Integrating Multiple Scoring Functions to Improve Protein Loop Structure Conformation Space Sampling

Yaohang Li, Ionel Rata, and Eric Jakobsson

Abstract — **In this article, we present a new protein structure modeling approach based on multi-scoring functions sampling. The rationale is to integrate multiple carefully-selected physics- or knowledge-based scoring functions to tolerate insensitivity and inaccuracy existing in an individual scoring function so as to improve protein structure modeling accuracy. We apply the multi-scoring function sampling approach to protein loop backbone structure modeling. Our computational results show that sampling the scoring function space of a physics-based soft-sphere potential function and a knowledge-based scoring function based on pairwise atoms distance has led to resolution improvement in the predicted decoy populations in a set of 12-residue benchmark loop targets.**

I. Introduction

ACCURATELY modeling protein or protein complex structures from protein sequences (*ab initio* protein modeling) is considered as one of the most significant grand challenges that have broad economic and scientific impacts. The theoretical foundation of *ab initio* protein modeling is the Anfinsen's thermodynamic hypothesis [1], which states that the native conformation is a unique, stable, and kinetically accessible minimum of the protein energy. Based on the Anfinsen's thermodynamic hypothesis, currently the *ab initio* protein modeling efforts are focused on globally optimizing a scoring function describing the protein energy to obtain models close to the native structures.

In practice, due to the difficulty of accurately calculating the energy of large protein or protein complex molecules, many scoring functions have been developed instead to approximate the protein energy. For example, the physics-based scoring functions are designed to estimate various physical interactions; the knowledge-based scoring functions are derived from the statistics of the experimentally-determined conformations within the protein data banks (PDB), and the regression-based scoring functions are developed to combine various physics- or knowledge-based scoring terms using the regression method. Moreover, simplification techniques by reducing protein representation are often used to "soften" or "simplify" the scoring function landscape to facilitate the global optimization process. These approximations or simplifications in scoring functions usually cause certain level of insensitivity or inaccuracy, which may lead to incorrect protein models.

In this article, we put forward a new approach for protein modeling by sampling multiple scoring functions with potentially different resolution. The multi-scoring functions sampling approach can tolerate the insensitivity in individual scoring function and thus increases the chance of discovering native-like, good conformations. We apply the multi-scoring functions sampling approach to protein loop backbone structure modeling by developing a population-based evolutionary algorithm to explore loop conformations using a scoring function based on pair-wise atoms distance and a soft-sphere potential function. Our computational results show RMSD (Root Mean Square Deviation) improvement in the decoys produced by the multi-scoring functions sampling algorithm, which agree with our theoretical analysis.

The rest of the article is organized as follows. The protein modeling scoring functions and their potential inaccuracy are analyzed in Section 2. In Section 3, we illustrate the insensitivity problem in current scoring functions. Section 4 provides the theoretical analysis of the multi-scoring functions sampling approach in protein modeling. We describe the implementation of the population-based evolutionary multi-scoring functions sampling algorithm for protein loop structure modeling and the computational results in Sections 5 and 6, respectively. Finally, Section 7 summarizes our conclusions and future research directions.

II. Protein Structure Modeling Scoring Functions and Their Potential Inaccuracy and Insensitivity

A. Physics-based Scoring Functions

Ideally, the protein energy would be evaluated with quantum mechanics, in which case the energy function could report the true energy. In computational practice, quantum mechanics is wildly intractable because of the size of protein molecules. As a compromise, artificial scoring functions (force fields) are developed based on classical physics to

Yaohang Li is with the Department of Computer Science, North Carolina A&T State University, Greensboro, 27411 USA (phone: 336-334-7245; fax: 336-334-7244; e-mail: yaohang@ncat.edu).

Ionel Rata is with the Center for Biophysics and Computational Biology, University of Illinois at Urbana-Champaign, Urbana-Champaign, IL 61801, USA (email: rata@uiuc.edu).

Eric Jakobsson is with the Department of Molecular and Integrative Physiology and National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign Urbana-Champaign, IL 61801, USA (e-mail: jakobsson@uiuc.edu).

978-1-4244-6766-2/10/ $26.00 © IEEE

approximate the true energy of molecular systems, such as CHARMM [2], AMBER [3], OPLS [4], and GROMOS [5]. Moreover, the physics-based scoring functions usually yield a rugged scoring function (energy) landscape, which is difficult to search. Simplification techniques such as reducing protein representation are often used to design a less rugged scoring function to facilitate the optimization process. These simplifications may also decrease the resolution of the physics-based scoring functions.

B. Knowledge-based Scoring Functions

Knowledge-based approaches evaluate the increasing number of experimentally determined conformations by statistical means to extract rules on preferred configurations and combinations. These rules are converted into "pseudo-potential" scoring functions for protein modeling. Compared with physics-based scoring functions, the knowledge-based potentials tend to be "softer" to tolerate structural imperfection – allowing better handling of the uncertainties and deficiencies of the computer-generated models [6]. The main problems of knowledge-based scoring functions are from their theoretical assumptions. Knowledge-based scoring functions derive their rules from experimental data typically by applying the inverse Boltzmann law, which is based on the following two important assumptions: 1) the set of known conformations are representative of proteins or protein complexes in general; and 2) the observed frequencies are independent of each other. The first assumption is questionable in most knowledge-based scoring functions – because compared to the unknown conformations, the known ones are an extremely small fraction [7, 8]. Thomas et al. [9] and Kocher et al. [10] also argued that inter-residue interactions are not independent. In consequence, all these may contribute to inaccuracy or insensitivity in knowledge-based scoring functions.

