NIH Public Access
Author Manuscript
J Chem Inf Model. Author manuscript; available in PMC 2012 July 25.

Published in final edited form as:
J Chem Inf Model. 2011 July 25; 51(7): 1656–1666. doi:10.1021/ci200143u.

Sampling Multiple Scoring Functions Can Improve Protein Loop Structure Prediction Accuracy

Yaohang Li[1], **Ionel Rata**[2], and **Eric Jakobsson**[3]

Yaohang Li: yaohang@cs.odu.edu; Ionel Rata: rata@uiuc.edu; Eric Jakobsson: jake@ncsa.uiuc.edu

[1]Department of Computer Science, Old Dominion University

[2]Center for Biophysics and Computational Biology, University of Illinois at Urbana-Champaign

[3]Department of Molecular and Integrative Physiology, Beckman Institute, and National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign

Abstract

Accurately predicting loop structures is important for understanding functions of many proteins. In order to obtain loop models with high accuracy, efficiently sampling the loop conformation space to discover reasonable structures is a critical step. In loop conformation sampling, coarse-grain energy (scoring) functions coupling with reduced protein representations are often used to reduce the number of degrees of freedom as well as sampling computational time. However, due to implicitly considering many factors by reduced representations, the coarse-grain scoring functions may have potential insensitivity and inaccuracy, which can mislead the sampling process and consequently ignore important loop conformations.

In this paper, we present a new computational sampling approach to obtain reasonable loop backbone models, so-called the Pareto Optimal Sampling (POS) method. The rationale of the POS method is to sample the function space of multiple, carefully-selected scoring functions to discover an ensemble of diversified structures yielding Pareto optimality to all sampled conformations. POS method can efficiently tolerate insensitivity and inaccuracy in individual scoring functions and thereby lead to significant accuracy improvement in loop structure prediction. We apply the POS method to a set of 4- to 12-residue loop targets using a function space composed of backbone-only Rosetta, DFIRE, and a triplet backbone dihedral potential developed in our lab. Our computational results show that in 501 out of 502 targets, the model sets generated by POS contain structure models are within subangstrom resolution. Moreover, the top-ranked models have Root Mean Square Deviation (RMSD) less than 1A in 96.8%, 84.1%, and 72.2% of the short (4~6 residues), medium (7~9 residues), and long (10~12) targets, respectively, when the all-atom models are generated by local optimization from the backbone models and are ranked by our recently developed Pareto Optimal Consensus (POC) method. Similar sampling effectiveness can also be found in a set of 13-residue loop targets.

1. Introduction

Prediction of the loop regions conformations in proteins is important in structural biology for its wide applications in homology modeling[1], segment defining[2], protein design[3], characterizing protein functions[4,5], and ion channel simulation[6,7]. Loop structure prediction without the use of a structural template typically involves several phases including sampling, clustering, refining, and ranking[8,9]. In the sampling phase, the loop conformation space is

Correspondence to: Yaohang Li, yaohang@cs.odu.edu.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

explored to produce a large ensemble of reasonable, coarse-grain models. Then, representative models are selected from the ensemble using a clustering algorithm to reduce redundancy. These representative models are used to build fine-grain models in the refining phase. Finally, in the ranking phase, the models are assessed and the top-ranked ones are determined as the predicted results. The sampling phase is of particular importance – if the sampling process cannot reach conformations close enough to the native, it is unlikely to obtain a high-resolution near-native model in the refining phase.

Although significantly smaller than that of a complete protein molecule, the loop conformation space still poses sampling challenge for loops of nontrivial length. To reduce the number of degrees of freedom to achieve sampling efficiency, coarse-grain approaches are often employed to construct loop conformations, in which the loop peptide can be modeled using a reduced representation. A popular approach is to approximately represent loop backbone conformations using a set of (θ, ψ) torsion angles[8–14]. Alternatively, fragment libraries[15,16] with backbone atoms and a few side chain atoms can also be used to buildup loop backbones. Loop closure constraints are satisfied by finding analytical solutions[46,47], using random tweak[17], inverse kinematics[18], or Cyclic Coordinate Descent (CCD)[19]. The recent kinematic closure approach developed by Mandell et al.[20] can lead to loop buildups with subangstrom resolution in long loops.

Coupled with reduced representations in loop modeling, coarse-grain energy (scoring) functions are developed to guide the sampling process toward native-like conformations. For example, Zhang et al.[22] developed a soft-sphere potential to fast construct loops. Rohl et al.[15] modeled loops by optimizing the Rosetta score using fragment buildup. We also developed a statistics-based potential[22] to assess loop torsions based on the distribution of adjacent θ-ψ backbone torsion angle pairs in the context of all possible residues as derived from structural data in a loop library. Jacobson et al.[8] built loop samples using a high resolution rotamer library and then filtered them with various heuristic scores/criteria. On the other hand, all-atom statistics- or physics-based energy (scoring) functions are used in the refining and ranking phases. DFIRE, the statistics-based scoring function, has proven to be successful in loop selection[26]. Cui et al. developed a grid-based force field, where their Monte Carlo sampling approach can obtain solutions less than 1.8A in a test set containing 14 protein loops[48]. Jacobson et al.[8], Zhu et al.[9], Rapp and Friesner[12], de Bakker et al.[13], Felts et al.[14], and Rapp et al.[23] modeled loops using physics-based potentials such as CHARMM[29], AMBER[30], and OPLS-AA[31] with various solvent models. Other terms can also be incorporated to improve the accuracy of the scoring functions. For example, Xiang et al.[11] developed a combined energy function with force-field energy and RMSD (Root Mean Square Deviation) dependent terms while Fogolari and Tosatto[24] took advantage of the concept of “colony energy” by considering the loop entropy (an important component in flexible loops) as part of the total free energy.

The scoring functions play a critical role in loop modeling. One of the problems in coarse-grain scoring functions is their potential inaccuracy due to the fact that many factors are treated implicitly and thus the coarse-grain score may not faithfully reflect the true energy surface. Although not always, distortion introduced by the coarse-grain scoring functions may potentially mislead the sampling process and thus ignore some important conformations. Even the all atom scoring functions have the insensitivity problem, i.e., the native or the near native-like conformations often do not necessarily exhibit the lowest scores. The scoring function inaccuracy or insensitivity have been shown and discussed in several loop modeling studies[8,10,11,25,27] with a variety of energy functions.

In this article, we report a Pareto Optimal Sampling (POS) method to sample loop conformation space. The fundamental idea of the POS method is to build a function space

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

composed of multiple, carefully-selected scoring functions and then sample an ensemble of diversified conformations yielding Pareto optimality. In certain sense, the POS method can be thought as a super consensus method –instead of finding a conformation with optimality in a scoring function or certain scoring functions combination, the POS method intends to cover all structurally diversified conformations with optimal consensus in every possible positive score combinations. This is achieved by obtaining diverse coverage of the Pareto optimal front of the function space composed of multiple scoring functions. The POS sampled conformations are expected to tolerate insensitivity in individual scoring functions, particularly those in coarse-grain, and thus increases the chance of discovering native-like, good conformations. We apply the POS method to a set of 4~12-residue loop targets and compare with the decoy sets generated by Jacobson et al.[8] using hierarchical all atom prediction approach. The loops in Jacobson's decoy sets are regarded as "difficult" targets[26,33] and Jacobson's decoy sets have been frequently used as a benchmark for loop prediction[26,33,34]. In addition to the loop targets in Jacobson's decoy sets, we compare POS sampling with recent results[9] on a set of 13-residue targets. Theoretical comparisons between POS and other consensus methods are also discussed.

2. Methods

2.1 Pareto Optimality

The theoretical foundation of the POS method is the Pareto optimality[36], whose definition is based on the dominance relationship. 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 $f_i(.)$, $f_i(u) \leq f_i(v)$ holds for all i;
- **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 other conformations in the conformation set form the Pareto-optimal solution set. The complete set of Pareto-optimal solutions is referred to as the Pareto-optimal front.

2.2 Function Space of Multiple Scoring Functions

We employ three scoring functions, including backbone Rosetta, DFIRE, and Triplet, to form the function space for loop sampling in POS. These scoring functions all have demonstrated their effectiveness in loop modeling. The Rosetta scoring function is a complex combination of physics- and statistics-based energy terms, which is one of the most successful tools in predicting overall backbone fold for protein domains that lacks any detectable structural analogs in PDB[38,39]. The Rosetta program has also demonstrated success in predicting protein loops[15,20]. DFIRE is a statistics-based potential energy which was derived from a high-resolution protein structure data set. DFIRE has previously proven to be successful in loop selection[26] as well as in filtering unreasonable models[34]. The Triplet torsion angle scoring function developed in our lab is based on the distribution of adjacent phi-psi backbone torsion angle pairs in the context of all possible adjacent residue triplets as derived from the coil library[37]. The Triplet scoring function has shown its effectiveness in distinguishing the native loop from the erroneous ones[22]. Due to the fact that they are developed using different methods and data sets, there is little correlation among these three scoring functions.

