k$ by recording their Cα atoms in an array. To reduce computation time of evaluating loop structure similarity, instead of calculating the Root Mean Square Deviation (RMSD) of all backbone atoms, we use the Cα RMSD between conformations in a population and the generated decoys as the distance function. Then, in diversity selection, all conformations in the current population are sorted according to their similarity to the generated decoys and the top μ% of the candidates are eliminated, where μ is a tunable constant.

3.3 Narrowing Gap Selection Scheme for Loop Closure

The so-called loop closure problem is defined as follows: given the N- and C-terminals, find a loop backbone conformation of a certain length that can bridge the ends seamlessly [19]. Inverse kinematics [23] is a common method to solve the loop closure problem. Unfortunately, inverse kinematics has difficulty to be applied to our population-based sampling approach because crossing over the dihedral angles of two or more conformations satisfying the loop closure condition does not automatically guarantee loop closure in the new conformation.

In our population-based sampling approach, we develop a narrowing gap selection scheme to produce decoys satisfying the loop closure condition. We fix the position of the N-terminal, produce the loop based on the dihedral angle values in loop conformation C_i, and then calculate the gap distance, $G(C_i)$, from the C-terminal in the generated loop to the target C-terminal. $G(C_i)$ is then used to measure the loop closure gap. To produce loops closely approximating the loop closure condition, in the gap selection scheme, we eliminate conformations C_i where $G(C_i) > \delta$. Here δ is a variable, which specifies the acceptable gap between the predicted C-terminal and the target C-terminal. At the beginning, δ is initialized to a large value to allow aggressive loop conformation sampling. The value of δ is decreasing in every iteration toward a small value so as to gradually eliminate conformations with gaps larger than δ and eventually lead to conformations approximately satisfying the loop closure condition. When the final conformation with the lowest score is selected to output as a decoy, the C-terminal gap can continue to be reduced by slightly adjusting the dihedral angles of ϕ_i, ψ_i, and ω_i in the loop structure.

3.4 Algorithm Description

By putting every piece of the puzzle together, the descriptive pseudo code of the population-based diversified sampling algorithm is described as follows. The algorithm can be repeatedly executed to produce multiple loop decoy structures.

```
Initialize N conformations, C1, …, CN, randomly and initialize δ
Repeat {
    Generate M new conformations, C1', …, CM', based on the previous
       population's N conformations using DE
    Run diversity selection scheme to eliminate conformations close to
       the already generated decoys stored in the decoy array
    Run gap selection scheme to eliminate conformations that G(C1'>δ)
    Evaluate the remaining C1', …, CM' use scoring function f(.)
    Replace C1, …, CN with top N conformations in C1, …, CN and the
       remaining C1', …, CM'
    Reduce δ
} Until convergence or reaching the expected iteration number
Produce the decoy in the current population with the lowest score
If there is serious steric clash or large loop closure gap
    Discard this decoy
Else {
    Save Cα atoms to the generated decoy array
    Minimize the terminal gap of by slightly adjusting the dihedral
       angles of •1, •2, and •3
    Output the loop decoy}
```

4. Computational Results

We applied our methods to the long loop benchmark targets specified in [14], including 17 12-residue, 35 11-residue, and 49 10-residue loops. Due to space restrictions, we can only report a fraction of our results in this article. Therefore, we use our computational results on 12-residue loop targets to illustrate the effectiveness of our population-based diversified sampling scheme. Our computations on the other targets actually yield similar results.

We use the path length of the Minimum Spanning Tree (MST) [20] based on the pair-wised RMSD matrix of the generated decoys to measure sampling diversity. Table 1 shows the comparison of the MST path lengths of the 1,000 decoys generated by our population-based loop conformation sampling algorithm with and without the diversity selection scheme. It is important to notice that the diversity selection scheme plays an important role in the sampling process, which leads to significantly larger MST path length in all 12-residue loop targets when the diversity selection scheme is employed. In other words, the decoys generated using the diversity selection scheme are more structurally different from each other than those without using the diversity selection scheme. This indicates that the sampling process with the diversity selection scheme has a broader coverage of the loop structure conformation space and leads to decoys with more diversified representation of structures.

