A Coarse-grained, Context-dependent Contact Potential for Protein Decoy Discrimination

Yaohang Li
Department of Computer Science
Old Dominion University
yaohang@cs.odu.edu

Keywords: Protein Structure Modeling, Knowledge-based Potential, Inter-residue Contacts

Abstract

Knowledge-based contact potentials have been popularly used in protein structure modeling. In this paper, we present a coarse-grained, context-dependent contact potential where the influences on inter-residue contacts from neighboring residues are taken into account. In particular, statistics of singlet-singlet, singlet-doublet, singlet-triplet, and doublet-doublet interactions are derived from known protein structures and then incorporated into the context-dependent contact potential. We use the Rosetta decoy set to test our context-dependent contact potential. In more than 83% targets, the context-dependent contact potential is able to successfully differentiate the near-native structures from the other decoys. The context-dependent contact potential yields comparable accuracy to the fine-grained contact potentials, but does not require all atoms information, which is particularly suitable for coarse-grained protein structure modeling and design.

1. Introduction

Amino acids are the basic structural building blocks of proteins. These amino acids exhibit very different physical-chemical properties, which form complicated contact preferences among each other. Inter-residue contacts control the stability of protein structure as well as their biological functions.

The formation of inter-residue contacts in proteins can be of different nature. First of all, inter-residue contacts can result from chemical bonds. For residues forming regular secondary structures such as α-helices and β-strands, hydrogen bonds are built between contacting residues [1]. Ionic bonds (salt bridges) are formed as residues bearing oppositely charged groups are placed together in the hydrophobic core of proteins. Side chains of Cysteine residues can establish a disulfide bridge, a covalent bond coupling two thiol groups [2]. The non-bonding inter-residue interactions also play an important role in formation of inter-residue contacts. The non-polar hydrophobic interactions cluster the hydrophobic residues to form the hydrophobic association shielded from interactions with solvent [3]. When atoms in a protein are packed closely enough to each other, van der Waals attractions start to take effect. Although van der Waals interaction is relatively weak compared to the bonded forces, the large number of van der Waals interactions occurred in large protein molecules contribute significantly to the folding of proteins in forming essential contacts [4]. Analysis of inter-residue contacts is important to understand protein structure stability, protein-protein interactions, and protein folding mechanisms.

The knowledge-based contact potentials (energies) are the most popularly used tools to investigate inter-residue contacts. Although the scope and limitation of the contact potentials are still vigorously debated and disputed [19-22], contact potentials have been successfully used in a variety of applications, including fold recognition [10], protein structure prediction [18], protein design [23], and protein-protein docking [24]. The fundamental idea of contact potentials is to derive statistical contact preferences from experimentally determined protein structures presented in Protein Data Banks (PDB) and then estimate the corresponding pseudo-potentials with respect to the defined reference states. Recently, quite a few new ideas have been proposed to enhance the contact potential. Alternative reference states are defined. Zhang et al. [11] developed random crystal reference states to remove compositional bias. Skolnick et al. [12] proposed reference states accounting for the constraints of chain connectivity and compactness. More precise contact definitions are also involved. Berrera et al. [13] found that the definition of contact based on van der Waals radii of Cα and heavy atoms in side chain yields better performance than other contact definitions. Reck and Vaisman [15] employed contact definition based on Delaunay tessellation. Zhang et al. [11] extended the Miyazawa-Jernigan procedure [6-9] to atom-atom contact potentials. McConkey et al. [14] incorporated calculations of solvent accessible surfaces into atom-atom contact potentials. Moreover, the amino acid environment is also taken into consideration. Zhang and Kim [16] showed that the strength of inter-residue contacts has strong correlation with its secondary structural environment. Duan and Zhou [17] considered residue hydrophobic environment in their contact potentials.