C. Regression-based (Weighted-Sum) Scoring Functions

To take advantage of the knowledge- and physics-based scores, a popular way is to build a linearly combined scoring function by adding up various empirical or physical terms with a particular set of weights. The regression-based scoring function is also called the weighted-sum scoring function. The weights are usually assigned by regression method [11, 12] – fitting predicted and experimentally determined models to a given set of training complexes. A major drawback of the regression method is its dependence on the size, composition and generality of the training set used to derive the weights. Moreover, it is also questionable that the weights are constants or the combinations are linear in general. An even more serious consequence of adding up scoring functions is the possible enlargement of the overall insensitivity regions and generation of more local minima.

III. SENSITIVITY OF SCORING FUNCTIONS

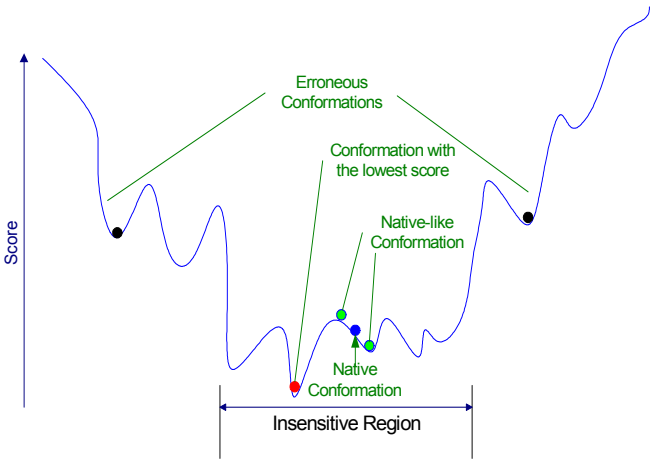
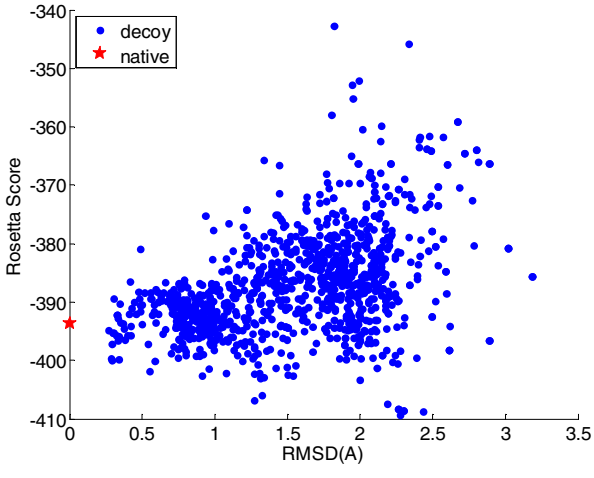
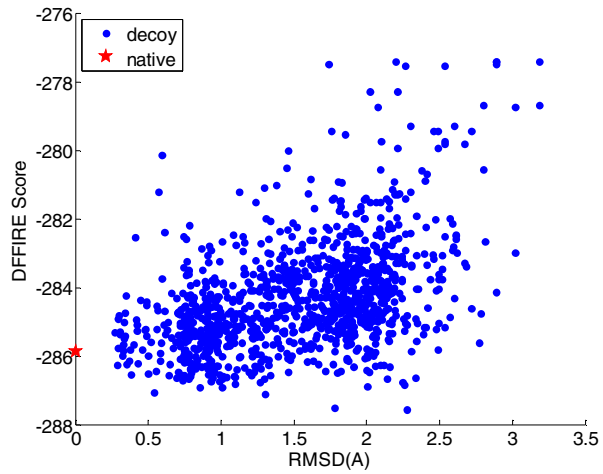


Figure 1: A Conceptual Illustration of Scoring Function Sensitivity

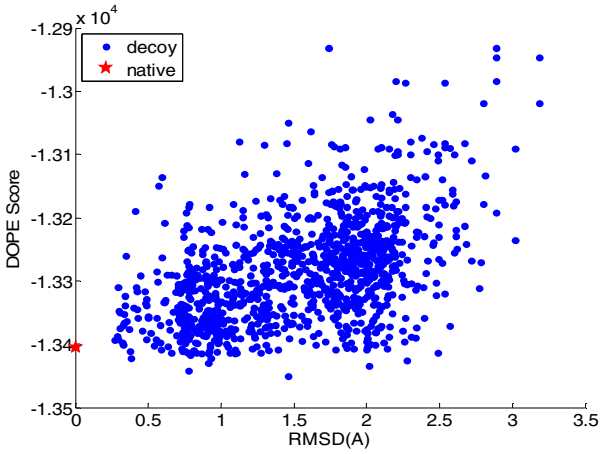
Currently, many existing physics-, knowledge-, or regression-based scoring functions for protein modeling have certain level of accuracy, which can distinguish a far deviated, erroneous conformation from a native-like, good conformation. When coupling with an effective optimization algorithm, a scoring function can usually drive the sampling process to some conformations relatively close to the native conformation. However, because the existing scoring functions are an approximation of the true energy function, when the score is low enough, the scoring functions' sensitivity decreases. That is, in low score regions, a conformation with a relatively higher score may in fact be a more reasonable structure than the one with a lower score. In practice, the native conformation usually does not exhibit the lowest score when it is put among the decoys generated by computer simulation [13]. Figure 1 shows the conceptual illustration of the scoring function insensitivity problem.