The diversified sampling of the loop conformation space directly improves the chance of generating decoys with close structure to the native one. Table 1 also compares the best decoys with the smallest backbone RMSD to the corresponding native structure generated with and without using the diversity selection scheme in 12-residue targets. One can find that in all loop targets except for 5nul(54:65), the population-based sampling process with the diversity selection scheme can consistently reach decoys with backbone RMSD less than 2A, which is within the experimental X-ray crystallization resolution. In contrast, sampling without the diversity selection scheme cannot reach decoys with RMSD less than 2A in 5 out of

the 17 targets. Moreover, there is averagely 0.37A RMSD shift in the best decoys of the 12-residue targets when the diversity selection scheme is used.

Our further structural analysis shows that the native 5nul loop (54:65) interacts with a flavin mononucleotide ligand. The distance-based scoring function used in our sampling program makes no assumption on any ligands. This explains why in 5nul(54:65) no decoys with RMSD under 2A are generated in our sampling approach even when the diversity selection scheme is used.

Table 1. MST path length of the pair-wise RMSD matrix and the best decoy with the smallest backbone RMSD of the 1,000 decoys generated in our population-based sampling with and without the diversity selection scheme in 12-residue loop targets

			With Diversity Selection Scheme		Without Diversity Selection Scheme	
Protein	Start Res.	End Res.	MST Path Length (A)	RMSD (A) of the Best Decoy	MST Path Length (A)	RMSD (A) of the Best Decoy
1ixh	160	171	1436	1.519	1297	2.786
1cex	40	51	1294	1.780	1201	2.265
5pti	36	47	1389	1.610	1295	1.844
1rge	57	68	1178	1.005	1132	1.403
1arb	74	85	1211	1.376	1115	1.500
7rsa	13	24	1343	1.509	1207	2.102
1xyz	813	824	1281	1.510	1151	1.687
1cyo	32	43	1368	1.341	1243	1.444
1akz	181	192	1370	1.197	1255	1.912
153l	98	109	1399	1.791	1308	2.245
1bkf	9	20	1296	1.102	1206	1.443
1dad	204	215	1382	1.423	1270	1.717
1dim	213	224	1262	1.109	1190	1.624
1kuh	90	101	1371	1.214	1258	1.188
2ayh	21	32	1392	1.678	1245	1.540
351c	15	26	1383	1.914	1271	1.612
5nul	54	65	1341	2.141	1235	3.218
Average			1335	1.483	1228	1.855
Standard Deviation			71	0.312	57	0.529

Due to the broad structure representations in the generated decoys, sampling with the diversity selection scheme will also lead to diversified clusters and representative decoys in the clustering phase of protein loop structure modeling. Figures 2 and 3 show the representative decoys in clustering the 1,000 decoys in loop target 1akz(181:192) using sampling with and without the diversity selection scheme, respectively. We use a simple agglomerative clustering algorithm [21] with 2.0A cutoff. One representative decoy is selected for each cluster. For the 1,000 decoys generated without using the diversity selection scheme, 9 clusters are produced. For each representative one of the 9 clusters, a similar structure can be found in the

representative decoys from the 19 clusters generated by sampling using the diversity scheme. However, several representative decoys exhibiting significantly different structures found in sampling using the diversity selection scheme are not presented in these 9 clusters generated by sampling without diversity selection scheme, including a native-like one with 1.25A RMSD.

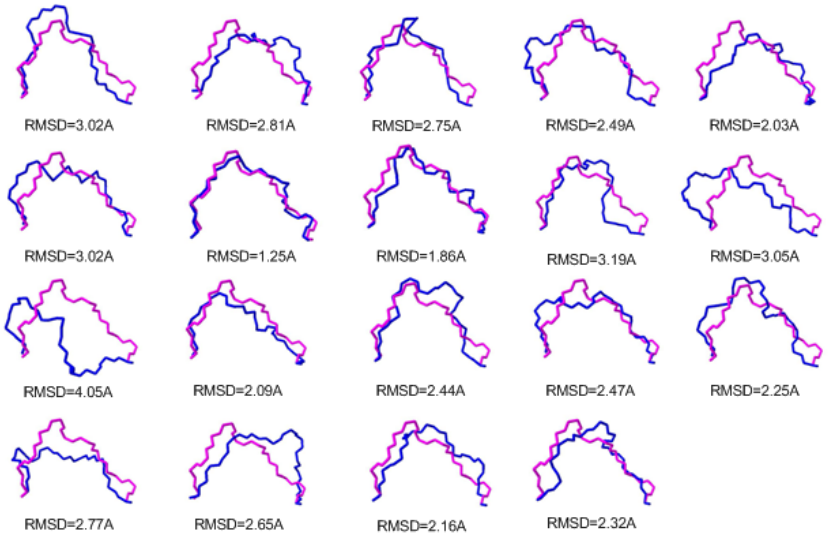

Fig. 2. Representative decoys in clustering the 1,000 decoys generated with diversity selection scheme in loop target 1akz(181:192). (purple – native conformation, blue – decoy conformation)