Nowadays, the number of experimentally-determined protein structures and sequences deposited in PDB continues to increase stably by around eight thousand a year [31]. The number of available inter-residue contact samples also increase accordingly, which provides a rich

information source to study the dependence of inter-residue contacts on their amino acid environment in a more precise manner than before. In this paper, we present a context-dependent contact potential where the surrounding amino acid environment of the contacting residues pair is explicitly taken into account. In particular, we are interested in a coarse-grained contact potential where only the information of the protein backbone and Cβ atoms is required, which can be conveniently used in coarse-grained protein structure modeling or protein design. The effectiveness of our context-dependent contact potential is tested on the Rosetta decoy set [25].

The rest of the paper is organized as follows. Section 2 describes our approaches of generating the context-dependent contact potential. Section 3 presents the computational results. Section 4 concludes our work.

2. Methods

2.1 Data Set

We use the protein chain dataset Cull16633 generated by the PISCES server [27] on 10/21/2011 to collect inter-residue contact samples to generate context-based statistics. Cull16633 contains 16,633 chains with at most 50% sequence identity, 3.0A resolution cutoff, and 1.0 R-factor. The Rosetta decoy set [25] is used to test the effectiveness of our context-dependent contact potential function. To ensure correctness of this work, the identities of target proteins in Rosetta decoy set are excluded from Cull16633 when the context-based statistics are derived.

2.2 Definition of Contacts

In literature, there are various definitions for inter-residue contacts, including measuring distances between Cα atoms, Cβ atoms, side chain centroids, van der Waals centers, or shortest distances between side chain heavy atoms. Accordingly, various distance cutoffs have been employed as well. In this paper, we adopt a simple definition of inter-residue contact to build a coarse-grained potential – a pair of residues separated by at least 9 positions in sequence is in contact if their distance d_{cutoff} between Cβ atoms is less than 6.5A.

2.3 Influences from Sequentially Neighboring Residues

It is well known that there exist general short range regularities in the primary structure of proteins [28]. Presumably, the sequentially neighboring residues have strong and probably deterministic influence to the chemical property of a pair of residues in forming inter-residue contact. Figure 1 shows the probability of two amino acids in contact when the hydrophobic residues (ILE and VAL) or hydrophilic residues (ARG and ASP) are presented as neighbors at different sequence distances ($i \pm 1$, $i \pm 2$, $i \pm 3$, and $i \pm 4$) in one of the contacting residues. One can find that the contact probability of two amino acids exhibits strong correlations with the types of residues at the neighboring positions. In most cases, the contact probability of two amino acids with hydrophobic neighbors is negatively correlated with the probability of those with hydrophilic neighbors. Although such correlations are weaker for neighbors at further sequence positions, they are non-negligible to build an accurate contact potential.

