A Method For Discriminating Native Protein-DNA Complexes From Decoys Using Spatial Specific Scoring Matrices

Wen Cheng
Department of Computer Science
North Dakota State University
Fargo, ND, USA
wen.cheng@my.ndsu.edu

Changhui Yan
Department of Computer Science
North Dakota State University
Fargo, ND, USA
changhui.yan@ndsu.edu

Abstract— **Decoding protein-DNA interactions is important to understanding gene regulation and has been investigated by worldwide scientists for a long time. However, many aspects of the interactions still need to be uncovered. The crystal structures of protein-DNA complexes reveal detailed atomic interactions between the proteins and DNA and are an excellent resource for investigating the interactions. In this study, we profiled the spatial distribution of protein atoms around six structural components of the DNA, which are the four bases, the deoxyribose sugar and the phosphate group. The resultant profiles not only revealed the preferred atomic interactions across the protein-DNA interface but also captured the spatial orientation of the interactions. The profiles are a useful tool for investigating protein-DNA interactions. We tested the strength of profiles in two experiments, discrimination of native protein-DNA complexes from decoys with mutant DNA and discrimination of native protein-DNA complexes from near-native docking decoys. The profiles achieved an average Z-score of 7.41 and 3.22 respective on benchmark datasets for the tests, both are better than other knowledge-based energy functions that model protein-DNA interaction based on atom pairs.**

Keywords— protein-DNA interaction; spatial specific scoring matrix; fragment-based method

I. INTRODUCTION

Knowledge about protein-DNA interactions is crucial for understanding important cell processes like gene regulation. Enormous efforts have been made to investigate various problems pertaining to protein-DNA interactions. Some interesting problems are finding DNA-binding proteins that will bind to DNA, detecting DNA-binding sites on proteins, and determining the binding mode between a given pair of protein and DNA.

Many computational methods have been developed for predicting DNA-binding sites on protein structures. Some methods focus on the geometrical and physiochemical properties of DNA-binding sites and use data-mining or statistical approach to predict potential DNA-binding sites on new protein structures [1-7]. These methods usually represent patches on the protein surface using vectors or graphs, and then compare them with known DNA-binding sites. These methods usually suffer relatively low accuracy, and some of them are very computational demanding. Other methods rely on structural alignment [8, 9]. These methods maintain a database of protein structures whose DNA-binding sites are known. To predict DNA-binding sites on a new protein (A.K.A. query), the new protein is used to query the database to find structures (A.K.A. templates) that share a high similarity with it. The query protein structure is then aligned with the templates. The region on the query protein that superimpose with the known DNA-binding sites on the templates will be predicted to be a DNA-binding site. The success of these methods strongly depends on the availability of templates and the level of similarity between the query and templates.

Other researchers using docking approach to predict the structure of the protein-DNA complex. The resultant complex structure not only reveals the DNA-binding sites on the protein structure but also shows the detailed atomic interaction between the protein and the DNA. A docking method searches the conformation space of the complex and uses an energy function to score the conformations. Different docking methods differ in the energy function used. Some use functions that model various physical and chemical forces between atoms [10-14]. Others use knowledge-based statistical energy functions derived from observed interacting atom pairs across the interface [15, 16].

Herein, we present a knowledge-based energy function for the study of protein-DNA interaction. We analyzed the distribution of protein atoms around each structural component of the DNA and developed spatial specific scoring matrices (SSSMs) based on the observed distribution. We showed that the SSSMs could be used as a knowledge-based energy function to discriminate native protein-DNA structures and various decoys.

II. METHODS AND MATERIALS

A. Datasets

The testing dataset for the first test, DNA mutation decoy test, was composed of 51 non-redundant complexes from [17]. For the second test, the near-native docking decoys were

978-1-4799-1389-3/13/$31.00 ©2013 IEEE

generated using FTDock [18] from 45 protein-DNA complexes collected by Robertson and Varani [16]. The training datasets for both tests were derived from the 212 protein-DNA complexes used in Xu et al. [13]，which were extracted from the PDB database and culled by the PISCES server [19] such that pairwise similarity is less than 35%. In both tests, we removed from the training set complexes that have higher than 35% similarity with any protein in the test sets. As a result, the training set for the DNA mutation decoy test contained 166 protein-DNA complexes and the training set for the near native docking decoy discrimination test contained 167 protein-DNA complexes.

