An Improved Statistics-based Backbone Torsion Potential Energy for Protein Loop Structure Modeling

Ionel Rata
National Institute for Physics and Nuclear Engineering (IFIN-HH), R-77125
Bucharest-Magurele, Romania
ionel.rata@nipne.ro

Kyle Wessells
Department of Biological Science
Old Dominion University
Norfolk, VA, USA
kwess002@odu.edu

Yaohang Li
Department of Computer Science
Old Dominion University
Norfolk, VA, USA
yaohang@cs.odu.edu

***Abstract*— Accurately modeling protein loops is critical to predicting three-dimensional structures and understanding functions of many proteins. Local interactions between neighboring residue configurations contribute significantly to forming loop conformations. In this paper, we improve our statistical potential energy function (J. Phys. Chem. B 114(5): 1859-1869, 2010) for loop backbone torsion angle conformations by taking influences from the second nearest neighbor effect (SNNE) into consideration. This is based on our study showing that the second nearest neighbors along a protein sequence still have non-negligible influences on the torsion angles conformation of a loop residue while such correlations from further neighbors are much weakened. A biologically meaningful reference state is also introduced. Accuracy and sensitivity enhancements of the new loop torsion potential energy are observed on a decoy set with 4- to 12-residue loop targets.**

Keywords- Protein Loop Structure, Backbone Torsion Potential, Local Interactions

I. INTRODUCTION

The loop regions in proteins are flexible segments of polypeptides connecting two conserved secondary structure elements, whose sizes and sequences may vary between homologous proteins. Loops play critical roles in stabilizing protein structures and performing important protein functions. Accurately modeling loop conformations is important in a wide variety of structural biology applications, including comparative modeling [1], NMR segment defining [2], protein modeling with Cryo-electron microscopy density map [3], protein design [4], characterizing protein functions [5, 6], and ion channel simulations [7, 8].

The coarse-grained structure of a protein can be described by a reduced representation of its backbone torsion angles ϕ and ψ. From the Ramachandran plots [18], it is well known that the backbone torsion angles ϕ and ψ are strongly correlated to an amino acid type, due to the balance between local and global interactions. Shortle [9] has shown that the $\phi - \psi$ angles propensities of each amino acid type conform to the Boltzmann hypothesis. Consequently, backbone torsion angle statistics can be used to build knowledge-based energy potential functions [10-16], which have demonstrated effectiveness in identifying correct, native-like structures as well as guiding prediction of protein structure when combined with other terms.

Secondary structures such as helices and sheets exhibit regular patterns in backbone torsion angles conformations [17]. Although not as obviously as helices and sheets, residues in loops also demonstrate favorability for certain torsion angle conformations. The distribution of ϕ and ψ angles in a propensity map of a residue reflects the local interactions in forming a loop structure. In [19], we developed a knowledge-based potential energy function according to the propensity maps of residues in a loop, where the nearest neighbor effects (NNE), i.e., structural influences from a residue's adjacent neighbors in sequence, are taken into account. The rationale is based on the fact that the nearest neighbors have the strongest correlations to account for substantial changes in loop structural conformation. Figure 1(a) shows the $\phi - \psi$ propensity map of LYS as a singlet in a loop, where two major clusters centered at (-75, 140) and (-76, -21) and one minor cluster at (60, 42) are observed. The cluster at (-75, 140) is the largest among them. When ASP is the adjacent left neighbor in sequence, as shown in Figure 1(b), the statistical propensity is strongly biased to the cluster at (-76, -21), which agrees with our assumption in [19]. When ASP is one position away, in Figure 1(c), although such bias is significantly weaker, the cluster centered at (-76, -21) is still the largest, suggesting that the second nearest neighbor ASP has non-negligible influences on the $\phi - \psi$ conformation of LYS in loop structure. In contrast, the propensity map of LYS with ASP two positions away in Figure 1(d) is almost indistinguishable to that of LYS as a singlet. This indicates that the influences from neighboring residues further than two positions away can be safely ignored. In conclusion, the second nearest neighboring effects (SNNE) can have non-negligible contributions on torsional angles conformation of a loop residue. Capturing SNNE correlation and incorporating them into the knowledge-based potential can potentially lead to accuracy and sensitivity improvements.