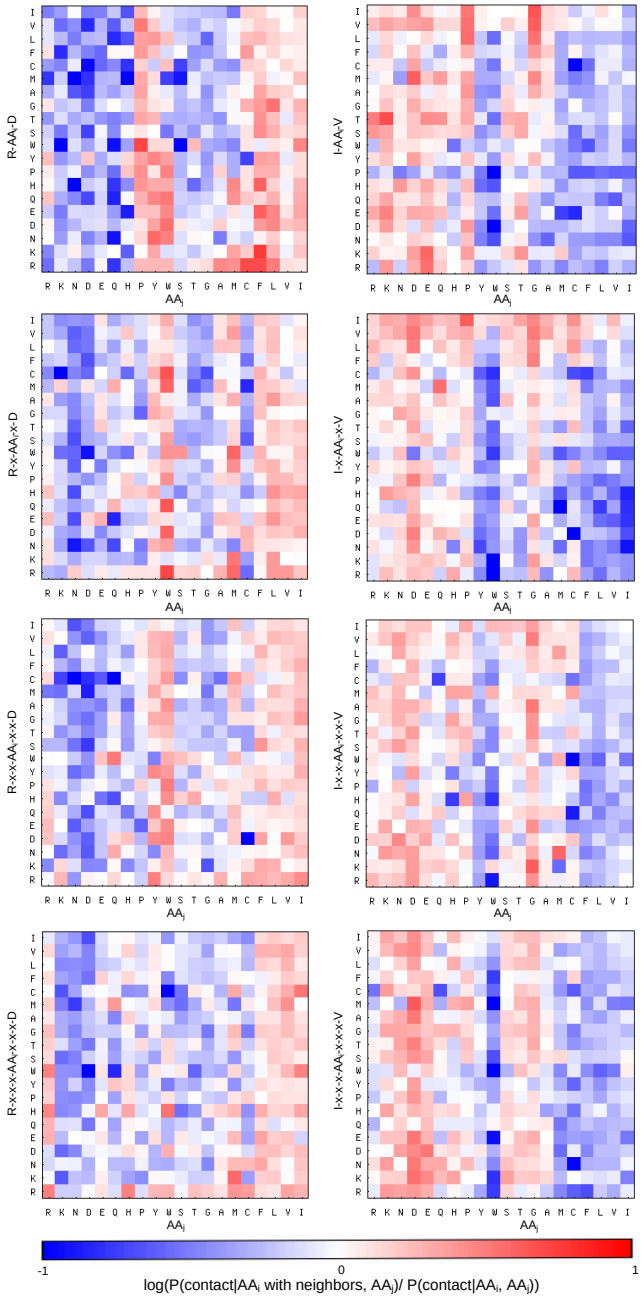
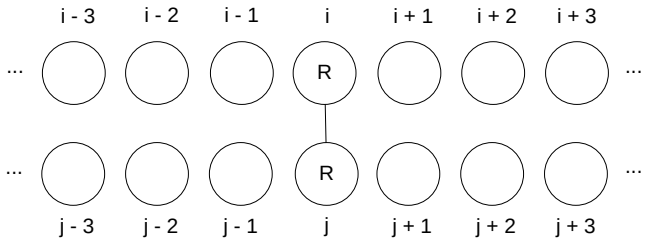


Figure 1: Probabilities of two amino acids in contact when the hydrophobic residues (ILE and VAL) or hydrophilic residues (ARG and ASP) are presented as neighbors at different sequence distances ($i \pm 1$, $i \pm 2$, $i \pm 3$, and $i \pm 4$)

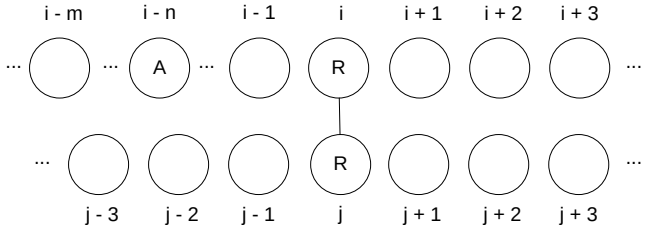
2.4 Context-based Statistics

In this paper, we adopt the potentials of mean force [27] approach to estimate the favorability of a pair of contacting residues within their amino acid environment. The mean-

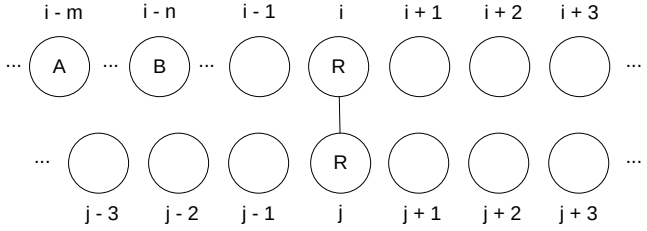
force potential is derived based on the statistics of correlations between the residues in contact and its nearby neighbors. In particular, the increasing number of experimentally determined protein structures in PDB recently has provided sufficient number of samples to enable derivation of statistics while neighboring residues are taken into account. In our method, we obtain statistics of singlet-singlet, singlet-doublet, singlet-triplet, and doublet-doublet interactions from the protein chains listed in Cull16633. Figures 2(a), 2(b), 2(c), and 2(d) illustrate singlet-singlet, singlet-doublet, singlet-triplet, and doublet-doublet contacts, respectively. Currently, the number of available protein structures in PDB is insufficient to derive meaningful higher order statistics.



