Pareto-Based Optimal Sampling Method and Its Applications in Protein Structural Conformation Sampling

Yaohang Li and Ashraf Yaseen

Department of Computer Science, Old Dominion University, Norfolk, VA, USA
{yaohang, ayaseen}@cs.odu.edu

Abstract

Efficiently sampling the protein conformation space is a critical step in *de novo* protein structure modeling. One of the important challenges in sampling is the inaccuracy of available scoring functions, i.e., a scoring function is not always sufficiently accurate to distinguish the correct conformations from the alternatives and thereby exploring the very minimum of a scoring function does not necessary reveal correct conformations. In this paper, we present a Pareto optimal sampling (POS) method to address the inaccuracy problem of scoring functions. The POS method adopts a new computational sampling strategy by exploring diversified conformations on the Pareto optimal front in the function space consisted of multiple scoring functions, representing consensus with different trade-offs among multiple scoring functions. Our computational results in protein loop structure sampling and protein backbone structure sampling have demonstrated the effectiveness of the POS method, where near-natives are found in the ensemble of Pareto-optimal conformations.

1. Introduction

One of the most important open problems in molecular biology is the prediction of the three-dimensional conformation of a protein from its primary structure, i.e., from the linear sequence of its amino acids, which is referred as the *de novo* protein folding problem. The native conformation of a given protein is the one observed under physiological conditions of temperature, pH, and ion balance. What distinguishes the native from the other conformations is that it has the minimum free energy of all accessible conformations, according to the Anfinsen's thermodynamic hypothesis (Anfinsen 1973). Taking Anfinsen's thermodynamic hypothesis as the theoretical foundation, a typical *de novo* protein structure prediction application builds a scoring (energy) function to approximate the protein folding energy landscape and then expects to find the reasonable structures close to the native by sampling the conformations yielding lowest scores in the scoring function. A representative success is the Rosetta program (Baker 2000), which has demonstrated its effectiveness in a series of CASP experiments (Simons et al. 1999, Chivian et al. 2005, Das et al. 2007, Raman et al. 2009).

Nevertheless, in *de novo* protein structure modeling, in addition to the tremendously large protein conformation space to explore, an another important challenge is the accuracy of a scoring function, i.e., whether a scoring function is sufficiently accurate to distinguish the correct conformation from the alternatives. Although a good scoring function can usually differentiate a correct conformation from those erroneous, far-deviated ones, in practice, the native often does not yield the minimum score in a scoring function (Dima, Banavar, and Maritan 2000). Moreover, exploring the global minimum of a scoring function does not always guarantee correct conformations either (Li et al. 2008), particularly when the coarse-grained scoring functions and reduced protein representations are employed during sampling. Consequently, instead of looking for the very deep global minimum of a scoring function, most *de novo* protein structure modeling methods focus their sampling efforts on obtaining an ensemble of structurally diversified conformations yielding low scores.

In this paper, we investigate a new strategy to sample protein conformation space by exploring the function space of multiple scoring functions. Assuming that a correct conformation should yield reasonably low scores, although not necessary the lowest, the rationale of our approach is to obtain an ensemble of structurally diversified conformations on the Pareto optimal front (Deb 2001) of multiple scoring function space. The Pareto optimal front represents the consensus with various trade-offs among multiple scoring functions. A population-based computational method, so-called Pareto optimal sampling

(POS), is developed to implement this new sampling strategy. In this paper, we first study the effectiveness of the POS method on sampling protein loop structures. Then, we discuss our preliminary results on using the POS method for backbone conformation sampling of a complete protein.

2. Background

2.1 Scoring Functions in Protein Structure Modeling

In protein structure modeling, the scoring functions are used to evaluate the feasibility of a particular protein or protein complex structure. According to the different ways these scoring functions are generated, they can be categorized into two major categories, physics- and knowledge-based scoring functions.