In this paper, we introduce an approach to incorporate SNNE into our loop backbone torsion potential described in [19] to improve its accuracy and sensitivity. Torsion probability density functions that quantify any adjacent $\phi - \psi$ pair distribution in the context of all possible

combinations of local residue types in between the second nearest neighbors are constructed. These torsion probability density functions are used to derive NNE and SNNE terms. Biologically meaningful reference states are also introduced. We benchmark the new loop torsion potential energy on the 4- to 12-residue loop decoys in Jacobson's decoy set [20]. Further analysis and discussions are also provided.

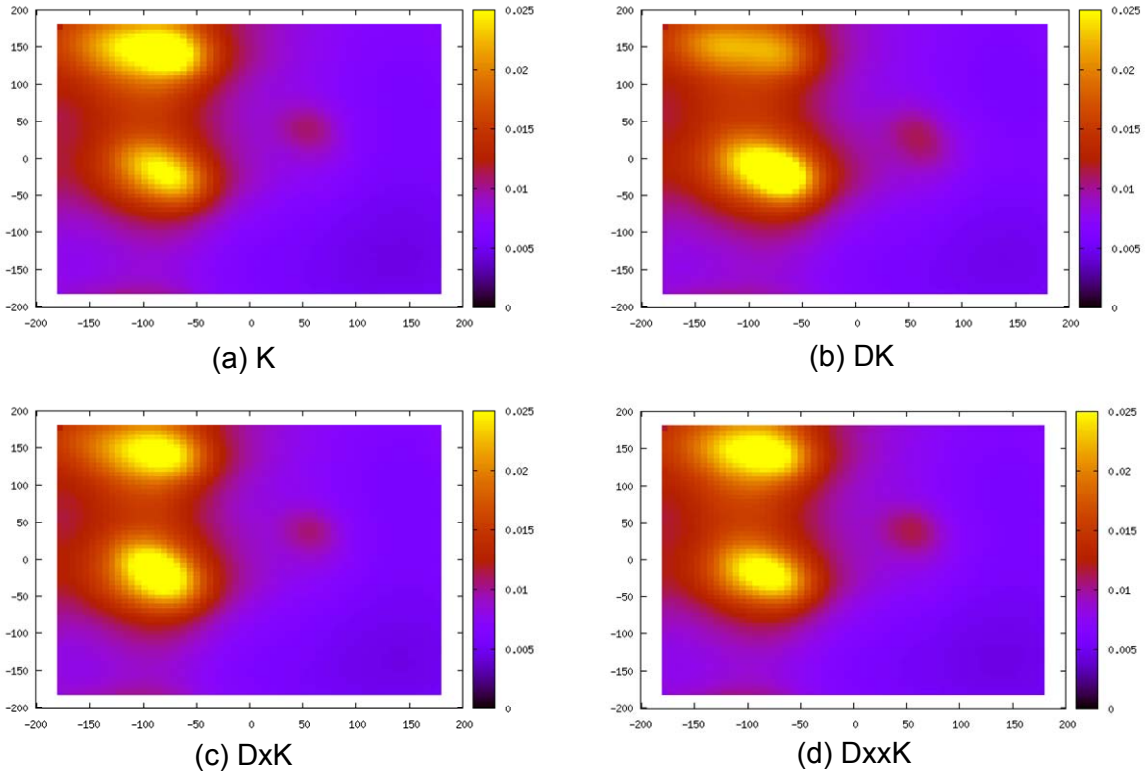


Figure 1: $\phi - \psi$ propensity maps of LYS in the loops in presence of left neighboring ASP. (a): LYS as a singlet. (b, c, d): LYS with ASP as the nearest, one position away, and two positions away neighbors in sequence, respectively. The nearest and second nearest neighbors have strong influences to the backbone torsion angle conformations of LYS and the influences from further neighbors are significantly weakened.