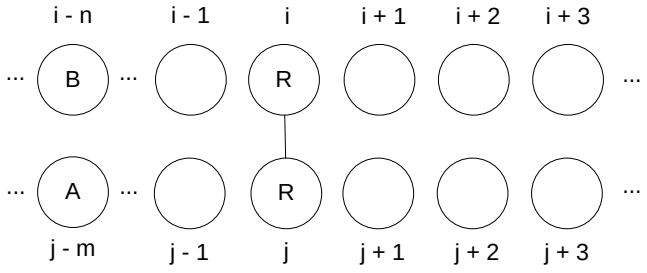
(a) Singlet-Singlet Interaction



(b) Singlet-Doublet Interaction



(c) Singlet-Triplet Interaction



(d) Doublet-Doublet Interaction

Figure 2: Singlet-singlet, singlet-doublet, singlet-triplet, and doublet-doublet interactions

2.5 Context-dependent Contact Potential

The context-dependent contact potential is generated based on the potentials of mean force method [26]. According to the inverse-Boltzmann theorem, we calculate the mean-force potential $U(R_i, R_j, Contact)$ to treat the interaction between singlet residues R_i and R_j in forming an inter-residue contact,

$$U_{singlet-singlet}(R_i, R_j, Contact) = -RTln\frac{P_{obs}(Contact|R_i, R_j)}{P_{ref}(Contact|R_i, R_j)}.$$

Here R is the gas constant and T is the temperature. $P_{obs}(Contact|R_i, R_j)$ is the observed probability of R_i-R_j contact, which is estimated by the fraction of R_i-R_j contacts among all observed inter-residue contacts, i.e.,

$$P_{obs}(Contact|R_i, R_j) = \frac{N_{contact}(R_i, R_j)}{\sum_{i'}\sum_{j'} N_{contact}(R_{i'}, R_{j'})},$$

where $N_{contact}(R_i, R_j)$ is the total number of observed R_i-R_j contacts. $P_{ref}(Contact|R_i, R_j)$ is the referenced probability, which is estimated by

$$P_{ref}(Contact|R_i, R_j) = \frac{N(R_i)N(R_j)}{(\sum_{i'} N(R_{i'}))^2},$$

where $N(R_i)$ is the total number of observed residues R_i in the protein database.

Similarly, when neighboring residues are taken into consideration, the mean-force potentials of singlet-doublet ($U_{singlet-doublet}$), singlet-triplet ($U_{singlet-triplet}$), and doublet-doublet ($U_{singlet-triplet}$) interactions are calculated as

$$U_{singlet-doublet}(R_i, R_jR_{j+k}, Contact) = -RTln\frac{P_{obs}(Contact|R_i, R_jR_{j+k})}{P_{ref}(Contact|R_i, R_jR_{j+k})} + RTln\frac{P_{obs}(Contact|R_i, R_j)}{P_{ref}(Contact|R_i, R_j)},$$

$$U_{singlet-triplet}(R_i, R_jR_{j+k_1}R_{j+k_2}, Contact) = -RTln\frac{P_{obs}(Contact|R_i, R_jR_{j+k_1}R_{j+k_2})}{P_{ref}(Contact|R_i, R_jR_{j+k_1}R_{j+k_2})} + RTln\frac{P_{obs}(Contact|R_i, R_jR_{j+k_1})}{P_{ref}(Contact|R_i, R_jR_{j+k_1})} + RTln\frac{P_{obs}(Contact|R_i, R_jR_{j+k_2})}{P_{ref}(Contact|R_i, R_jR_{j+k_2})} - RTln\frac{P_{obs}(Contact|R_i, R_j)}{P_{ref}(Contact|R_i, R_j)},$$