2.3 Population-based Sampling Guided by Multiple Scoring Functions

In most optimization methods used in protein structure modeling, the goal is to minimize single scoring (energy) function to search for a solution with lowest energy; whereas in POS

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

sampling involving multiple scoring functions, there are two equally important goals: 1) sampling a set of solutions near the Pareto optimal front; and 2) obtaining a diverse coverage of solutions near the Pareto optimal front. The POS sampling also differs from the classical multi-objective optimization (MOO) problems[36], in that POS is not only interested in the conformations at the Pareto optimal front, but also in those near the Pareto optimal front with diversified structures.

The POS method is a population-based approach to sample the loop backbone conformations in the function space of multiple scoring functions. In the population-based sampling method, initially, population P with N loop conformations, $C_1, \ldots, C_N$, is randomly generated. Each loop structure conformation C_i with n residues is represented by a vector $(\theta_1, \ldots, \theta_{2n})$, which represents the dihedral angles of $(\theta_1, \psi_1, \ldots, \theta_n, \psi_n)$. The dihedral angles of ω_i are kept constant at their average value of 180° while the bond angles and bond lengths also remain constant. Fitness is designed to measure the dominance relationship and popularity of a conformation in the population. Differential evolution style[41] crossover for continuous torsion angles is developed to generate new conformations satisfying loop closure condition. Adaptive Monte Carlo is used to determine acceptance of new conformations based on the Metropolis rule. The population is updated by the Metropolis transition and only the top-ranked conformations are selected for reproduction, which will evolve the population toward the Pareto optimal front. Moreover, the similarities between conformations in the population can be exploited, which will allow the sampling process to discriminate the popular ones so as to achieve solution diversity.

2.4 Fitness Assignment

The POS sampling program uses a fitness assignment scheme based on the number of external non-dominated points[40]. Considering a population P with N individual conformations, $C_1, \ldots, C_N$, 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

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

The conformations with fitness less than 1.0 are the non-dominated ones in the population. For the nondominated conformations, the fitness function achieves diversity by biasing to those with less dominated conformations while discriminating the popular conformations dominating a lot of conformations[40]. For the dominated conformations, preferences are given to those dominated by fewer conformations with less strength. Conformations in a population are sorted according to their fitness and only the top m ones are allowed for reproduction in POS.

2.5 Differential Evolution-Style Crossover

Differential Evolution (DE)[41] is used in POS sampling to produce new conformations by crossing over the selected conformations from the current population in loop conformation sampling. We select the "DE/rand/2/exp" scheme listed in DE[41], which is a robust scheme particularly suitable for scoring functions with complicated landscapes. In "DE/rand/2/exp" scheme, for each configuration $C_i(\theta_1, \ldots, \theta_{2n})$, a mutant vector V_i is typically formed by

$$V_i = C_{r5} + F(C_{r1} + C_{r2} - C_{r3} - C_{r4}),$$

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

where *i*, *r1*, *r2*, *r3*, *r4*, and *r5* are mutually distinct integer random numbers in the interval [1, *m*],*m* refers to the top *m* conformations selected for reproduction, and $F > 0$ is a tunable amplification control constant as described in DE[41]. Other forms of mutant vector generation are also provided in DE[41]. Then, a new conformation $C_i'(\theta_1', \ldots, \theta_{2n}')$ is generated by the crossover operation on V_i and C_i:

$$x_j' = \begin{cases} v_j & j=\langle k\rangle_{2n}, \langle k+1\rangle_{2n}, \ldots, \langle k+L-1\rangle_{2n} \\ x_j & otherwise \end{cases},$$

where $\langle . \rangle_{2n}$ denotes the modulo operation with modulus 2*n*, *k* is a randomly generated integer from the interval [0, 2*n*-1], *L* is an integer drawn from [0, *n*-1] with probability $\Pr(L = l) = (CR)^l$, and $CR \in [0,1)$ is the crossover probability. Practical advice suggests that $CR = 0.9$ and $F = 0.8$ are suitable choices the DE scheme[41]. These parameter values are also adopted in our sampling algorithm.

After DE crossover, the loop closure condition of the new loop conformation is usually not satisfied. The greedy-heuristic Cyclic Coordinate Descent (CCD) algorithm[19] is applied to each new conformation to close the loop. To obtain minimum distortion of the newly crossover section, CCD starts from the torsion angle next to the crossover torsion segment. Figure 1 illustrates the procedure using DE-style crossover to generate new loop conformation.

2.6 Adaptive Monte Carlo

In POS sampling, Monte Carlo moves based on Metropolis transition are used to determine acceptance of newly generated conformations in the population. The transition probability depends only on the change of the fitness function value. The Metropolis-Hastings ratio is calculated as:

$$w(C_i \rightarrow C_i') = e^{\frac{-(fit(C_i') - fit(C_i))}{T}}.$$

The new configuration C_i' generated by DE is accepted with the probability

$$\min(1, w(C_i \rightarrow C_i')).$$

T is the simulated temperature, which is used to control the acceptance rate of the Metropolis transitions. Studies[43,44,49] in Monte Carlo method literature have shown that the Markov Chain Monte Carlo (MCMC) sampler yields the best possible performance with an acceptance rate of around 20%. In this paper, we also maintain an acceptance rate of 15%~25% by adjusting *T* in each iteration. When the acceptance rate is less than 15%, *T* is increased by 10%. When the rate is more than 20%, *T* is decreased by 10%.

Using DE in population-based MCMC is proved to guarantee detailed balance[55]. However, control the acceptance rate by adjusting the simulated temperature *T* in POS may impact the detailed balance nature of the underlying MCMC and thus the resulting samples may not strictly follow Boltzmann distribution. Nevertheless, complying with detailed balance condition is not necessary in this loop modeling application, since our goal is to obtain a good coverage of loop conformations near the Pareto optimal front.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

2.7 Convergence Analysis

We evaluate the convergence of the loop conformation sampling by measuring the hypervolume indicator[42]. For a population $P(C_1, \ldots, C_k, \ldots, C_N)$, the hypervolume indicator, $HYP(P)$, relative to a reference point R, is defined as

$$HYP(P) = \bigcup_k VOL(S_k, R).$$

R can be arbitrarily selected in the scoring function space where $\max R_i \geq \max_j (f_i(C_j))$ for all scoring functions $f_1, \ldots, f_i, \ldots, f_n$. For convenience, we select the worst scoring function values to build R. $VOL(U, V)$ is the Lebesgue measure of two points, U and V, in the scoring function space, which is defined as $VOL(U, V) = \prod_i |U_i - V_i|$. S_k is the corresponding point of conformation C_k in the scoring function space.

The hypervolume indicator measures the dominated volume of the non-dominated solutions. Generally, the dominated solutions do not contribute to the hypervolume indicator and the worse case computational complexity of computing hypervolume is $O(M^{d-1})$, where d is the number of scoring functions and M is the number of non-dominated solutions. Usually, for large d, calculation of hypervolume is very costly. However, in our loop conformation sampling program using POS, only three scoring functions are used and thereby calculation of the hypervolume indicator does not pose significant overhead to the overall sampling computation. The hypervolume indicator value is monitored during the sampling process. Since POS is based on MCMC transitions, we can calculate the autocorrelation function, $\Gamma_{HYP}(T)$, to evaluate the integrated autocorrelation time (IAT), τ_{HYP}, to estimate the number of iterations to achieve convergence. The autocorrelation function, $\Gamma_{HYP}(T)$, is defined as,

$$\Gamma_{HYP}(T) = \frac{\sum_{t=0}^{T} \mathrm{cov}(HYP(P_0), HYP(P_t))}{\mathrm{var}(HYP)},$$

where $HYP(P_t)$ refers to the hypervolume indicator of the POS population P at time t. The IAT can be estimated by

$$\tau_{HYP} \cong \frac{1}{2} + \sum_{t=1}^{\infty} \Gamma_{HYP}(t).$$

Literatures[45] in MCMC indicate that the number of iterations, NUM, for the sampler should be $NUM >> \tau_{HYP}$.

Moreover, due to the large conformation space, a single sampling run may not lead to sufficient coverage of the diversified conformations near the Pareto optimal front. As a result, we repeat the sampling procedure with different random number sequences until the overall hypervolume indicator of the all models generated stops growing. The outputted models from these sampling runs form the backbone model set.