II. METHODS

A. Data Sets

We use the protein chain dataset Cull16633 generated by the PISCES server [21] on 10/21/2011 to collect loop samples to construct torsion probability density functions for our loop torsion potential. Loops in Cull16633 with undetermined structures/residues are ignored. The secondary structure assignment for each residue is determined by the DSSP program [26]. Cull16633 contains 16,633 protein chains with at most 50% sequence identity, 3.0A resolution cutoff, and 1.0 R-factor.

B. Torsion Probability Density Functions with Second Nearest Neighbors

From the observed backbone correlations between the ϕ and ψ torsion angles and the residue types, we capture a direct sequence structure relationship in a loop configuration probability that is an explicit function of sequence residue types and torsion angles, $P(residues, torsions)$. We use this function for evaluating the backbone structures of unknown loops with given sequences, for which the rest of protein structure is known.

We extract local information on sequence-structure correlations from known loops in the Cull16633 and use this information to calculate sequence-structure probabilities. Local information is expressed in the form of Ramachandran-like maps of two successive torsion angles in the context of up to three consecutive residue types. They are then converted to observed occurrence frequencies in the generic forms $P(\phi_i\psi_i|R_i)$, $P(\phi_i\psi_i|R_{i+p}R_i)$, $P(\phi_{i+1}\psi_i|R_iR_{i+1})$, or $P(\phi_{i+1}\psi_i|R_{i+p}R_iR_{i+1})$, where R_{i+p} is a neighboring residue p positions away from R_i along the protein sequence. These are frequency functions in torsion angle and residue type variables as they occur in the native structures, which are used to estimate the probability of a $\phi - \psi$ or $\psi - \phi$ pair adopting a certain conformation under

its neighboring amino acid environment. They are computed for small subsystems that can be assembled with the probabilistic chain rules to estimate the probability of the larger loops to be in native conformations.

A powerful statistical procedure for constructing probability density functions (PDF) from Ramachandran plots has been developed in our previous work [19] and is used here as well. It allows us to generate smooth PDF surfaces even when the Ramachandran maps are very inhomogeneously and sparsely populated.

When SNNE is taken into consideration, the PDF of a $\phi-\psi$ pair within its neighboring amino acid environment is calculated by

$$P(\phi_i\psi_i|R_{i-2}R_{i-1}R_iR_{i+1}R_{i+2}) = \frac{P(\phi_i\psi_i|R_iR_{i-2})P(\phi_i\psi_i|R_iR_{i-1})P(\phi_i\psi_i|R_iR_{i+1})P(\phi_i\psi_i|R_iR_{i+2})}{P(\phi_i\psi_i|R_i)^3}. \quad (1)$$

Similarly, the PDF of a $\psi-\phi$ pair is estimated by

$$P(\phi_{i+1}\psi_i|R_{i-1}R_iR_{i+1}R_{i+2}) = \frac{P(\phi_{i+1}\psi_i|R_{i-1}R_iR_{i+1})P(\phi_{i+1}\psi_i|R_iR_{i+1}R_{i+2})}{P(\phi_{i+1}\psi_i|R_iR_{i+1}),}. \quad (2)$$