and

$$U_{doublet-doublet}\left(R_iR_{i+k_1}, R_jR_{j+k_2}, Contact\right) = -RTln\frac{P_{obs}\left(Contact|R_iR_{i+k_1}, R_jR_{j+k_2}\right)}{P_{ref}\left(Contact|R_iR_{i+k_1}, R_jR_{j+k_2}\right)} + RTln\frac{P_{obs}\left(Contact|R_iR_{i+k_1}, R_j\right)}{P_{ref}\left(Contact|R_iR_{i+k_1}, R_j\right)} + RTln\frac{P_{obs}\left(Contact|R_i, R_jR_{j+k_2}\right)}{P_{ref}\left(Contact|R_i, R_jR_{j+k_2}\right)} - RTln\frac{P_{obs}\left(Contact|R_i, R_j\right)}{P_{ref}\left(Contact|R_i, R_j\right)},$$

respectively. Accordingly, the reference probabilities for doublets and triplets are

$$P_{ref}\left(Contact|R_i, R_jR_{j+k}\right) = \frac{N(R_i)N(R_jR_{j+k})}{\sum_{i'}N(R_{i'})\sum_{j'}N(R_{j'}R_{j'+k})}$$

and

$$P_{ref}\left(Contact|R_i, R_jR_{j+k_1}R_{j+k_2}\right) = \frac{N(R_i)N\left(R_jR_{j+k_1}R_{j+k_2}\right)}{\sum_{i'}N(R_{i'})\sum_{j'}N\left(R_{j'}R_{j'+k_1}R_{j'+k_2}\right)}.$$

Finally, by integrating all interactions together, the overall contact potential is calculated as

$$U_{protein} \cong \sum_{protein} U_{singlet-singlet} + \sum_{protein} U_{singlet-doublet} + \sum_{protein} U_{singlet-triplet} + \sum_{protein} U_{doublet-doublet}.$$

3. Results

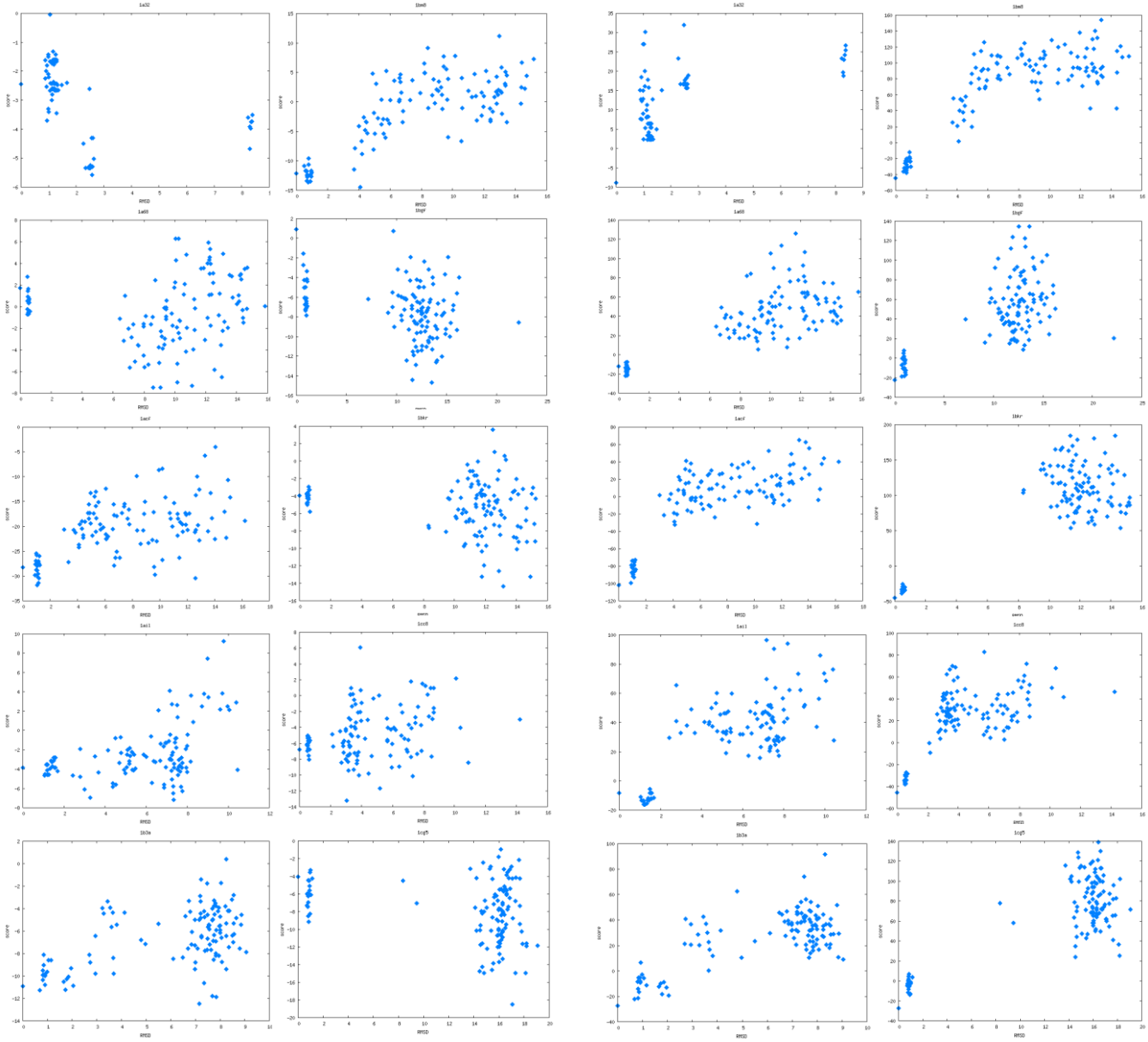
Table 1 shows the accuracy of the context-dependent contact potential on the Rosetta decoy set [25]. For each target, the decoy set includes 100 predicted models generated by the Rosetta program [29, 30] as well as 20 "relaxed" models generated by refining the native structure by the all-atom Rosetta program. The structures of the relaxed models are typically near to the native. Using the relaxed models can avoid the problem of potentially biasing to the native structures in knowledge-based potentials since the statistics of the contact potential is derived from the native structures.