(a)



(b)



(c)

Figure 2: Scoring functions analysis for 1,081 loop decoys 90-101 in 1kuh. (a) Plot of Rosetta score vs. RMSD (b) Plot of DDFIRE score vs. RMSD (c) Plot of DOPE score vs. RMSD

Figure 2 illustrates an example of scoring function insensitivity in protein loop decoys in 1kuh(90:101) provided in Jacobson's loop decoys library [14] generated by comparative modeling, where the score-RMSD plots of 1,081 decoys using the Rosetta [15] full-atom scoring function, the dDFIRE (dipolar Distance-scaled, Finite-Ideal gas Reference) [16] scoring function, and the DOPE (Discrete Optimized Protein Energy) [17] scoring function are shown. The scores are obtained by evaluating the loop decoys together with the rest of the protein. In any one of these scoring functions, there are a small portion of decoys showing lower scores than the native conformation. Moreover, neither the native conformation nor the native-like decoys (within 0.5A RMSD from the native conformations) yield the lowest scores in these scoring functions. Such phenomenon can also be found in many other targets listed in Jacobson's loop decoys library.

Figure 3 shows the interpolated surface of the 1,081 decoys of 1kuh(90:101) represented by points within a function space composed of the Rosetta, dDFIRE, and DOPE scoring functions. The color of the surface represents the RMSD of the corresponding decoy. Although an individual scoring function may exhibit certain insensitivity, decoys that are very close to the native conformation can be found in the common low score areas of the scoring functions, as indicated in Figure 3. This strongly indicates that sampling the common low score areas of multiple scoring functions may improve the chance of reaching conformations with good quality and thereby improve the resolution of the predicted models.

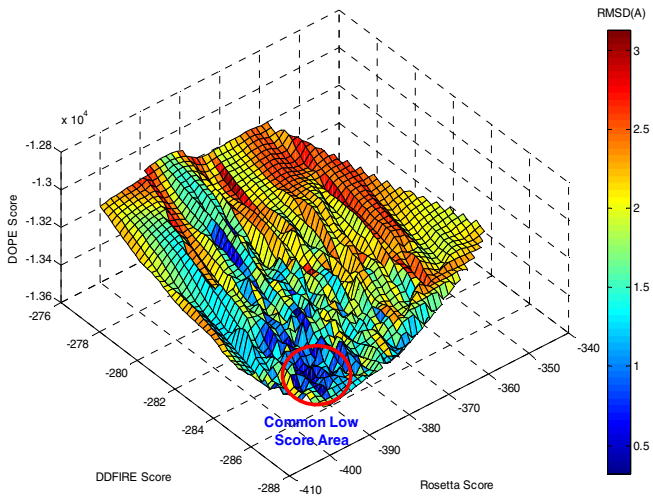


Figure 3: Common low score area reveals decoys in good quality in 1kuh(90:101)

IV. Sampling Multiple Scoring Functions

Due to the protein modeling scoring function insensitivity problem, an optimization process, which is looking for the absolute global minimum in the scoring function, may be misled to undesired conformations and ignore good ones with low but not the lowest scores. To handle this problem, we advocate a new protein structure modeling approach based on sampling multiple physics- or knowledge-based scoring functions instead of globally optimizing an individual scoring function. This is based on our assumption that when multiple good scoring functions are present, the correct, native-like conformations should satisfy most of them by yielding low scores while an incorrect conformation usually may yield low scores in some but not all scoring functions. This assumption is reasonable because most current "good" scoring functions for protein structure modeling do show certain level of accuracy. Efficiently sampling multiple scoring functions can increase the chance of reaching native-like, good conformation candidates, which will eventually lead to protein modeling accuracy improvement.

In multi-scoring functions sampling, the scoring functions $s_1(x)$, $s_2(x)$, …, $s_n(x)$, involved form a scoring function space,

$S(x) = [s_1(x), s_2(x), \ldots, s_n(x)]$, where $x \in C$ is a feasible protein conformation. A conformation $x^* \in C$ is Pareto optimal if and only if there is no $x \in C$ such that $s_i(x) \le s_i(x^*)$ for all $i \in \{1,2,...,n\}$, with at least one strict inequality. The set of x^* forms an optimal surface called the Pareto optimal front [18], which include conformations at the global minimum in each individual scoring function as well as non-dominated conformations in different combinations of scores.