2.2 Physics-based Scoring Functions

Ideally, the protein or protein complex 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 large size of protein molecules. As a compromise, artificial scoring functions (force fields) are developed based on classical physics to approximate the true energy of molecular systems. Generally, the energy functions have the following generic form:

$$E(\mathbf{R}) = \sum_{bonds} B(\mathbf{R}) + \sum_{bondangles} A(\mathbf{R}) + \sum_{torsions} T(\mathbf{R}) + \sum_{non\text{-}bonds} N(\mathbf{R}),$$

where $\mathbf{R}$ is the conformation vector of the protein molecule. $B(\mathbf{R})$ is the bond energy term corresponding to the stretching and compression of the bond length. $A(\mathbf{R})$ is the bond angle energy term corresponding to changes in the angle between bonds. $T(\mathbf{R})$ is the torsional energy term with respect to torsion angles in series of three bonds. The non-bonding term $N(\mathbf{R})$ usually includes Van der Waals interactions, steric clash, and electrostatic interaction. In addition, terms such as hydrogen bonding, hydrophobic, and salvation, are often used in various energy functions. Physics-based scoring functions popularly used in protein structure modeling include CHARMM (Brooks et al. 1983), AMBER (Cornell et al. 1996), OPLS (Damm et al. 1997), and GROMOS (Gunsteren and Berendsen 1987). Recent study by Raval et al. (Raval et al. 2012) on protein structure refinement has shown that the current physics-based force fields are still not accurate enough to refine a protein structure homology model to near-natives even when very long (>100μs) Molecular Dynamics (MD) simulations are carried out.

In order to lower the degrees of freedom in protein structure sampling, reduced representations of protein structures (Sun 1993) have been popularly used. For example, ϕ-ψ angles, backbone atoms, Cα atoms, or side chain centers of mass are often used to represent all atoms of a protein chain. Correspondingly, simplified and coarse-grained scoring functions are developed to facilitate the reduced representations of protein structures, where further accuracy loss is inevitably associated.

2.3 Knowledge-based Scoring Functions

The knowledge-based approaches evaluate the increasing number of experimentally determined conformations in PDB (Protein Data Bank) to extract statistics on preferred configurations and combinations. Sippl's potentials of mean force approach (Sippl 1990) has been popularly used to derive knowledge-based scoring functions from these statistics. According to the inverse Boltzmann theorem, the knowledge-based energy (score) is calculated as

$$\Delta E = -kT\ln(f_1 / f_2),$$

where k is the Boltzmann constant, T is the temperature, f_1 is the frequency of observation of a certain conformation in the data set, f_2 is the referenced frequency, and ΔE is the energy difference related to the ratio of two state's occupancies (f_1 and f_2). There exist numerous knowledge-based scoring functions based on various aspects of protein structures or data sets, including pair-wise atom-atom distance (Zhang, Liu, and Zhou 2004), secondary structures (Li et al. 2013), residue contacts (Miyazawa and Jernigan 1996), side chain orientation (Lu, Dousis, and Ma 2008) , torsion angle distribution (Rata, Li, and Jakobsson 2010) , etc.

2.4 Combining Multiple Scoring Functions

Park and Levitt (Park and Levitt 1996) found that certain combinations of scoring functions lead to improved accuracy. More generally, a linearly combined scoring function can be built by adding up various scoring terms with a particular set of weights (Eramian et al. 2006). These weights are assigned by regression. The combined scoring function is typically more accurate than an individual scoring function.