Protein	Chain	RMSD (A) of Top Decoy	Best RMSD (A) of Top 5 Decoys	Rank of Native
1a19	A	11.79	11.16	3
1a32		1.14	1.00	1
1a68		0.47	0.47	17
1acf		0.90	0.90	1
1ail		1.23	1.19	24
1aiu		1.53	1.08	5
1b3a	A	0.68	0.68	1
1bgf		0.72	0.72	1
1bk2		7.04	7.04	33
1bkr		0.47	0.47	1
1bm8		0.71	0.51	1
1bq9	A	3.46	3.46	57
1c8c	A	0.59	0.59	4
1c9o	A	2.57	2.32	7
1cc8	A	0.53	0.51	1
1cei		10.70	10.23	20
1cg5	B	0.84	0.77	1
1ctf		1.21	0.93	15
1dhn		0.81	0.81	1
1e6i	A	0.84	0.83	1
1elw	A	0.65	0.57	1
1enh		2.26	0.74	31
1ew4	A	1.05	0.77	1
1eyv	A	1.68	1.31	2
1fkb		0.64	0.64	13
1fna		0.73	0.73	1
1gvp		2.06	1.24	1
1hz6	A	0.78	0.78	1
1ig5	A	7.48	3.14	6
1iib		3.37	0.88	2
1kpe	A	0.75	0.73	2
1lis		1.27	1.22	1
1lou	A	13.11	6.17	83
1nps		0.60	0.59	1
1opd		0.49	0.47	4
1pgx		0.77	0.77	11
1ptq		0.81	0.81	1
1r69		1.76	1.42	1
1rnb	A	0.71	0.54	5
1scj	B	7.70	6.99	104
1shf	A	0.70	0.65	3
1ten		0.58	0.58	8
1tif		1.07	0.68	3
1tig		0.70	0.70	1
1tul		0.69	0.64	1
1ubi		4.56	2.68	4
1ugh	I	1.00	0.99	4
1urn	A	1.19	0.69	9
1utg		1.30	0.96	2
1vcc		1.01	0.96	2
1vie		0.44	0.42	1
1vls		1.91	1.43	1
1who		11.73	11.73	120
256b	A	1.40	1.24	1
2acy		0.63	0.53	6
2chf		0.61	0.61	1
2ci2	I	0.63	0.63	20
4ubp	A	1.00	0.96	1
5cro	A	0.69	0.60	4

