Use of Amino Acid-Nucleotide Base Pair Potentials in Screening Protein-DNA Docked Complexes

Dongmin Liu, Shan Chang, Jian Chen, and Xuhong Tian

***Abstract*—Amino acid-nucleotide base pair potentials are used to screen docked complexes generated by DOT. The pair potential algorithm designed in this paper is applied to screening 10 systems selected from protein-DNA benchmark set. For all the systems, a correct docking was placed within the top 6% of the pair potential score ranked complexes. Also, over 60% correct answers rank in the top 10% of the docked results for most of the systems.**

***Index Terms*—Amino acid-nucleotide base pair potentials, protein-DNA dock, screening.**

I. Introduction

Protein-DNA interactions regulate many cellular processes involving gene expression, DNA replication and repair [1]. Since DNA play very important roles in cells, they are molecular targets of many clinically used drugs, such as anticancer drugs and antibiotics [2]. Study on the protein-DNA interactions would be meaningful for drugs design on the nucleic acids. However, the determination of the protein-DNA structure is still difficult through the biochemistry experiment directly. Until January 5, 2013, there are over 87,000 structures deposited in the PDB (Protein Data Bank) [3], where the number of protein-NA structure is just almost 4,000. Hence, computational techniques such as protein-DNA docking will become an increasingly important way to help understand the molecular mechanisms of biological systems [4].

Although a number of approaches are available for predicting the docking of protein-protein, the treatment of docking of small ligands and proteins with other proteins, the treatment of the docking of nucleic acids onto proteins lags far behind [5]. Two particular problems have hampered the development of efficient docking methods: the sparsity of the information to define the DNA-binding interface and the inherent flexibility of DNA [6]. Nonetheless, study groups such as Takeda *et al*. [7] and Liu *et al*. [8] have already made achievements on modeling protein–DNA interactions using different computational methods. However, these treatments have never applied amino acid-nucleotide base pair potentials on the docking of protein-DNA even if empirical residue-residue pair potentials [9]-[11] have already played an important role in protein-protein dockings for a long time.

II. Materials and Methods

A. The Program DOT

The program DOT [12], which uses the sum of both a Poisson-Boltzmann electrostatic energy and a van der Waals energy as its energy function, quickly finds low-energy docked structure for two macromolecules by performing a systematic search over six degrees of freedom. A major objective of the DOT program is to provide a method that is fast enough for routine use and cheap enough to be used in highly speculative modes.

In our earlier work, we applied DOT to dock 10 systems selected from protein-DNA benchmark set [13] and the results showed DOT was able to produce correct answers efficiently. However, these results were not ranking forward enough to guide the design of experiments to test the suggested interactions. Hence, amino acid-nucleotide base pair potentials were considered to screen the docked complexes to raise the ranks of the DOT results.

B. Generation of Amino Acid-Nucleotide Base Pair Potential Matrix

Anna Marabotti etc. [14] explored the specificity of the interaction between amino acids and nucleotide bases by determining, in a dataset consisting of 100 high-resolution protein-DNA structures, the frequency and energy of interaction between each amino acid and base, and the energetic of water-mediated interactions. The normalized HINT score fraction between 4 kinds of nucleotide base and 20 kinds of amino acid (AA), which is a non-Newtonian force field encoding both enthalpic and entropic contributions, is adopted as the pair potential matrix as Table I to score interactions across an interface of protein-DNA docked complexes in this study.

C. The Screening Algorithm

DOT was set to produce top 1000 solutions after protein and DNA is docked. According to the matrix, a score was calculated for each complex by summing the appropriate scores of pairs that spanned the interface of the docked complex produced by DOT. The complex structure would be only included the heavy atoms as well as polar hydrogen atoms according to how the potentials matrix were generated. The pairs were considered to exist if the distance between the atoms in the base and residue is within the cutoff distance 6 Å. That is,

Manuscript received January 14, 2013; revised March 14, 2013. This work was supported by the National Natural Science Foundation of China (31200990, 11047183), Foundation for Distinguished Young Talents in Higher Education of Guangdong (LYM10042), Undergraduates Innovating Experimentation Project of Guangdong Province (1056411055).

The authors are with College of Informatics, South China Agricultural University, 510642, Guangzhou, China (email address: ldmfs@126.com, schang@scau.edu.cn, c-j1987@163.com, tianxuhong@scau.edu.cn).