Unlike global optimization, which looks for a conformation $x_{\min s_i} = \min(s_i(x))$ in a single scoring function $s_i(x)$, the goal of multi-scoring functions sampling in protein modeling is to explore the diverse conformations close to solution set x^* at the Pareto optimal front, which yield solution optimality in various combinations of the involved scoring functions. As a result, compared to globally optimizing a single scoring function, efficiently sampling multiple scoring functions to explore the conformations at or near the Pareto optimal front will lead to broader exploration of the protein conformation space, including not only the ones best satisfying individual scoring function but also those satisfying most scoring functions by yielding low scores. This tolerates insensitivity in individual scoring function and thus increases the chance of discovering native-like, good conformations.

The protein energy function has a well-known rugged, "funnel-like" landscape with large number of deep local minima hierarchically disposed [19]. In protein modeling using global optimization, an optimization process may be trapped in a deep local minimum and cannot escape in practical time. This phenomenon is referred to as the "waiting-time dilemma." [20] Compared to global optimization, multi-scoring functions sampling has a faster barrier-crossing capability. This is due to the fact that when multiple scoring functions are involved, a deep local minimum in one scoring function may not be a local minimum in another. When a population-based operation such as genetic-algorithm-style crossover or replica exchange [21] is applied, a conformation in a deep local minimum of a scoring function may be switched to another scoring function where its score is no longer a local minimum. This will increase the chance of jumping out of deep local minima in a scoring function. Figure 4 depicts a scenario that a replica exchange between two scoring functions helps escape deep local minima. Taking advantage of multiple available scoring functions to help the sampling process escape the deep local minima in one scoring function are also discussed in [27].

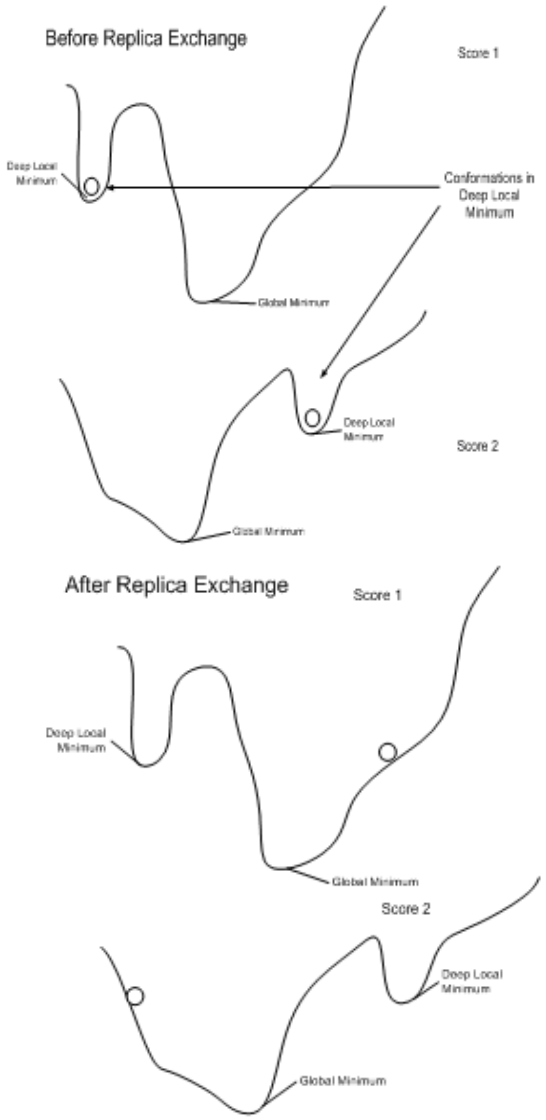


Figure 4: Replica exchange helps simulation process to escape deep local minima

Compared to optimizing a weighted-sum scoring function that combines multiple scoring terms, the multi-scoring functions sampling does not require estimation of weights. More importantly, multi-scoring functions sampling has potentially broader exploration of the protein loop conformation space with low scores than weighted-sum scoring function optimization. This is due to the fact that the weighted sum approach cannot find certain Pareto optimal conformations in the case of a concave Pareto optimal front [18].

The minor drawback of multi-scoring functions sampling is its computational cost, which is higher than optimizing a single scoring function due to the requirement of evaluating multiple scoring functions.

V. Multi-scoring Functions Sampling vs. Sampling a Weighted-Sum Scoring Function

Optimizing a weighted-sum scoring (energy) function is a popular approach in protein structure modeling to combine multiple scores. The weighted-sum scoring function is built by linearly combining various knowledge- or physics-based scoring terms by a particular set of weights, which are usually derived by the regression methods. Our analysis shows, theoretically, multi-scoring functions sampling has major advantages compared to weighted-sum scoring function optimization.