Table 1: Accuracy of the context-dependent contact potential on protein targets in Rosetta decoy set

The context-dependent contact potential is effective in discriminating decoys in good quality in the Rosetta decoy set. In 49 (83%) out of the 59 targets, the top-ranked decoys have Root Mean Square Deviation (RMSD) values less than 3A. Also, in 51 (86%) out of the 59 targets, the best RMSD values of the top-5 ranked decoys are less than 3A. Moreover, when the native structures are mixed into the decoy sets, in 52 (88%) out of the 59 targets, the natives are ranked in top 20. The accuracy of our context-dependent contact potential on the Rosetta decoy set is comparable to that of the all-atom contact scoring function [14]; however, our context-dependent potential is coarse-grained, where only information of Cβ atoms is required.

Figure 3 compares the performance of pair-wise contact potential and context-dependent potential on targets 1a68, 1a32, 1acf, 1ail, 1b3a, 1bgf, 1bm8, 1bkr, 1cc8, and 1cg5 in the Rosetta decoy set. Clearly, when influences of the neighboring residues are taken into account, the context-dependent contact potential is not only more accurate but also more sensitive than the pair-wise contact potential.



(a) Pair-wise Contact Potential (b) Context-Dependent Contact Potential

Figure 3: Comparison of pair-wise contact potential and context-dependent contact potential on decoys for targets of 1a68, 1a32, 1acf, 1ail, 1b3a, 1bgf, 1bm8, 1bkr, 1cc8, and 1cg5 in the Rosetta decoy set

4. Conclusions

In this paper, a coarse-grained, context-dependent contact potential integrating statistics of singlet-singlet, singlet-doublet, singlet-triplet, and doublet-doublet interactions is presented. This context-dependent contact potential has demonstrated its effectiveness on the Rosetta decoy set, where in more than 83% targets, the near-native decoys are differentiated from the other decoys as the top-ranked models. The accuracy of our context-dependent contact potential is comparable to the other fine-grained contact potentials, but only information of backbone atoms is needed, which is suitable for coarse-grained protein structure modeling and protein design.

Acknowledgements

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

References

[1] E. N. Baker, R. E. Hubbard, "Hydrogen bonding in globular proteins," Prog. Biophys. Mol. Biol., 44: 97-179, 1984.

[2] C. Branden, J. Tooze, "Introduction to protein structure," Garland publishing, 1999.

[3] K. A. Dill, "Dominant forces in protein folding," Biochemistry, 29: 7133-7155, 1990.

[4] R. L. Baldwin, "Weak interactions in protein folding: hydrophobic free energy, van der Waals interactions, peptide hydrogen bonds, and peptide solvation," Protein Folding Handbook, Wiley-CCH, 2005.

[5] S. Tanaka, H. A. Scheraga, "Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins," Macromolecules, 9: 945-950, 1976.

[6] S. Miyazawa, R. L. Jernigan "Estimation of effective interresidue contact energies from protein crystal structures: quasichemical approximation," Macromolecules, 18: 534-552, 1985.

[7] S. Miyazawa, R. L. Jernigan, "Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading," J. Mol. Biol., 256: 623-644, 1996.

[8] S. Miyazawa, R. L. Jernigan, "Self-consistent estimation of interresidue protein contact energies based on an equilibrium mixture approximation of residues," Proteins, 34: 49-68, 1999.

[9] S. Miyazawa, R. L. Jernigan, "An empirical energy potential with a reference state for protein fold and sequence recognition," Proteins 36: 357-369, 1999.