C. *Loop Torsion Potential Energy*

In this work, we employ Sippl's potentials of mean force approach [24] to obtain the statistics-based loop torsion potential energy functions. According to the inverse-Boltzmann theorem, the knowledge-based energy potential U is calculated as

$$U = -kTln\frac{P_{obs}}{P_{ref}},$$

where k is the Boltzmann constant, T is temperature, P_{obs} is the observed probability, and P_{ref} is the referenced probability. The observed probability of a torsion pair is calculated by equations (1) or (2). In order to calculate the referenced probability, we also generate the PDFs $P(\phi_i\psi_i|*_i)$, $P(\phi_i\psi_i|R_{i+p}*_i)$, $P(\phi_{i+1}\psi_i|R_{i+p}*_i*_{i+1})$, and $P(\phi_{i+1}\psi_i|R_{i+p}*_i*_{i+1})$ from the known loops in Cull16633, where "*" represents any one of the twenty amino acid types. Then, following Samudrala and Moult's principle [25], the referenced probabilities of a $\phi-\psi$ pair and a $\psi-\phi$ pair with SNNE are estimated as

$$P(\phi_i\psi_i|R_{i-2}R_{i-1}*_iR_{i+1}R_{i+2}) = \frac{P(\phi_i\psi_i|*_iR_{i-2})P(\phi_i\psi_i|*_iR_{i-1})P(\phi_i\psi_i|*_iR_{i+1})P(\phi_i\psi_i|*_iR_{i+2})}{P(\phi_i\psi_i|*_i)^3}$$

and

$$P(\phi_{i+1}\psi_i|R_{i-1}*_i*_{i+1}R_{i+2}) = \frac{P(\phi_{i+1}\psi_i|R_{i-1}*_i*_{i+1})P(\phi_{i+1}\psi_i|*_i*_{i+1}R_{i+2})}{P(\phi_{i+1}\psi_i|*_i*_{i+1})},$$

respectively. Then, the torsion pair potentials with SNNE terms are calculated as

$$U(\phi_i\psi_i) = -kTln\frac{P(\phi_i\psi_i|R_{i-2}R_{i-1}R_iR_{i+1}R_{i+2})}{P(\phi_i\psi_i|R_{i-2}R_{i-1}*_iR_{i+1}R_{i+2})}$$

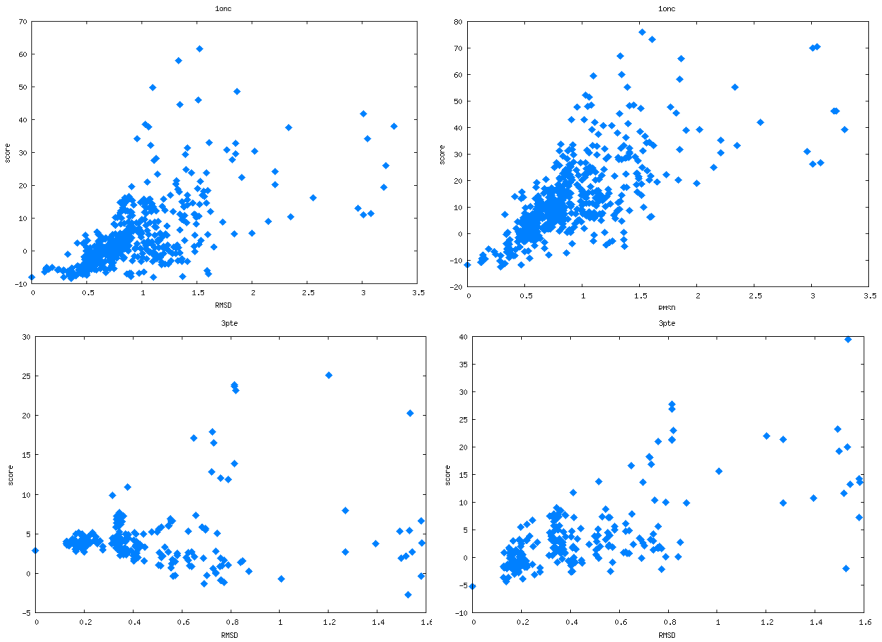
and

$$U(\phi_{i+1}\psi_i) = -kTln\frac{P(\phi_{i+1}\psi_i|R_{i-1}R_iR_{i+1}R_{i+2})}{P(\phi_{i+1}\psi_i|R_{i-1}*_i*_{i+1}R_{i+2})},$$

correspondingly.