3. Methods

3.1 Pareto Optimality

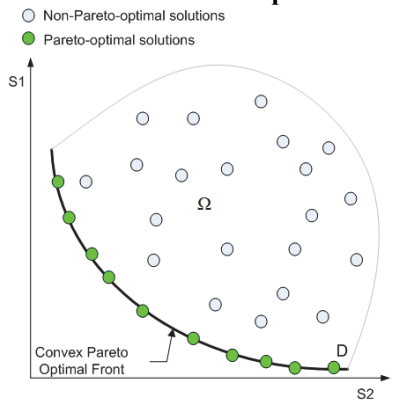
The theoretical foundation of POS method is Pareto optimality (Deb 2001), whose definition is based on the dominance relationship. Given a set of scoring functions $f_i(.)$, without loss of generality, assuming that minimization is the optimization goal for all scoring functions, a conformation u is said to dominate another conformation v ($u \prec v$) if both conditions i) and ii) are satisfied:

i) for each scoring function $f_i(.)$, $f_i(u) \le f_i(v)$ holds for all i; and
ii) there is at least one scoring function $f_j(.)$ where $f_j(u) < f_j(v)$ is satisfied.

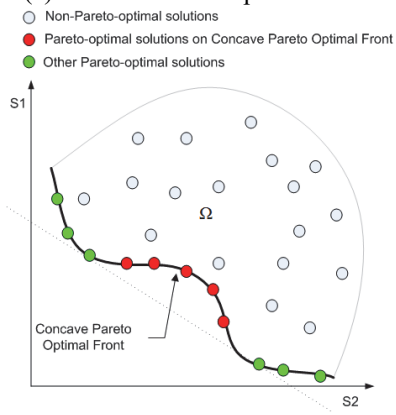
By definition, the conformations which are not dominated by any others in the conformation set form the Pareto-optimal solution set. The complete set of Pareto-optimal conformations is referred to as the Pareto-optimal front,

representing consensus with all possible trade-offs among multiple scoring functions.

3.2 Convex and Concave Pareto Optimal Fronts



(a) Convex Pareto Optimal Front



(b) Concave Pareto Optimal Front

Figure 1: Illustration of Convex and Concave Pareto Optimal Fronts in the Function Space of Two Hypothetical Scoring Functions

The Pareto optimal front of the function space of multiple scoring functions can be either convex or concave. Figures 1(a) and 1(b) illustrate a convex Pareto optimal and concave Pareto optimal front in the function space of two hypothetical scoring functions $S_1(.)$ and $S_2(.)$, respectively.

Denoting Ω the solution space for all accessible conformations, each point on a convex Pareto optimal front shown in Figure 1(a) corresponds to a conformation C that minimizes

$$w_1 S_1(C) + w_2 S_2(C),\ C \in \Omega,$$

for a specific pair of weights w_1 and w_2. However, this does not hold for a concave Pareto optimal front. The conformations on the concave surface (red points shown in Figure 1(b)) cannot be reached by any linear combination of $S_1(.)$ and $S_2(.)$ (Deb 2001, Li et. al. 2010). Consequently, optimizing a scoring function with linearly combined score terms will miss the conformations on the concave surface of the Pareto optimal front. In conclusion, comparing with the regression-based methods, a key advantage of the POS method is its capability of covering all Pareto optimal solutions, including those on the concave surface of the Pareto optimal front.

3.3 Fitness Assignment

In Pareto optimal sampling, there are two equally important goals:

1) optimality – finding the Pareto optimal and near Pareto optimal conformations; and
2) diversity – obtaining diversified coverage of the Pareto optimal front.

In multi-objective optimization literature, there has been a variety of fitness assignment methods designed to balance these two goals (Deb 2001). In our POS implementation, we adopt the following simple fitness assignment scheme to ensure satisfaction of both goals.

Considering a population P with N individual conformations, C_1, …, C_N, the fitness calculation is based on the strength s_i of each non-dominated conformation C_i, which is defined as the proportion of conformations in P dominated by C_i. As a result, the fitness of an individual C_i is defined as

$$fit(C_i) = \begin{cases} s_i & C_i \text{ is non-dominated} \\ 1 + \sum_{j}^{i \succ j} s_j & C_i \text{ is dominated} \end{cases}.$$

The conformations with fitness less than 1.0 are the non-dominated ones in the population. The fitness function biases to the non-dominated ones with less dominated configurations while those with a lot of neighbors in their dominating niche are discriminated (Vrugt et al. 2003).