DOI: 10.7763/IJBBB.2013.V3.247

$$\begin{cases} pDNA_k = baseID\_ATOM_i & dist_{i,j} < D \& (baseID\_ATOM_i < pDNA_{k-1} \| \\ pPROT_k = resID\_ATOM_j & resID\_ATOM_j < pPROT_{k-1}) \end{cases} \quad (1)$$

$$score_n = \sum_{k=1}^{k=MaxPairs} matrix[pDNA_k][pPROT_k] \quad (2)$$

where $pDNA_k$ and $pPROT_k$ record the ID of the corresponding base or residue respectively where their atoms are pairing in the pair k, $baseID\_ATOM_i$ is the base ID of atom i in the DNA and $resID\_ATOM_j$ is the residue ID of atom j in the protein, *D* is the distance cutoff in (1). MaxPairs is the number of the pairs in the docked complex in (2).

TABLE I: The Matrix of Amino Acid-Nucleotide Base Pair Potential

Base AA	A	C	G	T
Als	0.00	0.17	-0.43	-0.20
Arg	3.25	1.12	39.41	10.12
Asn	6.49	3.36	1.78	-3.27
Asp	1.05	7.27	-0.74	-1.68
Cys	-0.02	-0.05	0.14	0.01
Gln	3.87	0.87	0.77	-2.36
Glu	1.33	9.06	-1.76	-2.05
Gly	0.34	-0.39	0.66	-0.22
His	0.64	0.66	1.22	0.31
Ile	0.23	-0.16	-0.21	0.20
Leu	0.29	-0.19	-0.05	-0.80
Lys	0..85	0.46	14.24	1.11
Met	-0.13	-0.28	-0.18	-0.73
Phe	0.33	0.07	0.54	0.87
Pro	0.10	0.05	0.03	-0.15
Ser	1.04	0.08	2.47	-0.61
Thr	0.09	-0.45	0.32	-0.47
Trp	0.03	0.30	0.01	0.07
Tyr	0.40	0.35	0.39	0.16
Val	-0.01	-0.67	-0.31	-0.62

In order to speed up the algorithm, once the pairing of two atoms was successful, the rest of the atoms in the same base or residue would not to be judged. The procedure of the algorithm was as follows:

Step 1: Calculate the distance between atoms in the base and residue if the corresponding base or residue is not marked

Step 2: Judge as pairs if the distance is within cutoff range

Step 3: Mark the corresponding base and residue of the pairing atoms

Step 4: Add up the score of the pair to the sum of the complex

Step 5: Turn to Step 1 until all bases and residues in the DNA and protein are checked

Step 6: Rank the complexes after all of them are scored

D. Systems Studied

To test interactions spanning the range from those dominated by shape and hydrophobicity to those governed by electrostatics, we selected protein-DNA systems that differed considerably in size, charge and amount of surface area buried upon complex (Table II). For example, 1A74 interface buries over 2000Å^2, and involves protein with net charges of only +3, yet DNA with -40. On the other hand, 3CRO interface buries only 216Å^2, yet involves protein net charges of +8 and DNA with -37.

TABLE II: Properties of Systems Studied

System	Net charge(e)		Mean change in ASA(Å^2)[a]	Level[b]
	Protein	DNA		
1A74	+3	-40	2175	I
1AZP	+6	-14	825	I
1F4K	+6	-40	626	I
1HJC	+6	-30	702	E
1JJ4	+14	-30	1495	I
1K79	+6	-28	1032	I
2FL3	+1	-18	2151	D
2OAA	-2	-19	2044	D
3BAM	-2	-21	1730	D
4KTQ	-7	-23	2746	I

[a]The mean change in solvent-accessible surface(ASA) that occurs upon complex in the crystallographically determined solution.

[b]The Level column partitions the complexes into the E for easy, I for intermediate and D for difficult notation assigned by van Dijk and Bonvin, which estimates the degree of conformational change upon docking.

III. Results and Discussion

Compared with the results of DOT, Table III shows the results of each system after screening by the pair potentials in this study.

TABLE III: The Screened Result of DOT

System	No.[a]	DOT result		Re-rank result	
		No.[b]	Best rank /RMSD[c]	No.[b]	Best rank /RMSD[c]
1A74	65	14	81/2.44	45	4/2.30
1AZP	32	5	47/3.07	18	10/2.21
1F4K	9	3	4/4.75	5	57/4.75
1HJC	7	0	-	7	4/4.1
1JJ4	5	0	-	5	22/4.04
1K79	4	1	40/5.0	2	5/5.0
2FL3	28	3	14/4.92	9	14/4.72
2OAA	116	13	27/2.45	14	19/2.09
3BAM	2	0	-	2	54/5.0
4KTQ	7	1	61/4.78	2	2/4.84

[a]Number of the top 1000 DOT solutions before screening within the RMSD cutoff 5Å of the crystallographic position.