2.8 Sampling the Pareto Optimal Front

Figure 2 shows the multiple scoring functions coordinate plots of the model sets in targets 1alc(34:41), 1fnd(262:269), 1ptf(65:71), and 1onc(70:78) obtained by the POS method. All scores are linearly normalized in [0, 1]. As shown in 2(a), 2(b), and 2(c), respectively,

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

models yielding the lowest scores in DFIRE, Triplet, and Rosetta have the best RMSD values (< 0.5A), while those with lowest scores in the other two scoring functions are far deviated with RMSD > 2.0A. This strongly indicates that neither DFIRE, Triplet, nor Rosetta is a perfect scoring function – one may be effective in some targets while fail in some others. More often, as shown in the example in 2(d), the native-like models do not exhibit minimum scores in either one of these scoring functions, but exhibit Pareto optimality in certain scoring functions combination. Our computational results on 502 targets in Jacobson's decoy set show that, in 41 (8.2%) and 22 (4.4%) targets, the Pareto optimal models with best RMSD values in the model set generated by POS sampling are at least 0.5A and 1.0A better in RMSD than the models with lowest DFIRE, Triplet, or Rosetta scores, respectively. Ideally, the POS method intends to capture all structurally diversified conformations on the Pareto optimal front in the functional space of the scoring functions.

2.9 All-atom Prediction after Sampling

After sampling using POS, three major steps are applied to the generated backbone model sets to obtain final all-atom prediction results. First of all, the backbone model set is clustered with a 0.5A cutoff and one representative candidate is selected for each cluster. Secondly, the PLOP package[8] is used to add side chains to the representative backbone models and then perform local optimization using OPLS/AA force field with SGB solvent models to generate the all-atom model set. Finally, we use our recently developed Pareto Optimal Consensus (POC) method[27] to rank the all-atom models. Five scoring functions, including all-atom Rosetta, DFIRE, DOPE, OPLS/AA, and triplet, are used to form the function space in POC. The top-ranked models are selected as the predicted models.

3. Results

3.1 Sampling Efficiency

Figure 3 shows the comparison of average RMSD of best models as well as the model set size in POS model set and Jacobson's decoy set. One can find that for the short (4~6 res) and medium (7~9 res) targets, POS model set is averagely smaller than Jacobson's while for long targets (9~12 res), the average sizes of the two model sets are approximately the same. By comparing the resolution of models, it is important to notice that the average RMSD's of the best models in POS model set are lower than those in Jacobson's decoy set, particularly for medium and long targets. If models within 1A to the native are considered as native-like models, Table 1 lists all targets in which either POS or Jacobson's decoy set do not contain at least one native-like model. There are 20 out of 502 targets that Jacobson's decoy set misses a native-like model. In contrast, the model set generated by the POS method has only one miss in target 1poa(79:83), which demonstrates the efficiency of Pareto optimality-based sampling by integrating multiple scoring functions.

The 1poa(79:83) has some particularities that make it hard to predict especially by the statistical potentials, which are trained on mostly "regular" structures. Normally, a loop's backbone folds by packing itself against the rest of the protein. The loop usually has buried hydrophobic residues that help its folding, while the polar residues are exposed to the solvent. By contrast, 1poa(79:83) has no hydrophobic residues and a polar Asparagine residue is partially buried forming two favorable hydrogen bonds inside the protein (the black dashed lines in Figure 4). Also, the loop's backbone is solvated and flexible (with two successive Glycine residues). The stability of this structure is helped by two disulfide bridges (yellow bars in Figure 4) formed by Cystine residues situated just near the two ends of the loop, which is also a rare situation.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

In addition to the sensitivity problem that the scoring functions might have, the 1poa(79:83) case poses a sampling problem. Our sampling approach is hierarchical where the good backbone conformations are firstly found and then the side-chains are later added and optimized. Nevertheless, in the case of 1poa(79:83), the Asparagine's side chain is the key to the loop's structure formation and the backbone adapts to Asparagine's structure. As a result, the backbone sampling approaches have difficulty in finding structures in the vicinity of the native.

In addition to the targets listed in Jacobson's decoy set, we also apply the POS method to a set of 13-residue targets listed in Zhu et al. Table 2 compares the sampling results. For comparison convenience, in Table 2, as Zhu et al. does, we also use superimposed RMSD instead of regular RMSD. The POS method results in significantly lower superimposed RMSD compared to Zhu et al. In 9 targets, including 1ojq(A167:A179), 1ock(A43:A55), 1bhe(121:133), 1cnv(110:122), 2hlc(A:91:A103), 1ako(203:215), 1g8f(A72:A84), 1f46(A64:A76), and 1jp4(A153:A165), POS can produce models within 1A resolution while Zhu et al. has sampling results over 1A or more.

There are three targets, 1eok(A147:A159), 1hxh(A87:A99), and 1qqp(2_161:2_173), where the POS method is significantly worse compared to Zhu et al's approach. Our further analysis finds that native loop 1eok(A147:A159) interacts with two sulfate ions and an asymmetric unit (Figure 5). Native loops 1hxh(A87:A99) and 1qqp(2_161:2_173) deeply interact with external chains as shown in Figure 6. Because the scoring functions we use in the POS method make no assumption of interactions with ligands, other asymmetric units, or external chains, the native or native-like loops are not scored well DFIRE, Triplet, Rosetta, or combination of these scoring functions. As a result, the native-like conformations in these targets are not present in the Pareto optimal front in the function space of these scoring functions and thereby the POS method has a high chance to miss those close to the native.

3.2 All-atom Prediction Accuracy

To identify the models with best quality within the model set, we use our recently developed Pareto Optimal Consensus (POC) method[27] for model ranking. In POC ranking, we evaluate models with all-atom scoring functions including Rosetta, DOPE, DDFIRE, I-TASSER, and OPLS-AA/SGB as well as Triplet, identify the models at the Pareto optimal front, and then rank them based on the fuzzy dominance relationship to the rest of the models in the model set. Table 3 lists the average RMSD of top-ranked model as well as the best top-5-ranked model in POS model set and Jacobson's decoy set. One can notice that in short, medium, and long loop targets, there is 0.05A~ 0.2A shift to the native in the model set generated by the POS method compared with Jacobson's decoy set. This is due to the fact that, in quite a few targets such as those listed in Table 1, the POS model set contains models with significantly higher resolution than those in Jacobson's decoy set. These models are identified as top models in the POC ranking process. Correspondingly, the POS method also leads to improved percentages of top-ranked and top-5-ranked models with RMSD less than 1A.

4. Discussion

4.1 Pareto Optimal Front

Various methods have been developed to take advantage of multiple scoring functions to improve protein modeling accuracy. One approach is to combine models generated by sampling guided by different scoring functions. For example, Fujitsuka et al.[35] mixed models generated by SimFold and Rosetta separately and obtained higher prediction success rate than individual predictions.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

We use a conceptual scenario illustrating a function space of two scoring functions S_1 and S_2 shown in Figure 7 to compare POS and sampling of individual scoring functions. Successfully sampling an individual scoring function typically leads to clustered solutions near its correspondent global minimum, as shown in Figure 7. However, some Pareto optimal solutions compromised between S_1 and S_2 will likely be ignored by the sampling process due to their higher scores compared to those near the global minimum. In contrast, POS intends to explore not only solutions near the respective global minima of individual scoring functions, but also the other Pareto optimal solutions in the multiple scoring function space.

A more popular approach is to obtain a consensus scoring function by combining multiple scores via machine learning approaches. To name a few examples, Qiu et al.[28] used Support Vector Regression (SVR) to combine various terms into a consensus scoring function to rank protein models. Gao et al.[29] combined independent latent servers by reducing server correlation, which achieved significant accuracy improvement in contact prediction. We hereby compare POS with a consensus scoring function generated by the weighted-sum approach to theoretically show that certain Pareto optimal solutions may be unreachable in a weighted-sum consensus scoring function.

In the weighted-sum approach, weights are assigned to various scoring functions and a consensus scoring function is built by linearly combining the weighted score terms. Figure 8 shows a conceptual scenario of a nonconvex scoring function space of two scoring functions S_1 and S_2. 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 feasible solution space. However, there exists no contour line that can produce a tangent point with the feasible solution space in the region BC. 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. Even if nonlinear combination is used to integrate various score terms, some regions in the Pareto optimal front may still be unreachable. Compared to a consensus scoring function combining multiple scores, the POS method can explore solutions in both concave and convex Pareto optimal fronts, which can potentially lead to broader exploration of the loop conformation space.

Nonconvex Pareto optimal front are common in practical problems involving multiple scoring functions[50]. Nonconvexity in scoring functions and/or nonconvexity in feasible solution region can lead to formation of nonconvex Pareto optimal front[36], although this is not a sufficient condition. Proteins are generally thought to have a rough, funnel-like folding energy landscape with many deep local minima[52]. Thus, scoring functions approximating the protein energy are clearly nonconvex. Moreover, the Ramachandran plots suggest that certain torsion angles combinations in protein conformations are infeasible and the feasible solution regions are often nonconvex[51]. As a result, the Pareto optimal front formed by multiple scoring functions can potentially and likely have non-convex regions.