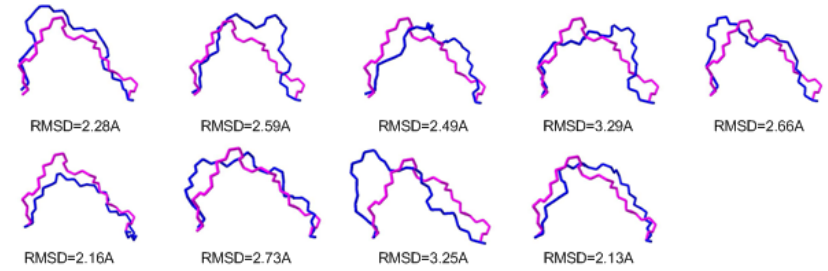

Fig. 3. Representative decoys in clustering the 1,000 decoys generated without diversity selection scheme in loop target 1akz(181:192). (purple – native conformation, blue – decoy conformation)

A minor disadvantage of the diversified sampling method is that it may also increase the production of "bad" decoys. This is due to the fact that the diversity selection scheme will increase the chance of discovering not only the "good", native-like conformations but also the "bad", far-deviated ones yielding low scores. As an example depicted in Figure 4 showing the RMSD distribution of the 1,000 decoys in loop target 1rge(57:68), sampling with the diversity selection scheme leads to larger population of decoys with RMSD less than 2.0A as well as those with RMSD higher

than 3.0A than sampling without diversity selection scheme. This problem can be relatively easy to address in the filtering phase of loop structure modeling – when a high-resolution, all-atom scoring function is employed, the "bad", far-deviated decoys can usually be identified and then eliminated.

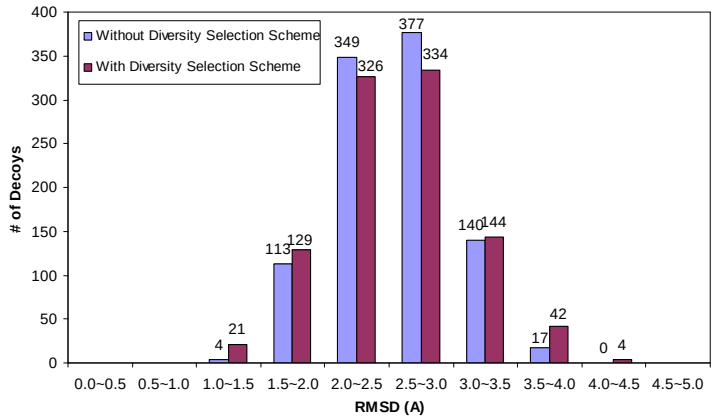

Fig. 4. RMSD distribution of the 1,000 decoys generated in sampling with the diversity selection scheme and without the diversity selection scheme in loop target 1rge(57:68)

5. Conclusions and Future Research Directions

In this article, we present a population-based sampling algorithm for diversified sampling protein loop backbone conformations. A diversity selection scheme is designed to diversify predicted decoys and a narrowing gap selection scheme is used to achieve loop closure condition satisfaction. Our computational results on 12-residue protein loop benchmark targets have shown diversified decoy structure distributions and improved chance of reaching native-like conformations.

It is important to notice that our approach only targets the backbone sampling phase in *ab initio* protein loop structure modeling and the decoy generation time is usually less than a minute. As a result, our decoys have relatively lower quality compared to the all-atom modeling method such as PLOP [15], which typically takes days to deliver a model. In the future, we are interested in studying how diversified backbone sampling can benefit all-atom simulation in high-resolution loop modeling.

Acknowledgement

The work is partially supported by NSF under grant number CCF-0829382.