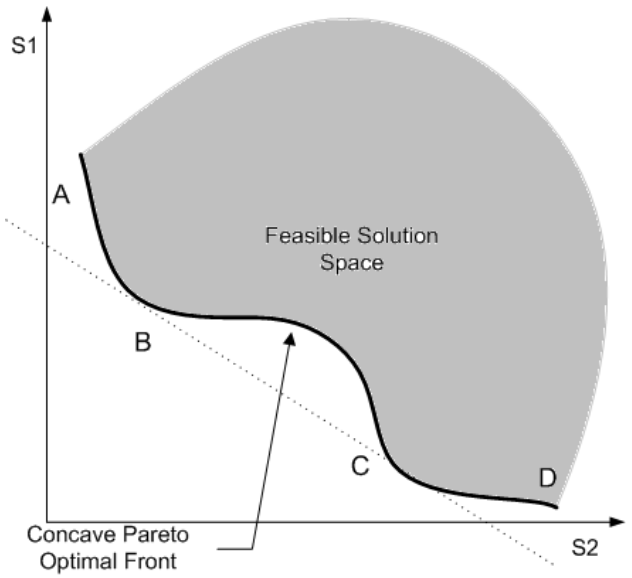


Figure 5: Scenario of a concave Pareto optimal front where a weighted-sum approach will fail to find some Pareto-optimal solutions

First of all, multi-scoring functions sampling has potentially broader exploration of the protein loop conformation space with low scores than weighted-sum scoring function optimization. This is due to the fact that the weighted sum approach cannot find certain Pareto optimal conformations in the case of a concave Pareto optimal front [20]. Figure 5 shows a conceptual scenario of a concave search space of two hypothetical scoring functions S1 and S2. When a set of weights are selected, a contour line is formed and the minimum solution of the weighted sum function corresponds to a solution on the Pareto optimal front, which is a tangent point of the contour line and the solution space. However, there exists no contour line that can produce a tangent point with the feasible solution space in the region BC in the Pareto optimal front. This is because before a tangent point is reached in BC, the contour line becomes a tangent at another point at AB or CD, which yields a lower weighted-sum function value. In other words, in weighted-sum function optimization, solutions in AB or CD will attract the optimization process to drive away from solutions in BC. The concave Pareto optimal front may still exist even a nonlinear function is used to combined various terms. In contrast, an efficient multi-scoring functions sampling approach can produce solutions at a concave Pareto optimal front, which thus leads to broader exploration of potentially good conformations in the protein conformation space.

Secondly, multi-scoring functions sampling can avoid problem due to scoring functions overlapping. Another problem of weighted-sum scoring function is its potentially over-counting of the same interaction scoring terms. This is due to terms, particularly the statistical ones, used in the weighted-sum scoring function that may have a common component. Without loss of generality, let us consider two scoring functions $S_1 = s_1 + s$ and $S_2 = s_2 + s$, where s is a common component and s_1 and s_2 are complementary components for S_1 and S_2, respectively. A weighted-sum approach using S_1 and S_2 will produce a scoring function $S = w_1S_1 + w_2S_2 = w_1s_1 + w_2s_2 + (w_1 + w_2)s$, where the common component s is over-counted. In contrast, multi-scoring functions sampling can address the over-counting issue naturally. Sampling multiple scoring functions S_1 and S_2 will lead to minimization of the common component s and sampling the Pareto optimal front of s_1 and s_2, where s will not be over-counted.

Finally, compared to weighted-sum scoring function optimization, multi-scoring functions sampling has the advantage of no weight determination required and no assumption of linear-combined weights.

VI. Population-based Evolutionary Multi-scoring Functions Sampling Algorithm for Protein Loop Structure Modeling

To demonstrate the effectiveness of multi-scoring functions sampling, we develop a population-based evolutionary algorithm to sample the loop backbone conformation space. A statistical pair-wise atoms distance-based scoring function [22] (knowledge-based) and a soft-sphere potential function [23] (physics-based) are employed to form the scoring function space.

Initially, a population with N conformations, $C_1, \ldots, C_N$, is randomly generated. Each loop structure conformation C_k with n residues is represented by a vector $(\theta_1, \ldots, \theta_{2n})$, which corresponds the backbone dihedral angles of $(\phi_1, \psi_1, \ldots, \phi_n, \psi_n)$. For simplicity, the dihedral angles of ω_k are kept constants at their average value of 180° while the bond angles and bond lengths are also remain constant. The greedy-heuristic Cyclic Coordinate Descent (CCD) algorithm [24] is applied to each conformation to adjust the dihedral angles so that the loop closure condition [26] is satisfied. Scores for multiple scoring functions $s_i(.)$ are evaluated for each conformation. Then, the relationship of each conformation pair C_p and C_q is evaluated according to

the dominance relation, which is generally defined as in [18] and below:

"A conformation C_p is said to (strongly) dominate another conformation C_q ($C_p \succ C_q$) if both conditions i) and ii) are satisfied:

i). for each scoring function $s_i(.)$, $s_i(C_p) \leq s_i(C_q)$ holds for all i;

ii). there is at least one scoring function $s_j(.)$ where $s_j(C_p) < s_j(C_q)$ is satisfied."

Let $D(C_k)$ denote the number of conformations in the population that C_k dominates. It is easy to prove that $D(C_p) > D(C_q)$ if $C_p \succ C_q$. Then, the N conformations in the population can be ranked in the decreasing order of $D(C_k)$ – a conformation dominating more other conformations will have a higher rank.