4.2 Selection of Scoring Functions

The selection of scoring functions is important in the POS method. We select the scoring functions for loop conformation sampling based on the following three criteria. First of all, the scoring functions selected should have certain accuracy. A poor scoring function may complicate the sampling process and bring erroneous models into the final solution set, although it will not prevent good models from being included in the final solution set when good scoring functions are present. The three scoring functions we selected in POS, Rosetta, DFIRE, and Triplet, have demonstrated their accuracy in loop modeling, selection, and/or

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

filtering[15,20,26,34,37]. Secondly, the scoring functions selected should exhibit low correlation. Incorporating additional highly correlated scoring functions into the POS scheme contributes little to sampling but increases the overall computational time in sampling, because scoring function calculation is usually costly. Table 4 shows the average Intraclass Correlation Coefficient (ICC)[55] of pairwise Rosetta, DFIRE, and Triplet in targets listed in Jacobson's decoy set. One can find that Rosetta-DFIRE has stronger correlation than Rosetta-Triplet and DFIRE-Triplet. This is because Triplet is scoring function in torsion space while no other atom-atom interactions are taken into consideration. Moreover, the scoring functions have stronger correlations in short targets than long ones. After all, the overall correlation among these scoring functions is low. It is important to notice that although having low overall correlation, the selected scoring functions likely agree on the good models with reasonable structures by yielding low scores due to their reasonably good accuracy. Figure 9 shows an example in 1nwp(A84:A91), where Rosetta and DFIRE have relatively low overall correlation (ICC = 0.165) (inconsistency particularly in poor models) but strong agreement models in native-like models. Finally, the number of scoring functions should be limited. The number of scoring functions strongly influences the performance of population-based multi-scoring functions optimization/sampling methods[53]. This is due to the fact that as the number of scoring functions increases, the possible solutions at the Pareto optimal front generally increase exponentially – as a result, sampling the solutions at the Pareto optimal front requires significantly larger population and more computational times to reach convergence.

The typical computational time for our serial program to model a 12-residue loop, including sampling, local optimization, and selection, ranges from 4 to 20 hours on an Intel Xeon 2.13GHz processor. Given the computational results shown in Session 4, we consider our selection of scoring functions effective.

5. Conclusions

By sampling the diversified conformations in the function space composed of Rosetta, DFIRE, and Triplet, our newly developed POS method is shown to be effective in exploring the loop conformation space. Theoretical analyses also prove that sampling the solution near the Pareto optimal front can lead to potentially broader exploration of the loop conformation space compared to sampling individual scoring functions or a consensus function. Consequently, the POS method produces effective model sets covering near-native conformations in targets in Jacobson's decoy set as well as those listed in Zhu's paper. As a result, the effective sampling directly leads to prediction resolution improvement in the top-ranked and top-5-ranked all-atom models.

One of the minor disadvantages of the POS method is its computational cost compared to sampling a single scoring function, where evaluations of multiple scoring functions are needed. Scoring function evaluation is usually the most costly component in a protein modeling program. However, the computations of multiple scoring functions are independent, which can be easily parallelized on modern high performance computing platforms. More interestingly, recent study by Handl et al.[32] also shows that optimizing multiple energy function components can even help reach the global energy minima in protein structure prediction by reducing the number of local minima as well as the local minima escaping time.

Our future research directions include incorporating side chain atoms into sampling and extending our Triplet scoring function by considering χ angles starting from χ_1. This may further enhance the accuracy of loop models using the POS method. We are also interested

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

in applying the POS method to other protein modeling applications, such as protein folding, protein-ligand docking, and protein-protein interactions.

Acknowledgments

We acknowledge support from NIH grants 5PN2EY016570-06 and 5R01NS063405-02 and from NSF grants 0835718, 0829382, and 1066471.

References

1. Bruccoleri RE. Ab initio loop modeling and its application to homology modeling. Methods in Molecular Biology. 2000; 143:247–264. [PubMed: 11084909]
2. Dmitriev OY, Fillingame RH. The rigid connecting loop stabilizes hairpin folding of the two helices of the ATP synthase subunit c. Protein Sci. 2007; 16(10):2118–2122. [PubMed: 17766379]
3. Martin AC, Cheetham JC, Rees AR. Modeling antibody hypervariable loops: a combined algorithm. Proc. Nat. Acad. Sci. 1989; 86(23):9268–9272. [PubMed: 2594766]
4. Jones S, Thornton JM. Prediction of protein-protein interaction sites using patch analysis. J. Mol. Biol. 1997; 272:133–143. [PubMed: 9299343]
5. Fetrow JS, Godzik A, Skolnick J. Functional analysis of the Escherichia coli genome using the sequence-to-structure-to-function paradigm: Identification of proteins exhibiting the glutaredoxin0thioredoxin disulfide oxidoreductase activity. J. Mol. Biol. 1998; 282:703–711. [PubMed: 9743619]
6. Tasneem A, Iyer LM, 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. 2005; 6(1):R4. [PubMed: 15642096]
7. Yarov-Yarovoy V, Baker D, Catterall WA. Voltage sensor conformations in the open and closed states in ROSETTA structural models of K+ channels. Proc. Natl. Acad. Sci. 2006; 103:7292–7297. [PubMed: 16648251]
8. Jacobson MP, Pincus DL, Rapp CS, Day TJF, Honig B, Shaw DE, Friesner RA. A Hierarchical Approach to All-atom Protein Loop Prediction. Proteins. 2004; 55:351–367. [PubMed: 15048827]
9. Zhu K, Pincus DL, Zhao S, Friesner RA. Long Loop Prediction Using the Protein Local Optimization Program. Proteins. 2006; 65:438–452. [PubMed: 16927380]
10. Fiser A, Do RKG, Sali A. Modeling of loops in protein structures. Protein Sci. 2000; 9:1753–1773. [PubMed: 11045621]
11. Xiang ZX, Soto CS, Honig B. Evaluating conformational free energies: the colony energy and its application to the problem of loop prediction. Proc. Natl. Acad. Sci. 2002; 99:7432–7437. [PubMed: 12032300]
12. Rapp CS, Friesner RA. Prediction of loop geometries using a generalized born model of solvation effects. Proteins. 1999; 35:173–183. [PubMed: 10223290]
13. 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. 2003; 51:21–40. [PubMed: 12596261]
14. Felts AK, Gallicchio E, Chekmarev D, Paris KA, Friesner RA, Levy RM. Prediction of Protein Loop Conformation Using the AGBNP Implicit Solvent Model and Torsion Angle Sampling. J. Chem. Theory Comput. 2008; 4(5):855–868. [PubMed: 18787648]
15. Rohl RA, Strauss CE, Chivian D, Baker D. Modeling structurally variable regions in homologous proteins with Rosetta. Proteins. 2004; 55:656–677. [PubMed: 15103629]
16. Fernandez-Fuentes N, Oliva B, Fiser AA. Supersecondary Structure Library and Search Algorithm for Modeling Loops in Protein Structures. Nucleic Acids Res. 2006; 34(7):2085–2097. [PubMed: 16617149]
17. Shenkin S, Yarmush DL, Fine RM, Wang H, Levinthal C. Predicting antobody hypervariable loop conformation. I. Ensembles of random conformations for ringlike structures. Biopolymers. 1987; 26:2053–2085. [PubMed: 3435744]