B. Spatial specific scoring matrices (SSSMs)

We first divided DNA into six repeating structural components, which were the four bases, the deoxyribose sugar and the phosphate group. We collected the protein atoms that contacted with these components and investigated how they distributed around the components in the space. For each component, we defined a new coordinate system that centered at it. Using the new coordinate system as grid, we divided the space into 16*16*16 cubes, with 16 bins on each axis. The size of the cubes was customized so that all the protein atoms contacting with the component fell into one of the cubes. We classified the atoms of protein into 14 types as described in [20] and then counted the number of different types of atoms falling in each cube. Therefore, the distribution of protein atoms around a component was described using a 16*16*16*14 matrix. The counts in the matrix were normalized by the total count. Therefore, each cell in the matrix corresponded to one atom type and one cube in the space, and the value in the cell showed how likely the atom would contact the DNA component from a location corresponding to the cube. These matrices were populated using protein-DNA complexes in the training set. The resultant six matrices (will be referred as SSSMs) were used as scoring matrices to discriminate native protein-DNA complexes from various decoys. For a given structure (native or decoy) of protein-DNA complex, a score was assigned to it using the following method.

$$S = \sum_{i=1}^{6} \sum_{j=1}^{16*16*16} \sum_{k=1}^{14} O_{ijk} P_{ijk} \quad (1)$$

where O_{ijk} is the number of atoms of type k that contact component i from the location corresponding to cube j, and P_{ijk} is the value in the cell of the scoring matrix for component i that corresponds to atom type k and cube j. Higher scores mean that the complex was more likely to be the native structure.

III. EXPERIMENTS AND RESULTS

A. Test 1: To discriminate native structures from DNA mutation decoys

For this test, 166 protein-DNA complexes were used as training set to derive the six scoring matrices and a disjoint test set consisting 51 protein-DNA complexes were used to generate decoys. For each of the native complex, we generated 50,000 decoys by replacing a nucleotide base with a different type of base with equal opportunity. The new base was placed in the same plane as the native one. Then we calculated the scores for the native complex and the decoys. Since the native complex only differed from the decoys in the bases, only the four SSSMs corresponding to bases were used in this test to calculate the scores. We used Z-score to evaluate the performance of discriminating native complex and decoys. Here Z-score=(S_{avg}-S_{native})/SD, where S_{avg} and SD are the average and standard deviation of scores of 50,000 decoys, and S_{native} is the score of the native structure. Since the native structure is expected to have higher score than the decoys, a lower negative Z-score means that the scoring system is able to distinguish the native structure from decoys with better performance. Our method achieved an average Z-score -7.41 on the test set. The Z-scores for each complex were shown in Table 1.

Many researchers have tried to develop knowledge-based energy functions for protein-DNA interactions based on observed atomic contacts across the interface. Zhou and Zhou [12] first applied a distance-scaled, finite ideal-gas (DFIRE) energy function for protein-DNA interaction. Gromiha et al. [21] also developed energy functions based on intermolecular and intramolecular contacts. Xu et al. [13] developed five variants of DFIRE energy functions, among which the variant (named vcFIRE) with low-count correction and volume correction achieved the best result. Xu et al. [13] evaluated and compared these methods using the same training and test datasets that were used in this study. Herein, we used the results from their study and compared our method with others. Table 1 showed that our method achieved better z-scores than all other methods in all but two complexes. The only exceptions are 1cjg, and 1xbr (shaded gray in Table 1). In 1xbr, our z-score was very close to the best. Paired t-test showed that our method outperformed all others with $p<0.0001$. The average Z-score for our method on the dataset is -7.41, which was much better than that of any other methods.

TABLE I. Z-SCORES FOR DIFFERENT METHODS IN THE TEST OF DISCRIMINATING NATIVE STRUCTURES FROM MUTATION DECOYS.