The probability that a conformation C_k is selected for reproduction, $P(C_k)$, is

$$P(C_k) = D(C_k) / \sum_{i=1}^{N} D(C_i).$$

We implement two methods to generate new conformations, including 1) mutation: permuting randomly selected dihedral angles of an old conformation; and 2) crossover: swapping part of the corresponding dihedral angle vectors between two old conformations. Again, the CCD algorithm is applied to every newly generated conformation by adjusting its dihedral angles to guarantee satisfaction of the loop closure condition. The top-ranked M conformations and the newly generated $N - M$ conformations form the new population and replace the old one.

The above procedure is repeated until convergence is observed in the population. Convergence is estimated by evaluating the likelihood of conformations within a population. The likelihood of two conformations C_p and C_q, $L(C_p, C_q)$, is measured by the RMSD value of the corresponding Cα atoms. The convergence in a population is reached if the M top-ranked conformations yield similar structures, i.e.,

$$\max_{1 \leq p \leq M,\, 1 \leq q \leq M,\, p \neq q} L(C_p, C_q) < \delta \quad ,$$

where δ is a threshold constant. The conformation with the highest rank is then outputted as a decoy.

The descriptive pseudo code of the population-based diversified sampling algorithm is described as follows. The program is executed repeatedly with different initial conformations to produce a set of decoys.

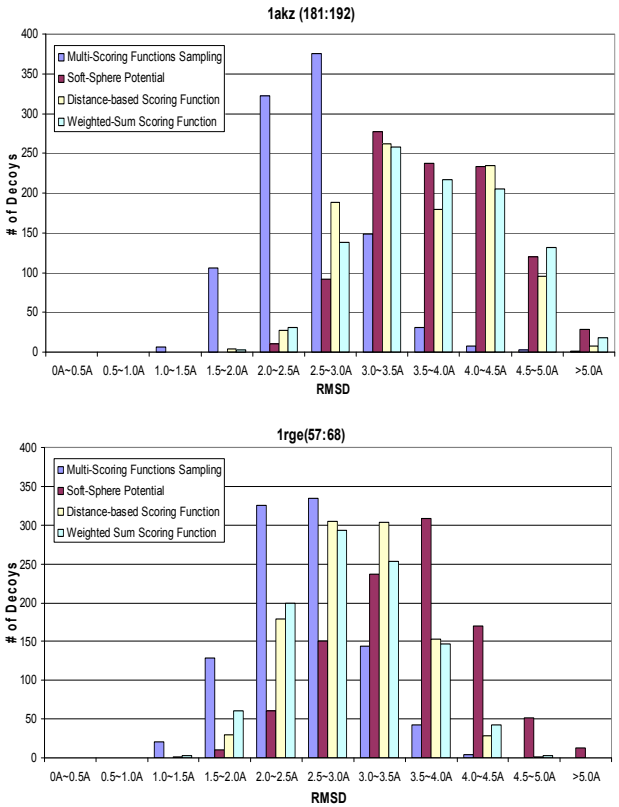
```
Initialize N conformations, C_1, …, C_N,
randomly
For each conformation C_k
   Adjust dihedral angles of C_k to satisfy
loop closure using CCD
Repeat {
   For each conformation C_k {
      For each scoring function s_i(.)
         Evaluate s_i(C_k)
      D(C_k) = 0
   }
   For each pair of conformations C_p and C_q {
      Evaluate the dominance relationship
      between C_p and C_q
      If C_p ≻ C_q, D(C_p)++
      If C_p ≺ C_q, D(C_q)++
   }
   Sort C_1, …, C_N so that D(C_1) ≥ D(C_2) ≥ … ≥
      D(C_N)
   Generate N - M new conformations, C_{N-M+1}',
      …, C_N'
   For each conformation C_k' in C_{N-M+1}', …, C_N'
      Adjust dihedral angles of C_k' to satisfy
loop closure
   Keep C_1, …, C_M in the population and
      replace C_{N-M+1}, …, C_N with C_{N-M+1}', …, C_N'
} Until convergence
Output the conformation with the highest rank
as a decoy
```

VII. Computational Results in Protein Loop Structure Modeling

Table 1 shows the average RMSD of the 1,000 decoys in the 17 12-residue targets listed in Jacobson's loop decoys library [14] generated based on multi-scoring functions sampling in comparison with those generated by optimizing the distance-based scoring function, the soft-sphere potential, and the weighted-sum scoring function. The weights of the weighted-sum scoring function are determined by a regression method based on the native loop structures in the benchmark library. Compared to optimizing single scoring function or weighted-sum scoring function, the multi-scoring functions sampling strategy yields averagely 0.16A~0.41A shift toward the native in the RMSD distribution of its decoys. In the targets of 1akz(181:192), 1ixh(160:171), 1dad(204:215), and 5nul(54:65), such RMSD shifts toward native in multi-scoring functions sampling are greater than or close to 0.5A. There is no significantly adverse RMSD shift in these targets in multi-scoring functions sampling. More importantly, sampling multiple scoring functions leads to enhanced chance of reaching decoys with good quality. Table 1 also shows improved average RMSD of the best decoys generated in multi-scoring functions sampling.