Finally, by integrating all torsion pair potentials together, the overall loop torsion potential is calculated as

$$U_{\text{loop}} \cong \sum_i U(\phi_i\psi_i) + \sum_i U(\phi_{i+1}\psi_i).$$



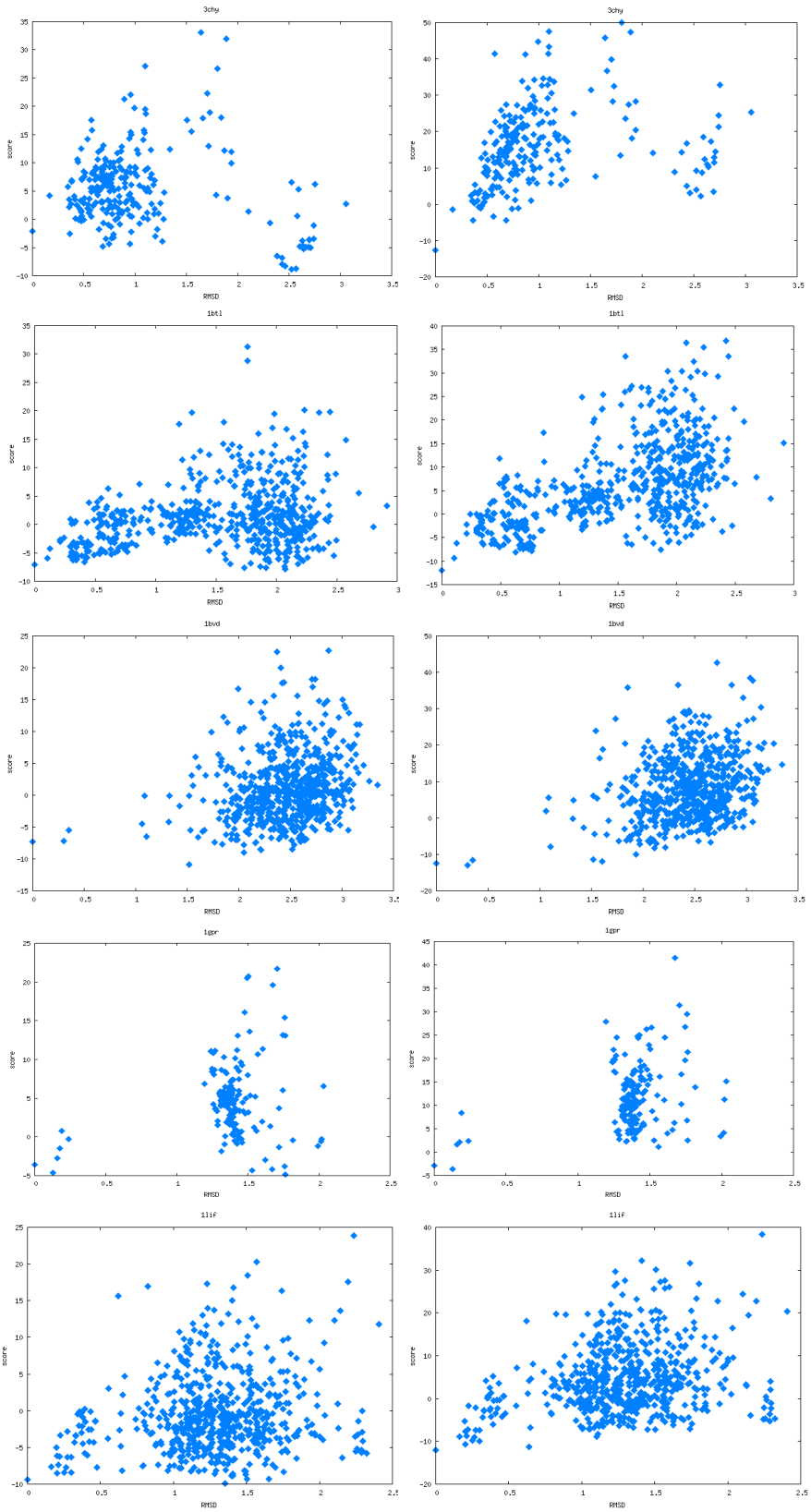


Figure 2: Comparison between potential with NNE terms only (left) and potential with both NNE and SNNE terms (right) on several 9-residue loop targets in Jacobson's loop decoy set.