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

18. Kolodny R, Guibas L, Levitt M, Koehl P. Inverse Kinematics in Biology: The Protein Loop Closure Problem. Int. J. Robot. Res. 2005; 24:151–163.
19. Canutescu AA, Dunbrack RL. Cyclic Coordinate Descent: A Robotics Algorithm for Protein Loop Closure. Protein Sci. 2003; 12:963–972. [PubMed: 12717019]
20. Mandell DJ, Coutsias EA, Kortemme T. Sub-Angstrom Accuracy in Protein Loop Reconstruction by Robotics-Inspired Conformational Sampling. Nat. Methods. 2009; 6:551–552. [PubMed: 19644455]
21. Zhang H, Lai L, Han Y, Tang Y. A Fast and Efficient Program for Modeling Protein Loops. Biopolymers. 1997; 41:61–72.
22. Rata I, Li Y, Jakobsson E. Backbone statistical potential from local sequence-structure interactions in protein loops. Journal of Phys. Chem. B. 2010; 114(5):1859–1869. [PubMed: 20070091]
23. Rapp CS, Strauss T, Nederveen A, Fuentes G. Prediction of protein loop geometries in solution. Proteins. 2007; 69:69–74. [PubMed: 17588228]
24. Fogolari F, Tosatto SCE. Application of MM/PBSA colony free energy to loop decoy discrimination: Toward correlation between energy and root mean square deviation. Protein Sci. 2005; 14(4):889–901. [PubMed: 15772305]
25. Smith CS, Honig B. Evaluation of the conformational free energies of loops in proteins. Proteins. 1994; 18(2):119–132. [PubMed: 8159662]
26. Zhang C, Liu S, Zhou Y. Accurate and efficient loop selections by the DFIRE-based all-atom statistical potential. Protein Sci. 2004; 13(2):391–399. [PubMed: 14739324]
27. Li Y, Rata I, Chiu S, Jakobsson E. Improving predicted protein loop structure ranking using a Pareto-optimality consensus method. BMC Struct. Biol. 2010; 10:22. [PubMed: 20642859]
28. Gao X, Bu D, Xu J, Li M. Improving Consensus Contact Prediction via Server Correlation Reduction. BMC Struct. Biol. 2009; 9:28–42. [PubMed: 19419562]
29. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM: a program for macromolecular energy, minimization and dynamics calculations. J. Comput. Chem. 1983; 4:187–217.
30. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM. A second generation force-field for the simulation of proteins, nucleic-acids, and organic-molecules. J. Am. Chem. Soc. 1995; 117:5179–5197.
31. Damm W, Frontera A, Tirado-Rives J, Jorgensen WL. OPLS All-Atom Force Field for Carbohydrates. J. Comput. Biol. 1997; 18(16):1955–1970.
32. Handl J, Lovell SC, Knowles J. Investigations into the Effect of Multiobjectivization in Protein Structure Prediction. Lect. Notes Comput. Sc. 2008; 5199:702–711.
33. Lin MS, Head-Gordon T. Improved Energy Selection of Nativelike Protein Loops for Loop Decoys. J. Chem. Theory Comput. 2008; 4:515–521.
34. Soto CS, Fasnacht M, Zhu J, Forrest L, Honig B. Loop Modeling: Sampling, Filtering, Scoring. Proteins. 2008; 70(3):834–843. [PubMed: 17729286]
35. Fujitsuka Y, Chikenji G, Takada S. SimFold Energy Function for de novo protein structure prediction: consensus with Rosetta. Proteins. 2006; 62(2):381–398. [PubMed: 16294329]
36. Deb, K. Multi-objective optimization using evolutionary algorithms. 1st ed.. John Wiley & Sons; 2001. Multi-Objective Optimization; p. 13-46.
37. Fitzkee NC, Fleming PJ, Rose GD. The protein coil library: a structural database of non-helix, non-strand fragments derived from the PDB. Proteins. 2005; 58(4):852–854. [PubMed: 15657933]
38. Simons KT, Bonneau R, Ruczinski I, Baker D. Ab initio Protein Structure Prediction of CASP III Targets Using ROSETTA. Proteins. 1999; 37(3):171–176. [PubMed: 10526365]
39. Baker D. Surprising simplicity to protein folding. Nature. 2000; 405:39–42. [PubMed: 10811210]
40. Vrugt JA, Gupta HV, Bastidas LA, Boutem W, Sorooshian S. Effective and Efficient Algorithm for Multiobjective Optimization of Hydrologic Models. Water Resour. Res. 2002; 39(8):1214–1232.
41. Storn R, Price K. Differential Evolution – A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces. J. of Global Optim. 1997; 11(4):341–359.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

42. Zitzler E, Thiele L. Multiobjective Evolutionary Algorithms: A Comparative Case Study and the Strength Pareto Approach. IEEE T. Evolut. Comput. 1999; 3(4):257–271.
43. Rathore N, Chopra M, de Pablo JJ. Optimal Allocation of Replicas in Parallel Tempering Simulations. J. Chem. Phys. 2005; 122(2) 024111.
44. Kone A, Kofke DA. Selection of Temperature Intervals for Parallel Tempering Simulations. J. Chem. Phys. 2005; 122(20) 206101.
45. Berg, BA. Markov Chain Monte Carlo Simulations and Their Statistical Analysis. 1st ed.. Singapore: World Scientific; 2004. Error Analysis for Markov Chain Data; p. 196-235.
46. Go N, Scheraga HA. Ring Closure and Local Conformational Deformations of Chain Molecules. Macromolecules. 1970; 3(2):178–187.
47. Wedemeyer WJ, Scheraga HA. Exact Analytical Loop Closure in Proteins Using Polynomial Equations. J. Comput. Chem. 1999; 20(8):819–844.
48. 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 Eng. Des. Sel. 2008; 21(12):729–735. [PubMed: 18957407]
49. Kincaid RH, Scheraga HA. Acceleration of Convergence in Monte Carlo Simulations of Aqueous Solutions using the Metropolis Algorithm. J. Comput. Chem. 1982; 3:525–547.
50. Corne DW, Deb K, Fleming PJ, Knowles JD. The Good of the Many Outweights the Good of the One: Evolutionary Multiobjective Optimization. IEEE Newsletter: connections. 2003; 1:9–13.
51. Ramachandran GN, Ramakrishnan C, Sasisekharan V. Stereochemistry of polypeptide chain configurations. J. Mol. Biol. 1963; 7:95–99. [PubMed: 13990617]
52. Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG. Funnels, pathways, and the energy landscape of protein folding: A synthesis. Proteins. 1994; 21(3):167–195. [PubMed: 7784423]
53. Brockhoff, D.; Zitzler, E. Dimensionality Reduction in Multiobjective Optimization: The Minimum Objective Subset Problem; Proceedings of Operations Research. Kalcsics, J.; Nickel, S., editors. Springer; 2007. p. 423-429.
54. Ter Braak CJF. A Markov chain Monte Carlo version of the genetic algorithm differential evolution: easy Bayesian computing for real parameter spaces. Stat. Comput. 2006; 16:239–249.
55. Bartko JJ. The intraclass correlation coefficient as a measure of reliability. Psychol. Rep. 1966; 19:3–11. [PubMed: 5942109]

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

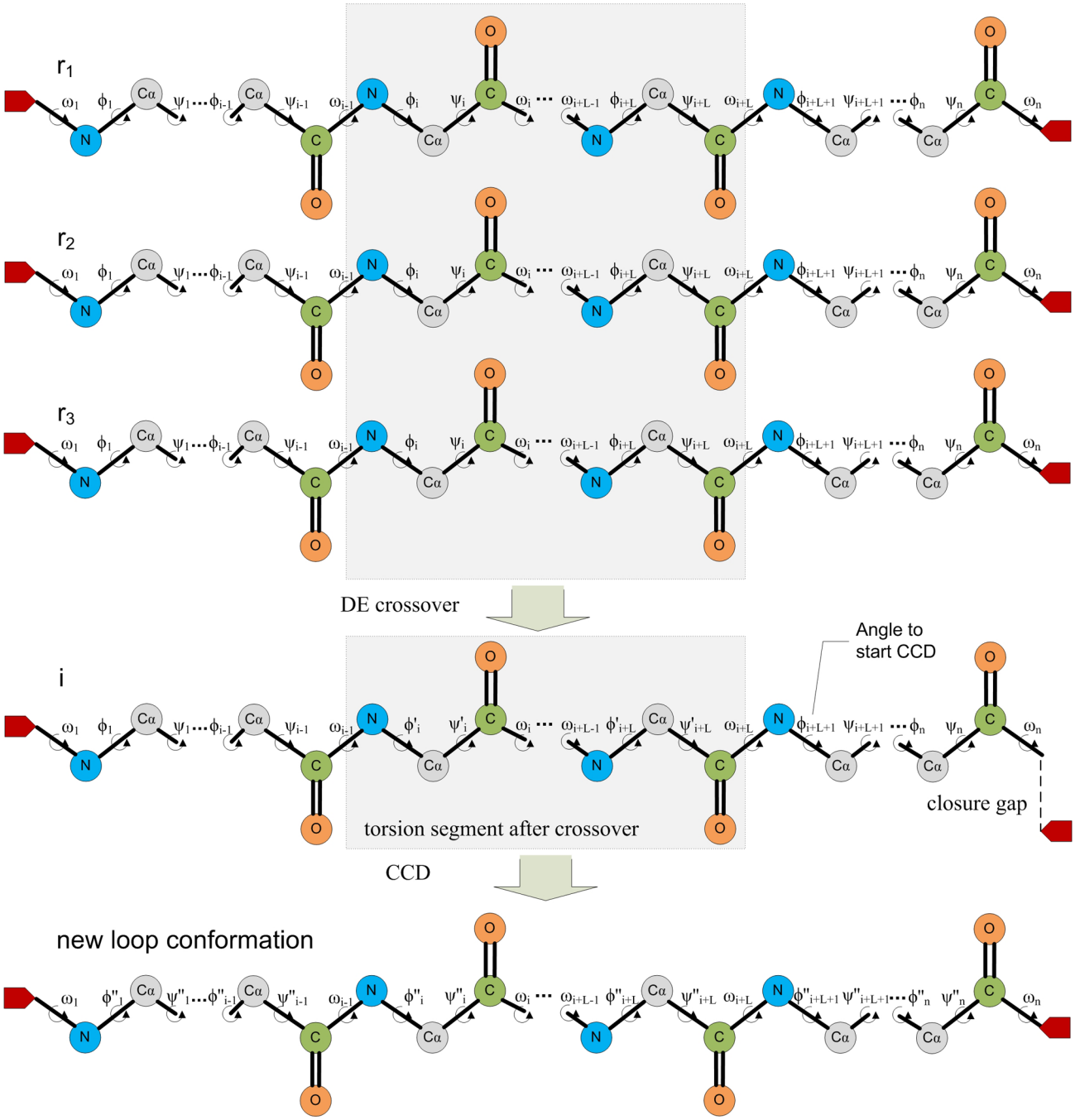
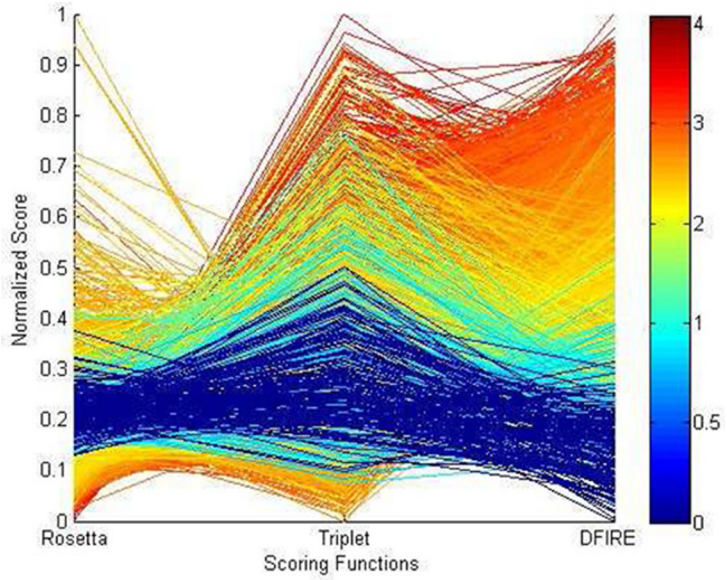