PDB ID	Gromiha et al. (2004)	Zhou and Zhou (2002)	Xu et al. (2009)	Our Method
1a02	-1.8	-2.27	-3.29	-18.27
1a74	0.7	1.50	-4.17	-5.50
1b3t	-2.1	-1.15	-2.38	-2.44
1bhm	-1.3	-0.05	-3.26	-6.20
1bl0	-2.5	-2.23	-3.25	-8.56
1cdw	-0.6	1.64	-0.02	-5.45
1cjg	-1.4	-2.58	-0.81	-0.10
1cma	-1.6	1.02	-1.59	-2.69
1e66	-1.7	-3.22	-3.12	-4.01
1dp7	-0.7	0.76	-3	-3.02
1ecr	-1.1	0.53	-1.58	-5.01
1fjl	-1	2.59	-2.63	-11.53

978-1-4799-1389-3/13/$31.00 ©2013 IEEE

1gat	-1.7	1.73	-1.27	-2.12
1gdt	-1.7	-0.04	-3.75	-10.70
1glu	-1.1	1.72	-1.95	-12.03
1hcq	-2.5	-0.85	-4.09	-10.11
1hcr	0.4	-0.25	-2.43	-3.70
1hdd	-1.8	0.95	-1.57	-6.48
1hlo	-1.6	0.29	-3.95	-5.83
1hry	-0.9	0.23	-1.33	-3.76
1if1	-1.7	-1.62	-1.96	-8.64
1ign	-2.2	-0.23	-5.32	-8.32
1ihf	-2.3	1.79	-1.81	-2.35
1j59	-0.8	-2.33	-3.79	-12.29
1lmb	-4.3	-1.48	-4.25	-7.04
1mdy	-2.5	2.81	-2.83	-14.06
1mey	-2.2	-1.52	-4.92	-9.84
1mhd	-1.9	0.56	-2.74	-7.12
1mnm	-3	0.20	-4.04	-8.24
1mse	-2	-0.69	-2.13	-4.46
1oct	-2.1	-0.37	-2.85	-8.96
1par	-1.7	-0.96	-2.42	-5.34
1pdn	-2.5	-1.06	-1.92	-8.41
1per	-1.1	0.20	-1.92	-8.53
1pue	-2.7	-1.27	-2.21	-11.13
1rep	-3.2	-2.2	-3.01	-12.57
1rv5	-0.3	0.11	-1.67	-3.99
1srs	-2.4	0.67	-3.62	-8.44
1svc	-2.2	-1.68	-4.27	-9.04
1tc3	-2.5	-0.24	-2.29	-6.46
1tf3	-2.3	-1.19	-3.56	-5.45
1tro	-3.1	-0.19	-4.05	-7.41
1tsr	-1.2	-2.38	-2.68	-8.74
1ubd	-2.1	-0.12	-4	-7.26
1xbr	-2.4	-2.76	-2.4	-2.21
1yrn	-2.9	-0.05	-3.78	-9.10
1ysa	-2.1	0.14	-4.01	-8.88
2bop	-1.7	-2.16	-3.12	-4.04
2drp	-2.3	1.40	-4.75	-21.02
3cro	0.3	-1.52	-0.57	-9.61
6cro	-2.3	-3.86	-3.79	-5.38
Mean	-1.8	-0.43	-2.86	-7.41

B. Test 2: To discriminate native structure from near-native docking decoys

This experiment was designed to test the ability of SSSMs to discriminate native complexes from near-native docking decoys. We created 10,000 docking decoys for each of the 45 native complexes using FTDock. 2,000 lowest-RMSD decoys were selected. These were near-native decoys. Six SSSMs were derived using 167 the protein-DNA complexes from the training set. Then, these SSSMs were used to compute scores for the native complex and near-native decoys.

For this test, we compared our method with the DFIRE-based methods developed by Zhou and Zhou [12] and Xu et al. [13] and an all-atom distance-based method developed by Robertson and Varani [16]. These methods were all evaluated using the same training and test datasets as used in this study. Our method achieved an average Z-score of -3.22, which was the best among all methods (Table 2). Our method achieved the best Z-score for 29 of the 45 protein-DNA complexes. Paired t-test confirmed that our method outperformed the others with $p<0.0001$.

TABLE II. COMPARISONS OF METHODS IN TERMS OF Z-SCORES FOR THE TEST OF DISCRIMINATING NATIVE STRUCTURES FROM NEAR-NATIVE DOCKING DECOYS.