III. RESULTS

The energy-RMSD (Root Mean Square Deviation) plots shown in Figure 2 illustrate the effectiveness of the loop torsion potential energy on several 9-residue loop targets when influences from the second nearest neighboring residues are incorporated. The left-hand-side plots in Figure 2 use the torsion potential energy with NNE terms only while the right-hand-side ones are based on the potential energy including both NNE and SNNE terms. Clearly, the additional SNNE terms improve the accuracy of the torsion potential, where the native or near-native structures are identified with the lowest energy values. Moreover, the sensitivity of the loop torsion potential is also improved due to SNNE terms, where a decoy cluster with smaller RMSD values typically exhibit lower energy.

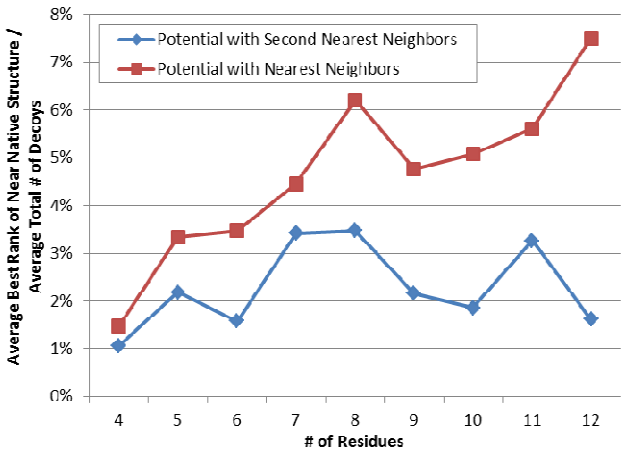


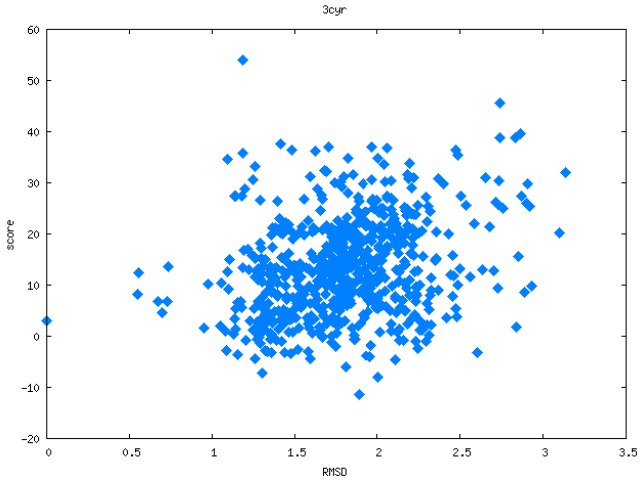
Figure 3: Comparison of ranking effectiveness of loop torsion potential with NNE terms only and the one with both NNE and SNNE terms on 4- to 12-residue loop targets in Jacobson's loop decoy set

Similar improvements in the loop torsion potential due to SNNE are also found in loop targets of other lengths. Figure 3 compares the average ranking of near-native (RMSD < 0.5A) on 4- to 12-residue loops in Jacobson's loop decoy set [20] when the torsion potentials with NNE terms only and the one with both NNE and SNNE terms are used. One can find that the loop torsion potential with NNE and SNNE terms yields better decoy discrimination capability than the one with NNE terms only. More interestingly, the loop torsion potential with both NNE and SNNE terms yields consistent performance as the loops get longer, where the average best ranking percentage of the near-native structures are kept under 4%.