3.4 Pareto Optimal Sampling Algorithm

By putting all the pieces together, a generic POS algorithm is described in the following pseudocode.

Algorithm: A Generic POS Algorithm

$P \leftarrow$ Initialize Population (N)

do {

  $P^* \leftarrow$ Proposed New Population (P)

  $P \leftarrow P \cup P^*$

  Evaluate pair-wise dominance relationship in P

  Assign fitness according to dominance relationship

  Sort conformations in P according to fitness

  $P \leftarrow$ {Select top N conformations in P}

}

until (convergence is reached)

Output non-dominated conformations in P

3. Results

3.1 Protein Loop Structure Modeling

We apply the POS method to model protein loop structures (Li, Rata, and Jakobsson 2011), which is

regarded as a "mini protein folding problem" (Fiser, Do, and Sali 2000) under geometric constraints, such as loop closure and avoidance of steric clashes with the remainder of the protein structure. Three scoring functions, including backbone Rosetta (Baker 2000), DFIRE (Zhang, Liu, and Zhou 2004), and Triplet (Rata, Li, and Jakobsson 2010) are incorporated. Differential Evolution (DE) (Storn and Price 1997) is employed to crossover dihedral angles in conformations in old population to propose new conformations and a Monte Carlo scheme (Li 2012) is used to determine acceptance of the new conformations. Cyclic Coordinate Descent (CCD) (Canutescu and Dunbrack 2003) is applied to each newly generated conformation to guarantee loop closure. A population size of 8,000 is used in modeling each loop target and the final conformations after sampling are generated as structure decoys.

PDB ID	# of Decoys on Pareto Optimal Front	Start Res.	End Res.	Best RMSD of Decoys on Pareto Optimal Front	RMSD of Decoy with lowest Rosetta Score	RMSD of Decoy with lowest Triplet Score	RMSD of Decoy with lowest DFIRE Score
1rhs	258	216	224	0.344	2.430	0.563	1.773
1npk	131	102	110	0.263	1.682	2.750	0.463
1pda	135	108	116	0.271	0.509	2.395	1.561
1tca	224	170	178	0.358	1.709	2.222	0.590
1php	102	91	99	0.252	0.821	3.602	0.550
2cpl	166	24	32	0.268	1.806	2.622	0.474
1mrk	195	53	61	0.716	3.122	2.934	2.640
1gpr	230	63	71	0.384	2.931	2.034	2.491
1amp	33	57	65	0.373	0.574	0.872	0.548
1pgs	236	117	125	0.366	2.036	3.337	3.320
1btl	120	102	110	0.463	2.740	1.939	0.629
1csh	316	252	260	0.234	2.792	1.710	0.660
1xnb	148	116	124	0.747	1.640	2.633	1.447
1ptf	212	10	18	0.355	0.813	2.922	1.998
1arb	139	168	176	0.529	0.644	2.832	2.684
1wer	203	942	950	0.682	2.771	3.069	2.984
1xyzA	239	568	576	0.398	2.061	2.000	3.061
2hbg	216	18	26	0.392	2.360	2.946	2.720
1fus	154	91	99	0.376	2.647	0.847	2.652
1isuA	130	30	38	0.455	2.487	2.479	2.613

Table 1: backbone RMSD values of decoys with lowest Rosetta, Triplet, and DFIRE scores as well as the best RMSD of decoys on the Pareto optimal front for 20 9-residue loop targets