PDBid	Zhou and Zhou (2002)	Xu et al. (2009)	Robertson and Varani, (2007)	Our method
1qna	-1.21	-1.79	-1.57	-2.36
1d02	-1.47	-2.63	-1.95	-4.91
1eon	-1.66	-3.09	-1.98	-3.52
1ckq	-1.02	-1.94	-1.14	-2.77
1dmu	-1.55	-4.16	-2.06	-3.06
1qpz	-2.2	-3.48	-2.55	-3.04
1au7	-1.52	-2.55	-1.96	-3.86
1je8	-1.85	-2.91	-2.04	-2.43
2cgp	-0.97	-1.99	-1.42	-2.07
1b3t	-1.38	-2.99	-1.94	-2.27
1tc3	-1.56	-2.67	-1.56	-3.02
1g9z	-2.63	-5.45	-3.29	-3.89
1zme	-2.13	-2.38	-2.26	-4.01
1a73	-1.85	-3.41	-2.3	-5.90
1jko	-1.77	-3.12	-2.16	-3.21
1bdt	-1.77	-3.19	-1.88	-3.13
2bop	-1.68	-2.97	-2.13	-2.55
1a1i	-1.44	-2.49	-1.98	-5.09
1bc8	-1.5	-2.67	-2.1	-3.22
1pdn	-1.45	-2.47	-2.17	-3.13
1skn	-1.23	-2.6	-2.06	-4.98
1mjo	-2.09	-2.55	-2.16	-3.12
1bl0	-0.96	-1.92	-1.4	-1.70
2dgc	-1.46	-2.36	-2.06	-1.30
3pvi	-1.65	-2.34	-1.86	-2.19

978-1-4799-1389-3/13/$31.00 ©2013 IEEE

2hdd	-2.37	-3.13	-2.7	-4.82
1ign	-1.74	-3.44	-2.3	-3.06
1qpi	-2.12	-3.67	-3.07	-3.26
1a3q	-1.46	-2.49	-1.91	-3.02
1dfm	-1.23	-2.6	-1.51	-1.97
1lq1	-1.94	-3.26	-2.38	-2.73
1tro	-1.43	-2.78	-2.05	-2.86
1fjl	-1.36	-2.12	-1.58	-3.45
1h8a_a	-1.29	-2.35	-2	-1.52
1h8a_b	-1.02	-2.18	-1.59	-4.71
1f4k	-1.16	-2.58	-2.1	-2.74
6pax	-1.21	-2.74	-1.96	-1.28
1hlv	-1.77	-3.17	-2.23	-2.48
1mnn	-1.59	-3.4	-2.49	-5.68
1dsz	-1.12	-2.38	-1.82	-2.79
1hwt	-1.77	-1.96	-2.4	-2.65
1per	-1.44	-2.7	-2.08	-3.62
1l3l	-1.76	-3.1	-2.42	-4.54
3hts	-0.95	-3.03	-2.05	-3.32
3bam	-1.66	-2.86	-1.99	-3.70
Mean	**-1.56**	**-2.8**	**-2.06**	**-3.22**

IV. CONCLUSIONS

We have developed a knowledge-based scoring function for assessing protein-DNA interactions. We divided the DNA into 6 repeating structural components and used spatial specific scoring matrices (SSSMs) to capture the distribution of protein atoms around these components in the 3D space. The proposed method was able to discriminate native protein-DNA complexes from various decoys with better performance than other knowledge-based energy functions. Compared with other energy functions derived from observed atom contacts, the proposed SSSMs not only reflect the preferences for atomic interactions across the protein-DNA interface but also capture the spatial orientation of the interactions. The SSSMs will be a useful tool for investigating protein-DNA interactions.

ACKNOWLEDGMENT

The project described was partially supported by NIH Grant Number P20 RR016471 from the INBRE Program of the National Institute of General Medical Sciences.

REFERENCES

[1] R. Chikhi, L. Sael, and D. Kihara, "Real-time ligand binding pocket database search using local surface descriptors," Proteins, 2010, vol. 78(9), pp. 2007-2028.

[2] F. Guo, S.C. Li, and L. Wang, "Protein-Protein Binding Sites Prediction by 3D Structural Similarities," J Chem Inf Model, 2011, vol. 51(12), pp. 3287-3294.