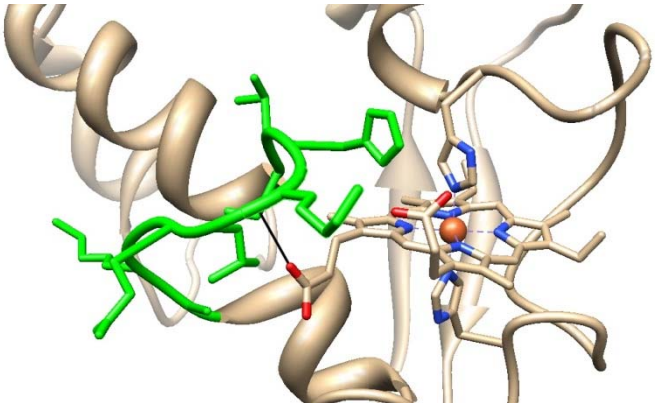
IV. DISCUSSIONS

Our loop torsion potential only accounts for local interactions from neighbor residues along the protein sequence. However, when strong global interactions are involved in loop structure formation, the loop torsion potential may become insufficient. Figure 4 shows the energy-RMSD plot of a 10-residue loop, 3CYR(70-79), where the native and near-native structures yield much higher energy than many decoys with RMSD ranging from 1.5A to 2.5A. Further analysis shows that the LYS(75) residue in 3CYR(70-79) has strong interaction with an HEM (Protoporphyrin IX) ligand containing a Fe atom, as shown in Figure 4(b), which accounts for substantial changes in the overall structure of loop 3CYR(70-79). Therefore, our loop torsion potential is ineffective on 3CYR(70-79) decoys since only local interactions are evaluated.

After all, in many protein loops, local interactions play dominating roles and thus improved accuracy in local interactions contributes to better modeling of loop structures. Combined with other potentials with global interaction terms, our loop torsion potential has been demonstrated effective in discriminating correct loop models [22] as well as predicting loop structures with high resolution [23].



(a) Energy-RMSD plot of 3CYR (70-79)



(b) Structural analysis of 3CYR (70-79) loop

Figure 4: The torsion potential has poor performance in 3CYR (70-79) due to the fact that Lys75 has strong interaction with the HEM ligand containing a Fe atom

V. SUMMARY

In this paper, we show that the second nearest neighboring residues still have non-negligible correlations with the backbone $\phi - \psi$ conformation of a loop residue. Hence, we improve our statistical loop torsion potential energy by incorporating SNNE terms. Accuracy and sensitivity enhancements are observed in 4- to 12-residue loop targets in Jacobson's decoy set [20].

Our loop torsion potential uses a reduced representation of protein loop structure, which is particularly suitable for coarse-grained loop structure modeling or loop sequence design. The loop torsion potential is effective on loops where the local interactions play a dominant role. When strong global interactions are present, using the loop torsion potential only may be insufficient. Nevertheless, together with other potential energy functions for global effects, our loop torsion potential also shows effectiveness in discriminating correct loop models [22] as well as predicting loop structures with high resolution [23].

ACKNOWLEDGEMENTS

This work is partially supported by NSF grant 1066471 to YL, ODU 2013 Multidisciplinary Seed grant to YL, and ODU Undergraduate Research Grant to KW.