References

1. Bruccoleri, R.E.: *Ab initio* loop modeling and its application to homology modeling. Methods in Molecular Biology, 143, 247-264 (2000).
2. Dmitriev, O.Y., Fillingame, R.H.: The rigid connecting loop stabilizes hairpin folding of the two helices of the ATP synthase subunit c. Protein Science, 16(10), 2118-2122 (2007).
3. Martin, A.C., Cheetham, J.C., Rees, A.R.: Modeling antibody hypervariable loops: a combined algorithm. PNAS, 86(23), 9268-9272 (1989).
4. Tasneem, A., Iyer, L.M., Jakobsson, E., Aravind, L.: Identification of the prokaryotic ligand-gated ion channels and their implications for the mechanisms and origins of animal Cys-loop ion channels. Genome Biol., 6(1) R4 (2005).
5. Monnigmann, M., Floudas, C.A.: Protein loop structure prediction with flexible stem geometries. Proteins: Structure, Function, and Bioinformatics, 61(4), 748-762 (2005).
6. Cui, M., Mezei, M., Osman, R.: Prediction of protein loop structures using a local move Monte Carlo approach and a grid based force field. Protein Engineering Design and Selection, 21(12), 729-735 (2008).
7. Anfinsen, C.B.: Principles that Govern the Folding of Protein Chains. Science, 181, 223-230 (1973).
8. Zhang, H., Lai, L., Han, Y., Tang, Y.: A Fast and Efficient Program for Modeling Protein Loops. Biopolymers, 41, 61-72 (1997).
9. van Vlijmen, H.W.T., Karplus, M.: PDB-based protein loop prediction: parameters for selection and methods for optimization. J. Mol. Biol., 289, 1469-1490 (1999).
10. Li, Y., Bordner, A.J., Tian, Y., Tao, X., Gorin, A.: Extensive Exploration of the Conformational Space Improves Rosetta Results for Short Protein Domains. In: 7th International Conference on Computational Systems Bioinformatics (2008).
11. Liu, Z., Mao, F., Li, W., Han, Y., Lai, L.: Calculation of protein surface loops using Monte Carlo simulated annealing simulation. Journal of Molecular Modeling, 6(1), 1-8 (2000).
12. McGarrah, D.B., Judson, R.S.: Analysis of the genetic algorithm method of molecular conformation determination. Journal of Computational Chemistry, 14, 1385-1395 (1993).
13. Storn, R., Price, K.: Differential Evolution – A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces. J. of Global Optimization, 11(4), 341–359 (1997).
14. Jacobson, M.P., Pincus, D.L., Rapp, C.S., Day, T.J.F., Honig, B., Shaw, D.E., Friesner, R.A.: A hierarchical approach to all-atom protein loop prediction. Proteins, 55(2), 351-367 (2004).
15. Zhu, K., Pincus, D.L., Zhao, S., Friesner, R.A.: Long loop prediction using the protein local optimization program. Proteins, 65, 438-452 (2006).
16. Rohl, C.A., Strauss, C.E., Chivian, D., Baker, D.: Modeling structurally variable regions in homologous proteins with rosetta. Proteins, 55, 656-677 (2004).
17. de Bakker, P.I.W., Depristo, M.A., Burke, D.F., Blundell, T.L.: *Ab initio* construction of polypeptide fragments. Proteins, 51, 21-40 (2002).
18. Rojnuckarin, A., Subramaniam, S.: Knowledge-based interaction potentials for proteins. Proteins: Structure, Function, and Genetics, 36, 54-67 (1999).
19. Canutescu, A.A., Dunbrack, R.L.: Cyclic Coordinate Descent: A Robotics Algorithm for Protein Loop Closure. Protein Science, 12, 963-972 (2003).
20. Xu, Y., Olman, V., Xu, D.: Minimum Spanning Trees for Gene Expression Data Clustering. Genome Informatics, 12, 24-33 (2001).
21. Lance, G.N., Williams, W.T.: A general theory of classificatory sorting strategies. The Computer Journal, 9, 373-380 (1967).
22. Price, K., Storn, R., Lampinen, J.: Differential Evolution – A practical approach to global optimization. Springer (2005).
23. Kolodny, R., Guibas, L., Levitt, M., Koehl, P.: Inverse Kinematics in Biology: the Protein Loop Closure Problem. International Journal of Robotics Research, 24(3), 151-163 (2005).