[b]Number of the top 100 solutions within the RMSD cutoff 5Å of the crystallographic position.

[c]Highest ranked solution within the RMSD cutoff 5Å and the RMSD of this solution from the crystallographic position.

A. The Applicability of Pairs Potentials

Since the systems we examined vary considerately in total charge and size of the interface (Table II), we investigated the average contributions of the electrostatic and van der Waals energies to the composite energy for each complex (Fig. 1) and they could be classified into the following categories: the first category is dominated by the electrostatic term as 1JJ4 (Fig. 1 (a)); the second category is dominated by the attractive van der Waals term as 1A74 (Fig. 1 (b)); the third category has roughly equal contributions from both terms as 1HJC (Fig. 1 (c)). However, the screening pair potential is proved to dock these various kinds of systems successfully. Fig. 2 shows their corresponding structure of the top solution and the native structure. It can be seen that the light grey for the best solutions and the grey for the native are nearly lap over each other.

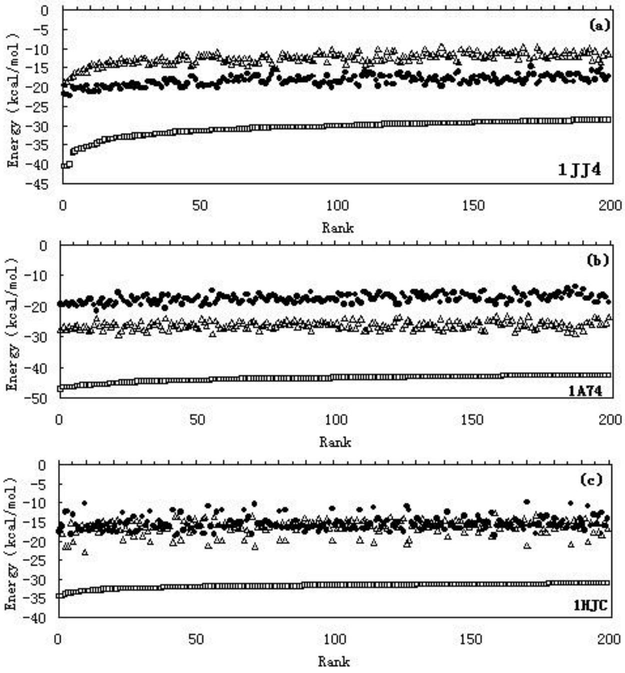

Fig. 1. Contributions of the electrostatic (●) and van der Waals attractive energies (▲) to the composite energy (□) for the top 200 solutions from the composite-energy list. (a) 1JJ4. (b) 2OAA. (c) 1HJC.

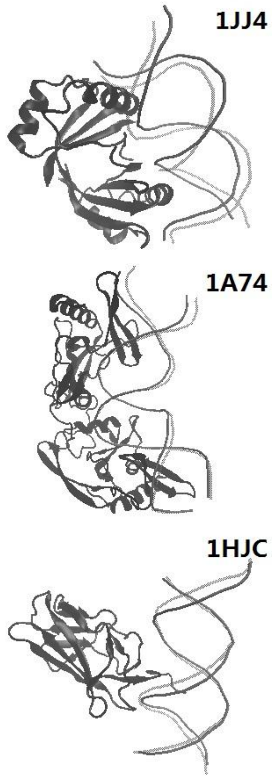

Fig. 2. The structure of 1JJ4, 1A74 and 1HJC (black for the protein, light grey for the best solution and grey for the native).

B. The Re-Rank of the Results

Table IV shows the rank contrast of the best solution between the DOT results and the screening results in this study. It can be seen that the rank of all the systems have improved to a varying degree after the screening of the pair potentials.

However, systems such as 1F4K didn't have an obvious rank promotion as others. Analyzed from the potential matrix, a preferential interaction of Arg and Lys with G, Asp and Glu with C, and Asn and Gln with A was found. Not only favorable, but also unfavorable contacts were Asn, Gln with T, Glu with G and T.

As Fig. 3, the score of the favorable pairs takes over 78% of the total score of 1K79. On the other hand, the score of the favorable pairs takes only 55% of the total score and there are also unfavorable pairs existed in the 1F4K. So it is not difficult to tell why screening results of some systems are better than the others. In the systems where the interface of the complex includes more favorable amino acid-nucleotide base pairs, it is likely to have better screening results.