[3] J. Ito, Y. Tabei, K. Shimizu, K. Tomii, and K. Tsuda, "PDB-scale analysis of known and putative ligand-binding sites with structural sketches," Proteins, 2012, vol. 80(3), pp. 747-763.

[4] L. Sael, and D. Kihara, "Detecting local ligand-binding site similarity in nonhomologous proteins by surface patch comparison". Proteins, 2012, vol. 80(4), pp. 1177-1195.

[5] V. Bianchi, P.F. Gherardini, M. Helmer-Citterich, and G. Ausiello, "Identification of binding pockets in protein structures using a knowledge-based potential derived from local structural similarities," BMC Bioinformatics, 2012, vol. 13 (S4), pp. S17.

[6] J. Konc, and D. Janezic, "ProBiS algorithm for detection of structurally similar protein binding sites by local structural alignment," Bioinformatics, 2010, 26(9):1160-1168.

[7] W. Zhou, and H. Yan, "A discriminatory function for prediction of protein-DNA interactions based on alpha shape modeling," Bioinformatics, 2010, vol. 26(20), pp. 2541-2548.

[8] M.N. Wass, L.A. Kelley, and M.J. Sternberg, "3DLigandSite: predicting ligand-binding sites using similar structures," Nucleic Acids Res, 2010, vol. 38(Web Server issue), pp. W469-473.

[9] K. Kinoshita, and H. Nakamura, "Identification of the ligand binding sites on the molecular surface of proteins," Protein Sci, 2005, vol. 14(3), pp. 711-718.

[10] Z. Liu, F. Mao, J.T. Guo, B. Yan, P. Wang, Y. Qu, and Y. Xu, "Quantitative evaluation of protein-DNA interactions using an optimized knowledge-based potential," Nucleic Acids Res, 2005, vol. 33(2), pp. 546-558.

[11] C. Zhang, S. Liu, Q. Zhu, and Y. Zhou, "A knowledge-based energy function for protein-ligand, protein-protein, and protein-DNA complexes," J Med Chem, 2005, vol. 48(7) pp. 2325-2335.

[12] H. Zhou, and Y. Zhou, "Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction," Protein Sci, 2002, vol. 11(11), pp. 2714-2726.

[13] B. Xu, Y. Yang, H. Liang, and Y. Zhou, "An all-atom knowledge-based energy function for protein-DNA threading, docking decoy discrimination, and prediction of transcription-factor binding profiles," Proteins, 2009, vol. 76(3), pp. 718-730.

[14] H. Zhao, Y. Yang, and Y. Zhou, "Structure-based prediction of DNA-binding proteins by structural alignment and a volume-fraction corrected DFIRE-based energy function," Bioinformatics, 2010, vol. 26(15), pp. 1857-1863.

[15] R. Samudrala, and J. Moult, "An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction," J Mol Biol, 1998, vol. 275(5), pp. 895-916.

[16] T.A. Robertson, and G. Varani, "An all-atom, distance-dependent scoring function for the prediction of protein-DNA interactions from structure," Proteins, 2007, vol. 66(2), pp. 359-374.

[17] H. Kono, and A. Sarai, "Structure-based prediction of DNA target sites by regulatory proteins," Proteins, 1999, vol. 35(1), pp. 114-131.

[18] E. Katchalski-Katzir, I. Shariv,M. Eisenstein, A.A. Friesem, C. Aflalo, and I.A. Vakser, "Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques," In Proceedings Of The National Academy Of Sciences Of The United States Of America, 1992, vol. 89(6), pp. 2195-2199.

[19] G. Wang, and R.L.J. Dunbrack, "PISCES: a protein sequence culling server," Bioinformatics, 2003, vol. 19, pp. 1589-1591.

[20] E. Petsalaki, A. Stark, E. Garcia-Urdiales, and R.B. Russell, "Accurate prediction of peptide binding sites on protein surfaces," PLoS Comput Biol, 2009, vol. 5(3), pp. e1000335.

[21] S. Ahmad, M.M. Gromiha, and A. Sarai, "Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information," Bioinformatics, 2004, vol. 20(4), pp. 477-486.

978-1-4799-1389-3/13/$31.00 ©2013 IEEE