Table 1 lists the backbone RMSD values of decoys with lowest Rosetta, Triplet, and DFIRE scores as well as the best RMSD of decoys on the Pareto optimal front for 20 9-residue loop targets. Considering decoys with RMSD less than 1A as near-natives, 1amp(57:65) is the ideal case, where all three scoring functions agree on the near-natives. Unfortunately, such consensus among these three scoring function does not happen very frequently. More often, the Rosetta, Triplet, and DFIRE scoring functions are effective on different loop targets. Particularly, in addition to 1amp(57:65), the Rosetta scoring function correctly identifies near-natives in loop targets 1pda(108:116), 1php(91:99), 1ptf(10:18), and 1arb(168:176). In contrast, the Triplet scoring function works best in 1rhs(216:224) and 1fus(91:99) while the DFIRE scoring function performs best in 1npk(102:110), 1tca(170:178), 1php(91:99), 2cpl(24:32), 1btl(102:110), and 1csh(252:260). Also quite often, for example, in loop targets 1mrk(53:61), 1gpr(63:71), 1pgs(117:125), 1xnb(116:124), 1wer(942:950), 1xyz_A(568:576), and 1isu_A(30:38), none of these three scoring functions can successfully identify the near-natives. Nevertheless, in all these loop targets, the decoy ensembles on the Pareto optimal front include near-native conformations. This indicates that POS can effectively tolerate inaccuracy among Rosetta, Triplet, and DFIRE scoring functions and thus sampling leads to near native conformations.

3.2 Modeling the Complete Protein

We use the POS method to sample backbone structures of complete proteins. Backbone Rosetta and DFIRE are used to build the multi-scoring function space. We use torsion angles φ and ψ to represent protein backbone structures. Bond lengths and bond angles are kept in ideal values. New conformations are generated using the Metropolis algorithm (Metropolis et al. 1953) by proposing new torsion angles according to propensity distributions of torsion angle pairs derived from Ramachandran plots. Predicted secondary structures with higher than 90% confidence are preserved during sampling. A population size of 1,000 is used.

Figure 2 shows the distribution of non-dominated conformations (after structurally clustering) for the A chain of 1ay7. The conformation with lowest Rosetta score has 12.05A RMSD (Figure 3a) and the one with lowest DFIRE score has 13.74A RMSD (Figure 3b). However, the best conformation (4.39A RMSD) lies on the Pareto optimal front with certain compromise between Rosetta and DFIRE scoring functions. The structure of this conformation is displayed in Figure 3c, where one can find that its secondary structures are correctly modeled and mostly aligned with the natives.

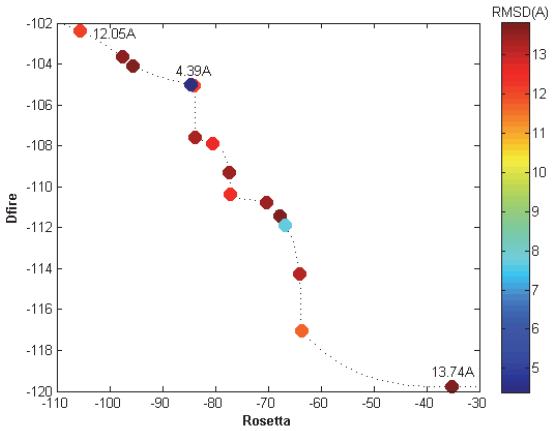
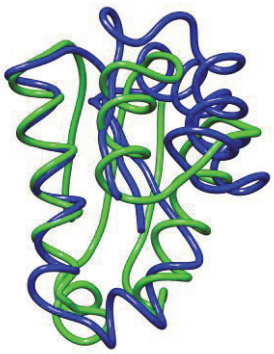
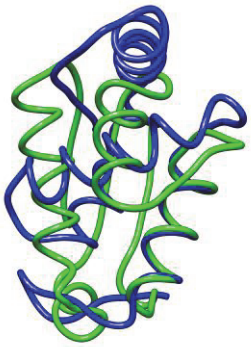


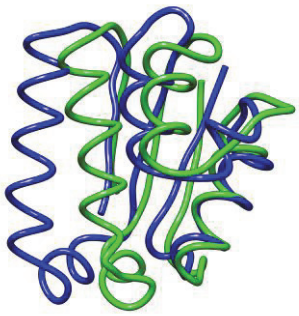
Figure 2: Distribution of Non-dominated Conformations in 1ay7 A Chain