Table 1: The best and average backbone RMSD(A) of the 1,000 decoys generated by multi-scoring functions sampling, optimizing in each individual scoring function as well as the weighted-sum scoring function in 12-residue loops

PDB	Start Res.	End Res.	Multi-Scoring Functions Sampling		Soft-Sphere Potential Function		Distance-based Scoring Function		Weighted-Sum Scoring Function	
			Best	Avg.	Best	Avg.	Best	Avg.	Best	Avg.
1akz	181	192	1.19	2.62	2.07	3.78	1.91	3.61	1.86	3.70
1ixh	160	171	1.51	3.93	2.80	4.76	2.79	4.54	2.37	4.59
1cex	40	51	1.78	3.77	2.28	4.10	2.27	3.84	2.28	3.99
5pti	36	47	1.61	3.25	1.66	3.86	1.84	3.54	1.69	3.65
1rge	57	68	1.00	2.61	1.71	3.52	1.40	2.90	1.29	2.97
1arb	74	85	1.37	2.66	1.46	2.92	1.50	2.62	1.32	2.65
7rsa	13	24	1.50	3.02	2.15	3.77	2.10	3.48	1.37	3.49
1xyz	813	824	1.51	3.67	1.60	3.52	1.69	3.34	1.38	3.42
1cyo	32	43	1.34	3.52	1.63	3.35	1.44	3.26	1.72	3.45
153l	98	109	1.79	3.32	2.07	3.82	2.25	3.49	2.05	3.55
1bkf	9	20	1.10	2.85	1.59	3.32	1.44	3.02	1.50	3.15
1dad	204	215	1.42	2.76	1.98	3.68	1.72	3.29	1.42	3.40
1dim	213	224	1.10	3.46	1.73	3.61	1.62	3.59	1.25	3.68
1kuh	90	101	1.21	2.55	1.64	2.95	1.19	2.30	1.29	2.60
2ayh	21	32	1.67	3.26	1.64	3.11	1.54	3.02	1.51	3.20
351c	15	26	1.91	4.30	1.37	3.72	1.61	3.96	1.57	4.22
5nul	54	65	2.14	4.38	2.39	5.13	3.22	4.87	2.81	5.14
Average			1.48	3.29	1.87	3.70	1.85	3.45	1.69	3.58
Standard Deviation			0.31	0.59	0.38	0.57	0.53	0.63	0.45	0.65



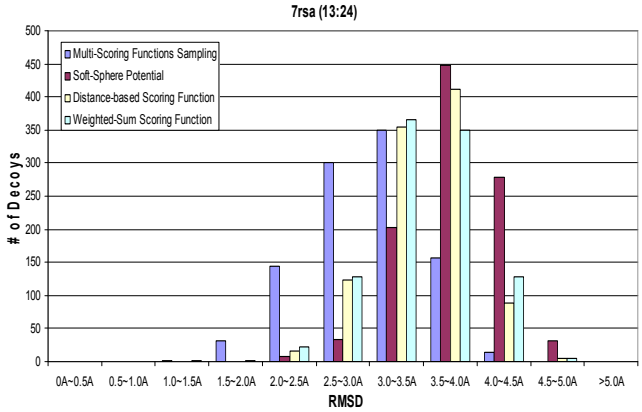


Figure 6: RMSD distributions in 1,000 decoys generated by multi-scoring functions sampling, optimizing distance-based scoring function, soft-sphere potential, and weighted-sum scoring function in protein loop targets of 1akz(181:192), 1rge(57:68), and 7rsa(13:24)

Figure 6 shows the RMSD distributions of 1,000 decoys generated by various sampling/optimization methods in loop targets 1akz(181:192), 1rge(57:68), and 7rsa(13:24) respectively. In these plots, one can find that in optimizing the soft-sphere potential function, distance-based scoring function, or the weighted-sum scoring function, although the "good", native-like conformations (RMSD < 2.0A) can be occasionally reached, the optimization method also produces large population of "bad" decoys (RMSD > 4.0A). This is because the "bad" decoys yield low scores in a single

scoring function due to the scoring function insensitivity problem. In contrast, the numbers of "bad" decoys are greatly reduced or even eliminated while the populations of native-like, "good" decoys are significantly larger in multi-scoring function sampling strategy. Our computational results indicate that the multi-scoring functions sampling strategy can tolerate insensitivity in an individual scoring function, because the decoys produced by the multi-scoring function sampling method are supposed to "satisfy" multiple scoring functions by yielding low scores. The increased "good" decoy population in multi-scoring functions sampling will enhance the chance of eventually generating high-resolution models in future clustering and structure refinement operations [25].

VIII. Conclusions and Future Research Directions

In this paper, we advocate a new protein conformation sampling approach based on multi-scoring functions sampling to address the scoring function insensitivity problem. By carefully selecting multiple physics- or knowledge-based scoring functions, the multi-scoring functions sampling approach intends to broadly explore the potentially "good" conformations in the protein conformation space and tolerate insensitivity in a single scoring function. We apply the multi-scoring functions sampling approach to modeling the backbone structure of long protein loops. Our population-based evolutionary algorithm using a distance-based scoring function and a soft-sphere potential function has shown increased "good" loop decoy production compared to optimizing an individual or weighted-sum scoring function.

Our future research directions include improving the efficiency and diversity of the multi-scoring functions sampling algorithms to sample conformations near/at the Pareto optimal front and applying the multi-scoring function strategy to more challenging protein modeling problems, such as *ab initio* protein folding, protein-ligand docking, and protein-protein interactions.