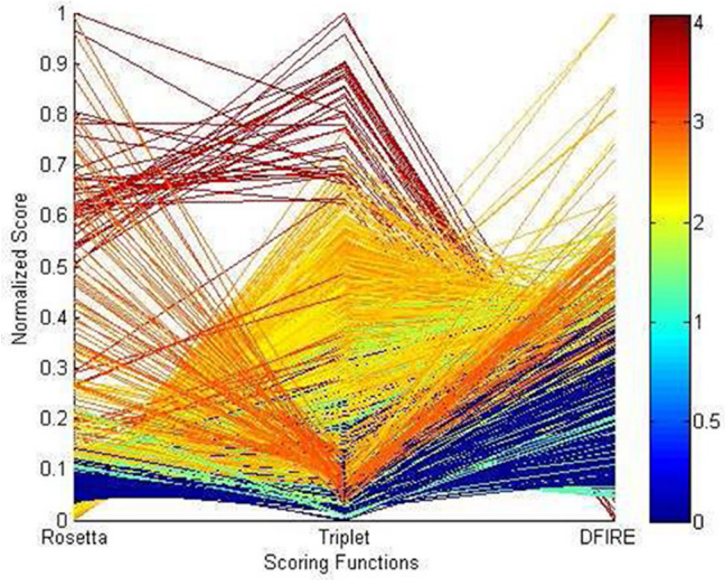
Figure 1.

DE crossover to produce new loop conformation. Segments from 5 randomly-selected conformations in the current population are used to build a mutant vector based on DE crossover. Then, for the *i*th loop conformation in the current population, the corresponding segment is replaced by the mutant vector to build a new conformation. The CCD algorithm is applied to the new conformation, starting from the torsion angle right after crossover, to ensure loop closure.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript



2(a): 1alc(34:41).



2(b): 1fnd(262:269)

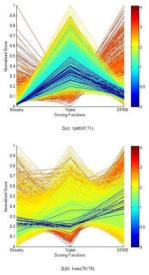


Figure 2.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Multiple coordinate plots of normalized Rosetta, Triplet, and DFIRE. Each line corresponds to a loop conformation in the generated model set and the color of the line is given by its RMSD value defined in the color bar (0.0A: blue and 4.0A: red). The Rosetta, Triplet, and DFIRE scores are linearly normalized in [0, 1], where the 0 and 1 correspond to the lowest and highest scores in individual scoring function, respectively. Models with lowest scores in Rosetta, Triplet, and DFIRE respectively in 1alc(34:41), 1fnd(262:269), and 1ptf(65:71) are native-like models with RMSD < 0.5A. In 1onc(70:78), models with best RMSD do not exhibit lowest scores in either Rosetta, Triplet, or DFIRE, but yield Pareto optimality in certain combination of these scoring functions.

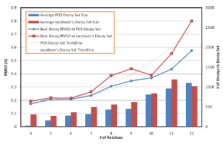


Figure 3.
Comparison of average RMSD of best models and model set sizes in POS model set and Jacobson's decoy set

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

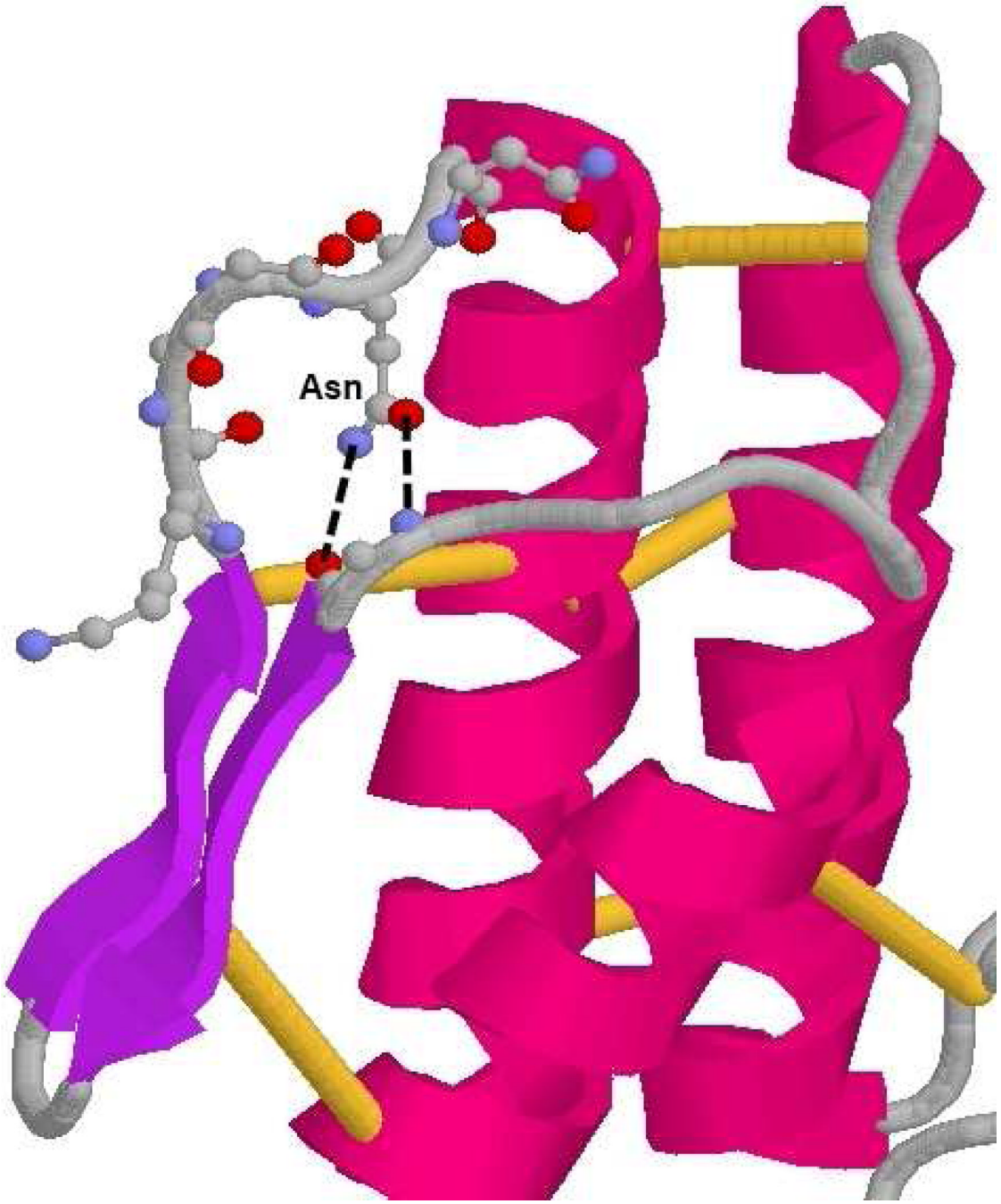


Figure 4.