(a) Conformation with lowest Rosetta score (RMSD 12.05A)



(b) Conformation with lowest DFIRE score (RMSD 12.05A)



(c) Best conformation (RMSD 4.39A) on the Pareto optimal front

Figure 3: Conformations with lowest Rosetta score, lowest DFIRE score, and lowest RMSD value on the Pareto optimal front. (Green: native structure; Blue: model)

4. Summary and Future Research Directions

In this paper, we develop a Pareto optimal sampling (POS) method to explore protein conformation space. The fundamental idea is to obtain diversified conformations on the Pareto optimal front in the function space of multiple scoring functions. The POS method has demonstrated its effectiveness in modeling protein loop structures. Our preliminary results have also shown that the POS method is able to discover near-native conformations in modeling protein backbone structures.

One of the main disadvantages of the POS method is its high computational cost. Multiple scoring functions have to be evaluated at each iteration step in POS. Moreover, to sufficiently cover the Pareto optimal front of multiple sophisticated scoring functions, a large population size is also needed. Nevertheless, taking advantage of the emerging large-scale high performance computing architectures, such as Graphics Process Units (GPUs), may greatly reduce the computational time of POS in sampling protein conformation space (Yaseen and Li 2012), which will be one of our future research directions.

Acknowledgements

This work is partially supported by NSF grant 1066471 and ODU 2013 Multidisciplinary Seed grant.

References

Anfinsen, C. B. 1973. Principles that govern the folding of protein chains. *Science* 181: 223-230.

Baker, D. 2000. A surprising simplicity to protein folding. *Nature* 405: 39-42.

Simons, K. T., Bonneau, R., Ruczinski, I., and Baker, D. 1999. Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins* 3: 171-176.

Chivian, D., Kim, D. E., Malmstrom, L., Schonbrun, J., Rohl, C. A., and Baker, D. 2005. Prediction of CASP6 structures using automated Robetta protocols. *Proteins* 61: 157-166.

Das, R., Qian, B., Raman, S., Vernon, R., Thompson, J., Bradley, P., Khare, S., Tyka, M. D., Bhat, D., Chivian, D., Kim, D. E., Sheffler, W. H., Malmstrom, L., Wollacott, A. M., Wang, C., Andre, I., and Baker, D. 2007. Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home. *Proteins* 8: 118-128.

Raman, S., Vernon, R., Thompson, J., Tyka, M., Sadreyev, R., Pei, J. M., Kim, D., Kellogg, E., DiMaio, F., Lange, O., Kinch, L., Sheffler, W., Kim, B. H., Das, R., Grishin, N. V., and Baker, D. 2009. Structure prediction for CASP8 with all-atom refinement using Rosetta. *Proteins* 77: 89-99.

Dima, R. I., Banavar, J. R., and Maritan, A. 2000. Scoring functions in protein folding and design. *Protein Science* 9: 812-819.

Li, Y., Bordner, A. J., Tian, Y., Tao, X., and Gorin, A. A. 2008. Extensive exploration of conformational space improves Rosetta results for short protein domains. In Proceedings of the 7th Annual International Conference on Computational Systems Bioinformatics (CSB08), 203-209. Stanford, Calif.: Comput Syst Bioinformatics Conf.

Deb, K. 2001. *Multi-objective optimization using evolutionary algorithms*. Chichester, New York.: John Wiley & Sons.

Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., and Karplus, M. 1983. Charmm - a Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *Journal of Computational Chemistry* 42: 187-217.

Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W., and Kollman, P. A. 1996. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society* 118: 2309-2309.

Damm, W., Frontera, A., TiradoRives, J., and Jorgensen, W. L. 1997. OPLS all-atom force field for carbohydrates. *Journal of Computational Chemistry* 18: 1955-1970.