Acknowledgement

The work is partially supported by NSF under grants of CCF-0829382 and CCF-0845702 and NCSA Summer Faculty Fellowship to Y. Li.

References

[1] C. B. Anfinsen, "Principles that Govern the Folding of Protein Chains," Science, 181:223-230, 1973.

[2] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, M. Karplus, "CHARMM: a program for macromolecular energy, minimization and dynamics calculations," J. Comput. Chem. 4:187-217, 1983.

[3] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, "A second generation force-field for the simulation of proteins, nucleic-acids, and organic-molecules," J. Am. Chem. Soc. 117:5179-97, 1995.

[4] W. Damm, A. Frontera, J. Tirado-Rives, W. L. Jorgensen, "OPLS All-Atom Force Field for Carbohydrates," Journal of Computational Chemistry, 18(16): 1955-1970, 1997.

[5] W. F. van Gunsteren, H. J. C. Berendsen, "Groningen Molecular Simulation (GROMOS) Library Manual," Groningen, The Nether.: BIOMOS, 1987.

[6] H. Gohlke, G. Klebe, "Statistical Potentials and Scoring Functions Applied to Protein-Ligand Binding," Current Opinion in Structural Biology, 11(2):231-235, 2001.

[7] A. Godzik, "Knowledge-based potentials for protein folding: what can we learn from known protein sequences?" Structure, 4:363-366, 1996.

[8] B. A. Naim, "Statistical potentials extracted from protein structures: are these meaningful potentials?" J. Chem. Phys. 107: 3698-3706, 1997.

[9] P. D. Thomas, K. A. Dill, "Statistical potentials extracted from protein structures: how accurate are they?" J. Mol. Biol. 257: 457-469, 1996.

[10] J. A. Kocher, M. J. Rooman, S. J. Wodak, "Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches," J. Mol. Biol., 235: 1598-1613, 1994.

[11] H. J. Bohm, "The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure," J. Comput. Aided Mol. Des, 8(3): 243-256, 1994

[12] A. N. Jain, "Scoring noncovalent protein-ligand interactions: continuous differentiable function tuned to compute binding affinities," J. Comput. Aided. Mol. Des, 10(5): 427–440, 1996.

[13] Y. Li, A. J. Bordner, Y. Tian, X. Tao, A. Gorin, "Extensive Exploration of the Conformational Space Improves Rosetta Results for Short Protein Domains", Proceedings of International Conference on Computational Systems Bioinformatics, Stanford, 2008.

[14] M. P. Jacobson, D. L. Pincus, C. S. Rapp, T. J. F. Day, B. Honig, D. E. Shaw, R. A. Friesner, "A hierarchical approach to all-atom protein loop prediction," Proteins: Structure, Function, and Bioinformatics, 55(2): 351-367, 2004.

[15] C. A. Rohl, C. E. Strauss, K. M. Misura, D. Baker, "Protein Structure Prediction using Rosetta," Methods in Enzymology, 383: 66-93, 2004.

[16] Y. Yang, Y. Zhou, "Ab initio folding of terminal segments with secondary structures reveals the fine difference between two closely-related all-atom statistical energy functions," Protein Science, 17:1212-1219, 2008.

[17] M. Shen, A. Sali, "Statistical potential for assessment and prediction of protein structures," Protein Science, 15: 2507-2524, 2006.

[18] K. Deb, "Multi-objective optimization using evolutionary algorithms," John Wiley&Sons, 2001.

[19] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, P. G. Wolynes, "Funnels, Pathways, and the Energy Landscape of Protein Folding: A Synthesis," Proteins, 21:167-195, 1995.

[20] W. H. Wong, F. Liang, "Dynamic Weighting in Monte Carlo and Optimization," Proceedings of the National Academy of Sciences of the USA, 94:14220-14224, 1997.

[21] A. Mitsutake, Y. Sugita, Y. Okamoto, "Generalized-ensemble algorithms for molecular simulation of biopolymers," Biopolymers, 60(2): 96-123, 2001.

[22] A. Rojnuckarin, S. Subramaniam, "Knowledge-based interaction potentials for proteins", Proteins: Structure, Function, and Genetics 36:54-67, 1999.

[23] H. Zhang, L. Lai, Y. Han, Y. Tang, "A Fast and Efficient Program for Modeling Protein Loops," Biopolymers, 41: 61-72, 1997.

[24] A. A. Canutescu, R. L. Dunbrack, "Cyclic Coordinate Descent: A robotics algorithm for protein loop closure," Protein Science, 12:963-972, 2003.

[25] K. Zhu. D. L. Pincus, S. Zhao, R. A. Friesner, "Long loop prediction using the protein local optimization program," Proteins, 65: 438-452, 2006.

[26] E. A. Coutsias, C. Seok, M. P. Jacobson, K. A. Dill, "A Kinematic View of Loop Closure," J. Comput. Chem. 25: 510-528, 2004.

[27] J. Handl, S. C. Lovell, J. Knowles, "Investigations into the Effect of Multiobjectivization in Protein Structure Prediction," Lecture Notes in Computer Science, 5199: 702-711, 2008.