The 1poa(79:83) loop has several unusual features: it is solvated and flexible having only polar residues and two consecutive glycines, it has a buried asparagine residue that makes two internal hydrogen bonds (represented by dashed lines), and it is flanked by two disulfide bonds near its ends (yellow bars).

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

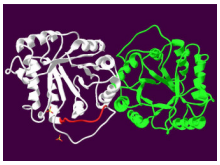
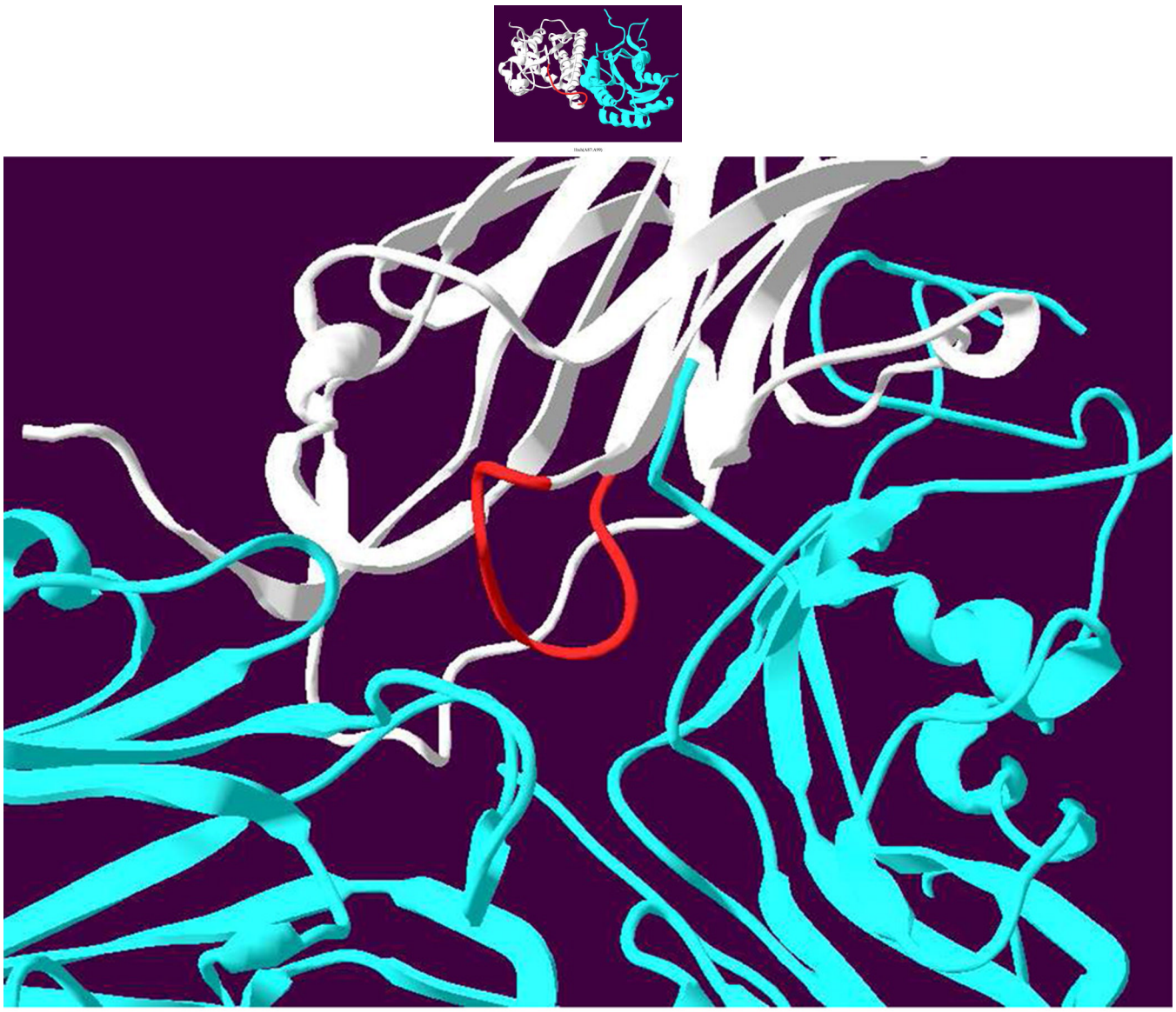


Figure 5.
Loop (red) in 1eok(A147:A159) interacts with sulfate ions and an asymmetric unit (green) in the unit cell

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript



1qqp(2_161:2_173)

Figure 6.
Loops (red) in 1hxh(A87:A99) and 1qqp(2_161:2_173) interact with external chains (blue)

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

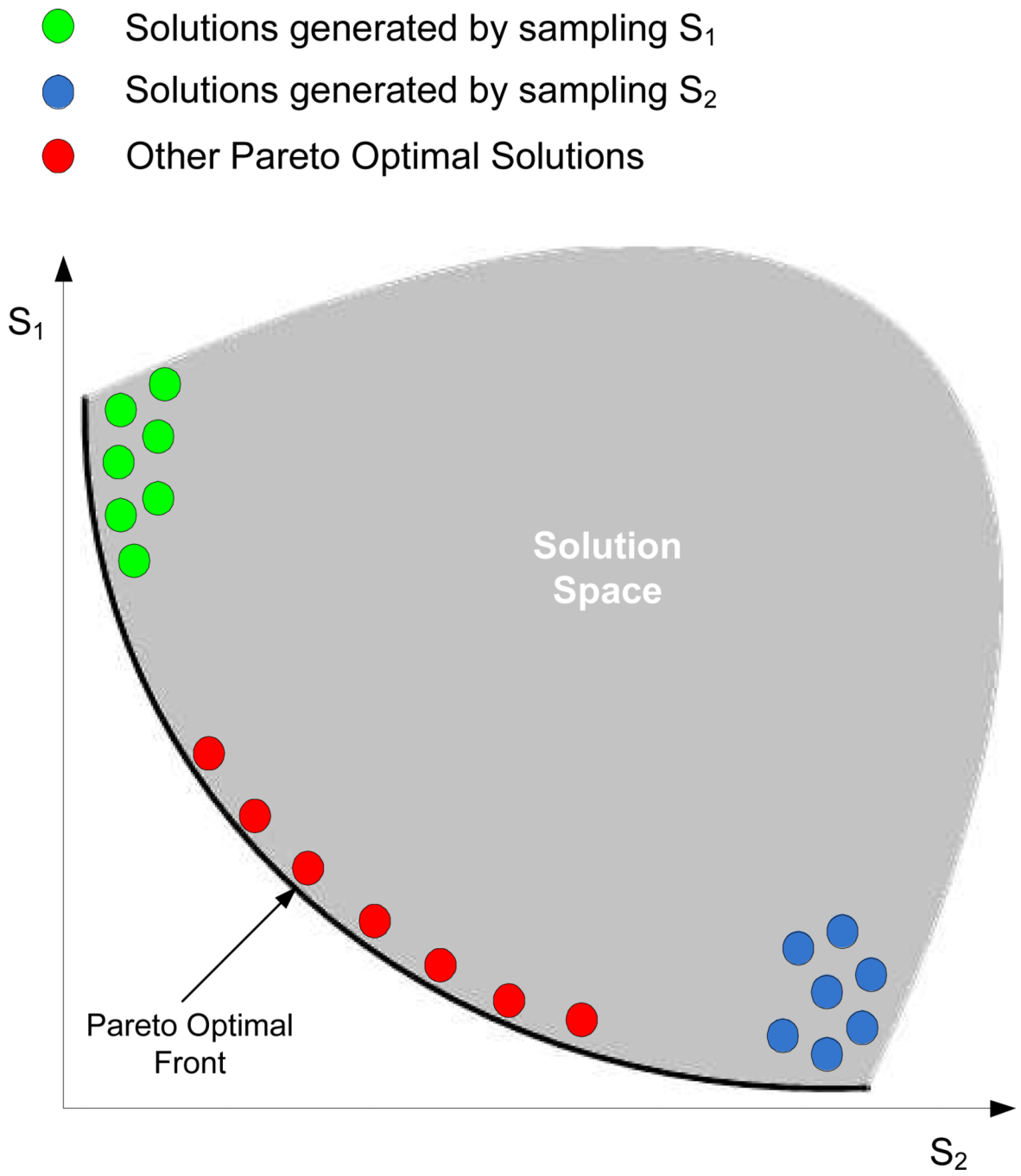


Figure 7.
Sampling individual scoring functions will likely ignore some Pareto optimal solutions. Sampling individual scoring functions typically leads to clustered solutions near the corresponding global minima (Green in S_1 and Blue in S_2) and ignores some other Pareto optimal solutions (Red).