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

Raval, A., Piana, S., Eastwood, M. P., Dror, R. O., and Shaw, D. E. 2012. Refinement of protein structure homology models via long, all-atom molecular dynamics simulations. *Proteins* 80: 2071-2079.

Sun, S. J. 1993. Reduced Representation Model of Protein-Structure Prediction - Statistical Potential and Genetic Algorithms. *Protein Science* 2: 762-785.

Sippl, M. J. 1990. Calculation of Conformational Ensembles from Potentials of Mean Force - an Approach to the Knowledge-Based Prediction of Local Structures in Globular-Proteins. *Journal of Molecular Biology* 213: 859-883.

Zhang, C., Liu, S., and Zhou, Y. Q. 2004. Accurate and efficient loop selections by the DFIRE-based all-atom statistical potential. *Protein Science* 13: 391-399.

Li, Y., Liu, H., Rata, I., and Jakobsson, E. 2013. Building a Knowledge-Based Statistical Potential by Capturing High-Order Inter-residue Interactions and its Applications in Protein Secondary Structure Assessment. *Journal of Chemical Information and Modeling* 53: 500-508.

Miyazawa, S., and Jernigan, R. L. 1996. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *Journal of Molecular Biology* 256: 623-644.

Lu, M., Dousis, A. D., and Ma, J. 2008. OPUS-PSP: an orientation-dependent statistical all-atom potential derived from side-chain packing. *Journal of Molecular Biology* 376: 288-301.

Rata, I. A., Li, Y., and Jakobsson, E. 2010. Backbone Statistical Potential from Local Sequence-Structure Interactions in Protein Loops. *Journal of Physical Chemistry B* 114: 1859-1869.

Gohlke, H., and Klebe, G. 2001. Statistical potentials and scoring functions applied to protein-ligand binding. *Current Opinion in Structural Biology* 11: 231-235.

Park, B., and Levitt, M. 1996. Energy functions that discriminate X-ray and near-native folds from well-constructed decoys. *Journal of Molecular Biology* 258: 367-392.

Li, Y., Rata, I., Chiu, S. W., and Jakobsson, E. 2010. Improving predicted protein loop structure ranking using a Pareto-optimality consensus method. *BMC Structural Biology* 10.

Vrugt, J. A., Gupta, H. V., Bastidas, L. A., Bouten, W., and Sorooshian, S. 2003. Effective and efficient algorithm for multiobjective optimization of hydrologic models. *Water Resources Research* 39: 1214-1232.

Fiser, A., Do, R. K. G., and Sali, A. 2000. Modeling of loops in protein structures. *Protein Science* 9: 1753-1773.

Storn, R., and Price, K. 1997. Differential evolution - A simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization* 11: 341-359.

Li, Y. 2012. MOMCMC: An efficient Monte Carlo method for multi-objective sampling over real parameter space. *Computers & Mathematics with Applications* 64: 3542-3556.

Canutescu, A. A., and Dunbrack, R. L. 2003. Cyclic coordinate descent: A robotics algorithm for protein loop closure. *Protein Science* 12: 963-972.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. 1953. Equation of State Calculations by Fast Computing Machines. *Journal of Chemical Physics* 21: 1087-1092.

Li, Y., Rata, I., and Jakobsson, E. 2011. Sampling Multiple Scoring Functions Can Improve Protein Loop Structure Prediction Accuracy. *Journal of Chemical Information and Modeling* 51: 1656-1666.

Eramian, D., Shen, M. Y., Devos, D., Melo, F., Sali, A. 2006. A composite score for predicting errors in protein structure models. *Protein Sci.* 15(7): 1653-1666.

Yaseen, A. and Li, Y. 2012. Accelerating Knowledge-based Energy Evaluation in Protein Structure Modeling with Graphics Processing Units. *Journal of Parallel and Distributed Computing* 72(2): 297-307.