REFERENCES

1. R. E. Bruccoleri, "Ab initio loop modeling and its application to homology modeling," Methods in Molecular Biology, 143: 247-264, 2000.
2. O. Y. Dmitriev, R. H. Fillingame, "The rigid connecting loop stabilizes hairpin folding of the two helices of the ATP synthase subunit c," Protein Sci., 16(10): 2118-2122, 2007.
3. J. Zhu, L. Cheng, Q. Fang, Z. H. Zhou, B. Honig, "Building and refining protein models with Cryo-electron microscopy density maps based on homology modeling and multi-scale structure refinement," J. Mol. Biol., 397(3): 835-851, 2010.
4. A. C. Martin, J. C. Cheetham, A. R. Rees, "Modeling antibody hypervariable loops: a combined algorithm," Proc. Nat. Acad. Sci., 86(23): 9268-9272, 1989.
5. S. Jones, J. M. Thornton, "Prediction of protein-protein interaction sites using patch analysis," J. Mol. Biol., 272: 133-143, 1997.
6. J. S. Fetrow, A. Godzik, J. Skolnick, "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., 282: 703-711, 1998.
7. A. Tasneem, L. M. Iyer, E. Jakobsson, L. Aravind, "Identification of the prokaryotic ligand-gated ion channels and their implications for the mechanisms and origins of animal Cys-loop ion channels," Genome Biol., 6(1): R4, 2005.
8. V. Yarov-Yarovoy, D. Baker, W. A. Catterall, "Voltage sensor conformations in the open and closed states in ROSETTA structural models of K+ channels," Proc. Natl. Acad. Sci., 103: 7292-7297, 2006.
9. D. Shortle, "Propensities, probabilities, and the Boltzmann hypothesis," Protein Sci., 12(6): 1298-1302, 2003.
10. D. Shortle, "Composites of local structural propensities: Evidence for local encoding of long range structure," Protein Sci., 11: 18–26, 2002.
11. O. Keshin, D. Yuret, A. Gursoy, M. Turkay, B. Erman, "Relationships between amino acid sequence and backbone torsion angle preferences," Proteins, 55: 992-998, 2004.
12. S. C. E. Tosatto, R. Battistutta, "TAP score: torsion angle propensity normalization applied to local protein structure evaluation," BMC Bioinformatics, 8: 155, 2007.
13. E. D. Amir, N. Kalis, C. Keasar, "Differentiable, multi-dimensional, knowledge-based energy terms for torsion angle probabilities and propensities," Proteins, 72(1): 62-73, 2008.
14. Y. Dehouck, D. Gilis, M. Rooman, "A new generation of statistical potentials for proteins," Biophysical J., 90(11): 4010-4017, 2006.
15. Q. Fang, D. Shortle, "A consistent set of statistical potentials for quantifying local side-chain and backbone interactions," Proteins, 60(1): 90-96, 2005.
16. J. E. Fitzgerald, A. K. Jha, A. Colubri, T. R. Sosnick, K. F. Freed, "Reduced Cβ statistical potentials can outperform all-atom potentials in decoy identification," Protein Science, 16(10): 2123-2139, 2007.
17. T. E. Creighton, (1996). Proteins: Structures and Molecular Properties (2nd edit.). W. H. Freeman: New York.
18. G. N. Ramachandran, V. Sasisekharan, "Conformation of polypeptides and proteins," Advan. Protein Chem., 23: 283–438, 1968.
19. I. Rata, Y. Li, E. Jakobsson, "Backbone Statistical Potential from Local Sequence-Structure Interactions in Protein Loops," Journal of Physical Chemistry B, 114(5): 1859-1869, 2010.
20. 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, 55: 351-367, 2004.
21. G. Wang, R. Dunbrack, "PISCES: a protein sequence culling server," Bioinformatics, 19(12): 1589-1591, 2003.
22. Y. Li, I. Rata, S. Chiu, E. Jakobsson, "Improving Predicted Protein Loop Structure Ranking using a Pareto-Optimality Consensus Method," BMC Structural Biology, 10: 22, 2010.
23. Y. Li, I. Rata, E. Jakobsson, "Sampling Multiple Scoring Functions Can Improve Protein Loop Structure Prediction Accuracy," Journal of Chemical Information and Modeling, 51(7): 1656-1666, 2011.
24. M. J. Sippl, "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, 1990.
25. R. Samudrala, J. Moult, "An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction," Journal of Molecular Biology, 275: 895-916, 1998.
26. W. Kabsch, C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen bonded and geometrical features," Biopolymers, 22: 2577-2637, 1983.