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

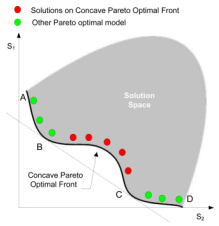


Figure 8.
Scenario of a nonconvex Pareto optimal front where a weighted-sum approach will fail to find some Pareto-optimal solutions

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

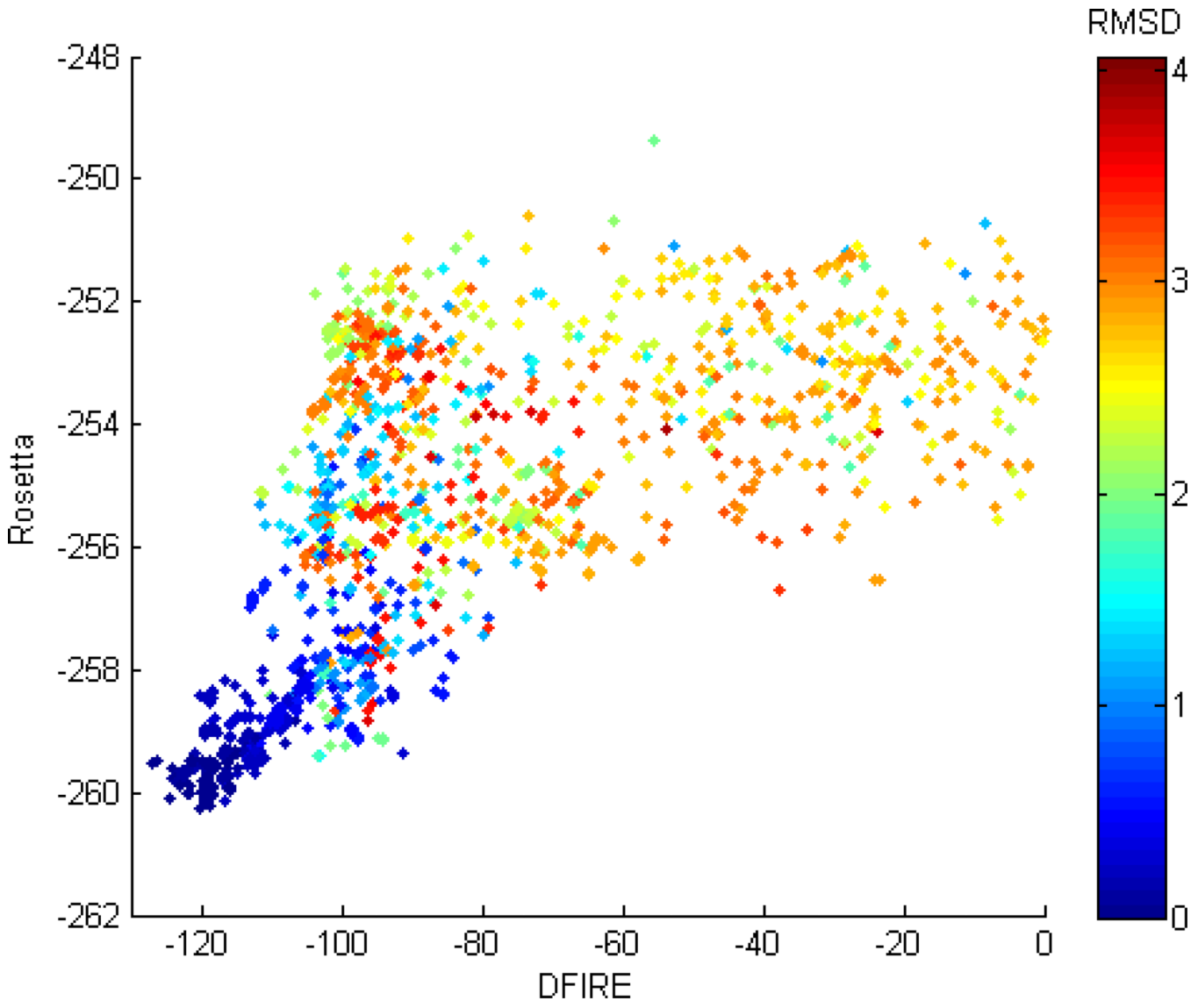


Figure 9.
Correlation between Rosetta and DFIRE on models in 1nwp(A84:A91)

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 1

List of all targets that the best models in either POS or Jacobson's decoy sets have RMSD over

PDB	RES: Start	RES: End	Best Model RMSD(A) in POS	Best Model RMSD (A) in Jacobson's Decoy Set
1dim	227	231	0.296	1.503
1rhs	21	27	0.376	1.092
1a2y	A:55	A:61	0.231	1.694
1cvl	15	21	0.334	1.019
1thw	18	25	0.17	1.111
1gof	606	613	0.225	1.797
1ppn	191	198	0.336	1.121
1a3c	92	99	0.156	1.253
1fus	31	39	0.435	1.761
1byb	246	254	0.275	2.223
1mla	194	202	0.227	1.716
1noa	99	107	0.47	1.011
1wer	942	950	0.522	1.452
1ixh	84	93	0.195	1.36
1dad	42	52	0.702	1.62
1fus	28	38	0.302	1.69
1a2p	A:76	A:86	0.413	1.63
1arb	74	85	0.599	1.45
153l	98	109	0.409	1.56
2ayh	21	32	0.807	1.24
1poa	79	83	1.32	0.311

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 2

Sampling results comparison in 13-residue targets listed in Zhu et al.

Target	RES: Start	RES: End	POS Results in Superimposed RMSD(A)	Zhu's Results in Superimposed RMSD(A)
1ojq	A:167	A:179	0.59	4.06
1dys	A:290	A:302	0.30	0.28
1dpg	A:352	A:364	1.08	1.27
1xyz	A:645	A:657	0.20	0.36
1eok	A:147	A:159	1.38	0.37
1ock	A:43	A:55	0.87	2.90
1hnj	A:191	A:203	1.76	3.11
1iir	A:197	A:209	0.13	0.21
1h4a	X:19	X:31	0.24	0.26
1bkp	A:51	A:63	0.25	0.83
1hxh	A:87	A:99	2.33	0.81
1nln	A:26	A:38	0.38	0.71
1bhe	121	133	0.65	2.45
1cnv	110	122	0.45	1.03
1gpi	A:308	A:320	0.27	0.70
2ptd	136	148	0.54	0.46
1lki	62	74	0.35	0.36
1d0c	A:280	A:292	0.26	0.30
1krh	A:131	A:143	0.41	0.72
2hlc	A:91	A:103	0.49	3.28
1ako	203	215	0.49	1.07
1ed8	A:67	A:79	0.06	0.26
1mo9	A:107	A:119	0.66	0.76
1g8f	A:72	A:84	0.42	1.41
1f46	A:64	A:76	0.51	1.27
1arb	182	194	0.67	0.85
1a8d	155	167	0.62	0.33
1kbl	A:793	A:805	0.62	0.48
1jp4	A:153	A:165	0.50	3.43
1l8a	A:691	A:703	0.54	0.25
1m8s	A:68	A:80	0.39	0.45
1qsl	A:389	A:401	1.01	3.61
1qqp	2:161	2:173	1.36	0.38
Average			0.63	1.18

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 3

Prediction accuracy comparison on 4~12-residue targets in Jacobson's decoy set (JDS) and model set generated by the POS method

		Short (4~6 Res)	Medium (7~9 Res)	Long (10~12)
Average RMSD (A) of Top-Ranked Models	POS	0.325	0.580	0.864
	JDS	0.372	0.655	1.076
Average RMSD (A) of Best Top-5-Ranked Models	POS	0.241	0.408	0.745
	JDS	0.311	0.513	0.904
Percentage of Top-Ranked Model with RMSD < 1A	POS	96.8%	84.1%	72.2%
	JDS	92.7%	81.1%	61.1%
Percentage of Best Top-5-Ranked Model with RMSD < 1A	POS	99.2%	91.5%	75.9%
	JDS	95.9%	89.1%	70.4%

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 4

Average ICC (Twoway, Consistency) in measuring correlation between pair-wise scoring functions

	Average ICC		
	Rosetta-DFIRE	**Rosetta-Triplet**	**DFIRE-Triplet**
Short (4~6)	0.232	0.004	0.011
Medium (7~9 Res)	0.186	0.003	0.011
Long (10~12 Res)	0.149	0.001	0.009