TABLE IV: THE RANK CONTRAST OF THE BEST SOLUTION

System	Best RMSD	Orginal-rank	Re-rank
1A74	1.92	831	51
1AZP	1.34	500	131
1F4K	2.98	319	174
1HJC	2.99	620	27
1JJ4	3.56	862	218
1K79	4.58	627	65
2FL3	2.20	875	191
2OAA	1.81	805	24
3BAM	4.67	921	96
4KTQ	2.53	812	120

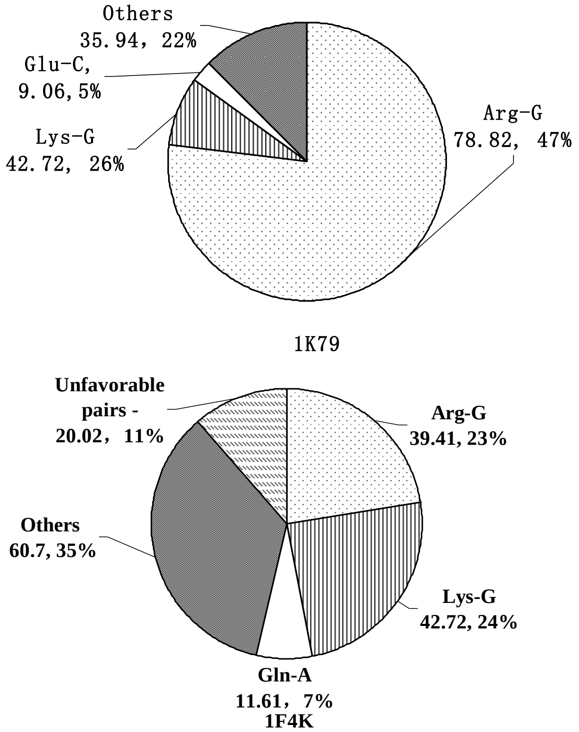

Fig. 3. The percentage of different pairs' score in the best screened result.

C. The Number of Correct Results

With the top 1000 solutions of DOT, the screening program obviously picked up more correct solutions in the forward ranks than DOT (Fig. 4). In the systems like 1AZP and 2FL3, over 96% of the correct answers were screened out in top 300 solutions. A method capable of generating larger number of correct solutions has important advantages. Correct solutions could be identified with stringent filtering using biochemical information, but some correct solutions were typically lost at each filtering step. In such cases, filters applied to a large number of close-to-correct solutions are likely to be more successful than application to a few close-to-correct solutions.

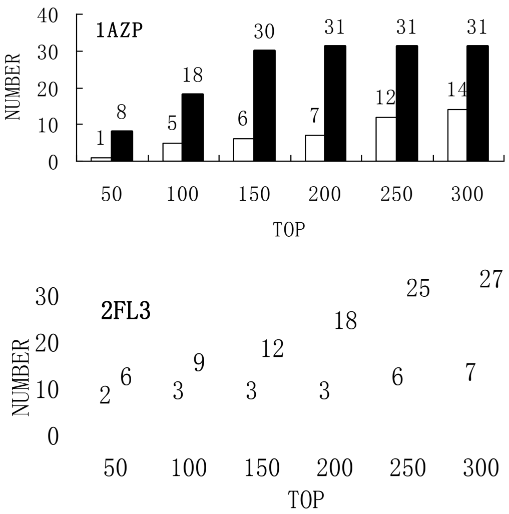

Fig. 4. The contrast of the number of correct result in top 50-300 between before and after screening (blank blocks for the DOT results and black blocks for the screened results).

IV. Conclusion

The key conclusion of this study is that amino-acid nucleotide base pair potentials have considerable power in correctly selecting correct dockings from a list of complexes. The experiment results of our study shows: Firstly, the potentials are adapted to a variety of protein-DNA systems. Secondly, the rank of the correct docked solutions has largely improved after the screening. Thirdly, the screening pair potentials are able to pick up almost all of the correct solutions within forward ranks.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (31200990, 11047183), Foundation for Distinguished Young Talents in Higher Education of Guangdong (LYM10042), Undergraduates Innovating Experimentation Project of Guangdong Province (1056411055).

References

[1] P. Setny, R. P. Bahadur, and M. Zacharias, "Protein-DNA docking with a coarse-grained force field," *BMC Bioinformatics*, vol. 13, no. 1, pp. 228, 2012.