[10] B. H. Park, M. Levitt, "Energy functions that discriminate X-ray and near-native folds from well-constructed decoys," J. Mol. Biol., 258: 367-392, 1996.

[11] C. Zhang, G. Vasmatzis, J. L. Cornette, C. Delisi, "Determination of atomic desolvation energies from the structures of crystallized proteins," J. Mol. Biol., 267: 707-726, 1997.

[12] J. Skolnick, L. Jaroszweski, A. Kolinski, A. Godzik, "Derivation and testing of pair potentials for protein folding, when is the quasichemical approximation correct?" Protein Science, 6: 676-688, 1997.

[13] M. Berrera, H. Molinari, F. Fogolari, "Amino acid empirical contact energy definitions for fold recognition in the space of contact maps," BMC Bioinformatics, 4: 8, 2003.

[14] B J. McConkey, V. Sobolev, M. Edelman, "Discrimination of native protein structures using atom-atom contact scoring," Proc. Natl. Acad. Sci. USA, 100(6): 3215-3220, 2003.

[15] G. M. Reck, I. I. Vaisman, "Decoy discrimination using contact potentials based on Delaunay tessellation of hydrated proteins," Proceedings of 4th International Symposium on Voronoi Diagrams in Science and Engineering, 2007.

[16] C. Zhang, S. H. Kim, "Environment-dependent residue contact energies for proteins," Proc. Natl. Acad. Sci. USA, 97(6): 2550-2555, 2000.

[17] M. J. Duan, Y. H. Zhou, "A contact energy function considering residue hydrophobic environment and its application in protein fold recognition," Geno. Prot. Bioinfo., 3(4): 218-224, 2005.

[18] S. Wu, A. Szilagyi, Y. Zhang, "Improving protein structure prediction using multiple sequence-based contact predictions," Structure, 19: 1182-1191. 2011.

[19] M. Vendruscolo, E. Domany, "Pairwise contact potentials are unsuitable for protein folding," J. Chem. Phys., 109:11101-11108, 1998.

[20] M. Vendruscolo, R. Najmanovich, E. Domany, "Can a pairwise contact potential stabilize native protein folds against decoys obtained by threading?" Proteins, 38: 134-148, 2000.

[21] L. Mirny, E. I. Shakhnovich, "How to derive a protein folding potential? A new approach to an old problem" J. Mol. Biol., 264: 1164–1179, 1996.

[22] J. Khatun, S. D. Khare, N. V. Dokholyan, "Can contact potentials reliably predict stability of proteins?" J. Mol. Biol., 336(5): 1223-1238, 2004.

[23] W. P. Russ, R. Ranganathan, "Knowledge-based potential functions in protein design," Current Opinion in Structural Biology, 12(4): 447-452, 2002.

[24] D. V. S. Ravikant, R. Elber, "Energy design for protein-protein interactions," J. Chem. Phys., 135: 065102, 2011.

[25] J. Tsai, R. Bonneau, A. V. Morozov, B. Kuhlman, C. A. Rohl, D. Baker, "An improved protein decoy set for testing energy functions for protein structure prediction," Proteins, 53(1): 76-87, 2003.

[26] G. Wang, R. Dunbrack, "PISCES: a protein sequence culling server," Bioinformatics, 19(12): 1589-1591, 2003.

[27] 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." J. Mol. Biol., 213: 859-883, 1990.

[28] F. Vonderviszt, G. Matrai, I. Simon, "Characteristic sequential residue environment of amino acids in proteins," Int. J. Peptide Protein Res., 27: 483-492, 1986.

[29] K. T. Simons, R. Bonneau, I. Ruczinski, and D. Baker, "Ab initio Protein Structure Prediction of CASP III Targets Using ROSETTA", Proteins: Structure, Function and Genetics, 37(3): 171-176, 1999.

[30] D. Baker, "A surprising simplicity to protein folding," Nature, 405: 39-42, 2000.

[31] M. Levitt, "Growth of novel protein structural data," Proc. Natl. Acad. Sci. USA, 104(9): 3183-3188, 2007.