[2] J. H. Gan, J. Sheng, and Z. Huang, "Chemical and structural biology of nucleic acids and protein-nucleic acid complexes for novel drug discovery," *Science China-Chemistry*, vol. 54, no. 1, pp. 3-23, 2011.

[3] S. Velankar *et al.*, "PDBe: Protein Data Bank in Europe," *Nucleic Acids Res*, vol. 39, Suppl. 1, pp. D402-D410, 2011.

[4] D. W. Ritchie, "Recent progress and future directions in protein-protein docking," *Curr Protein Pept Sci*, vol. 9, no. 1, pp. 1-15, 2008-02-01 2008.

[5] M. Parisien, K. F. Freed, and T. R. Sosnick, "On docking, scoring and assessing protein-DNA complexes in a rigid-body framework," *PloS one*, vol. 7, no. 2, pp. e32647, 2012.

[6] M. van Dijk, A. van Dijk, V. Hsu, R. Boelens, and A. Bonvin, "Information-driven protein-DNA docking using HADDOCK: it is a matter of flexibility," *Nucleic Acids Res*, vol. 34, no. 11, pp. 3317-3325, 2006.

[7] T. Takeda, R. I. Corona, and J. Guo, "A knowledge-based orientation potential for transcription factor-DNA docking," *Bioinformatics*, pp. 2325-2335, 2012.

[8] L. A. Liu and P. Bradley, "Atomistic modeling of protein – DNA interaction specificity: progress and applications," *Curr Opin Struc Biol*, pp. 397-405, 2012.

[9] G. Moont, H. A. Gabb, and M. J. E. Sternberg, "Use of pair potentials across protein interfaces in screening predicted docked complexes," *Proteins: Structure, Function, and Bioinformatics*, vol. 35, no. 3, pp. 364-373, 1999.

[10] O. Rahaman, T. P. Estrada, D. J. Doren, M. Taufer, C. L. Brooks III, and R. S. Armen, "Evaluation of several two-step scoring functions based on linear interaction energy, effective ligand size, and empirical pair potentials for prediction of protein – ligand binding geometry and free energy," *Journal of chemical information and modeling*, vol. 51, no. 9, pp. 2047-2065, 2011.

[11] D. Kozakov, R. Brenke, S. R. Comeau, and S. Vajda, "PIPER: An FFT - based protein docking program with pairwise potentials," *Proteins: Structure, Function, and Bioinformatics*, vol. 65, no. 2, pp. 392-406, 2006.

[12] J. G. Mandell, V. A. Roberts, M. E. Pique, V. Kotlovyi, J. C. Mitchell, E. Nelson, I. Tsigelny, and E. L. Ten, "Protein docking using continuum electrostatics and geometric fit," *Protein Eng*, vol. 14, no. 2, pp. 105-13, 2001.

[13] M. van Dijk and A. M. Bonvin, "A protein-DNA docking benchmark," *Nucleic Acids Res*, vol. 36, no. 14, pp. e88, 2008.

[14] A. Marabotti, F. Spyrakis, A. Facchiano, P. Cozzini, S. Alberti, G. E. Kellogg, and A. Mozzarelli, "Energy-based prediction of amino acid-nucleotide base recognition," *J Comput Chem*, vol. 29, no. 12, pp. 1955-69, 2008.

Dongmin Liu received her B.Sc. degree in Computer Science and Technology in 2010 from South China Agricultural University, Guangzhou, China. She is now a M.S. candidate in the college of Informatics, South China Agricultural University, Guangzhou, China. Her research interest includes protein-NA structure prediction, bioinformatics.

Shan Chang received his Ph.D. degree in Biomedical Engineering in 2009 from Beijing University of Technology, Beijing, China. He is now an associate professor in the Information College of South China Agricultural University, Guangzhou, China. His research interest includes protein molecular docking, scoring function, searching algorithm, amino acid network, evolving network model.

Jian Chen received his B.Sc. degree in Computer Science and Technology in 2010 from South China Agricultural University, Guangzhou, China. He is now a M.S. candidate in the College of Informatics, South China Agricultural University, Guangzhou, China. His research interest includes computer network, Agricultural Engineering.

Xuhong Tian received his Ph.D. degree in 2008 from South China University of Technology, Guangzhou, China. He is now a professor in the Information College of South China Agricultural University, Guangzhou, China. His research interest includes bioinformatics, graph and image processing, parallel processing.