BIOINFORMATICS ORIGINAL PAPER Vol. 26 no. 20 2010, pages 2541–2548
doi:10.1093/bioinformatics/btq478

Structural bioinformatics Advance Access publication August 23, 2010

A discriminatory function for prediction of protein–DNA interactions based on alpha shape modeling

Weiqiang Zhou* and Hong Yan

Department of Electronic Engineering, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Protein–DNA interaction has significant importance in many biological processes. However, the underlying principle of the molecular recognition process is still largely unknown. As more high-resolution 3D structures of protein–DNA complex are becoming available, the surface characteristics of the complex become an important research topic.

Result: In our work, we apply an alpha shape model to represent the surface structure of the protein–DNA complex and developed an interface-atom curvature-dependent conditional probability discriminatory function for the prediction of protein–DNA interaction. The interface-atom curvature-dependent formalism captures atomic interaction details better than the atomic distance-based method. The proposed method provides good performance in discriminating the native structures from the docking decoy sets, and outperforms the distance-dependent formalism in terms of the *z*-score. Computer experiment results show that the curvature-dependent formalism with the optimal parameters can achieve a native *z*-score of −8.17 in discriminating the native structure from the highest surface-complementarity scored decoy set and a native *z*-score of −7.38 in discriminating the native structure from the lowest RMSD decoy set. The interface-atom curvature-dependent formalism can also be used to predict apo version of DNA-binding proteins. These results suggest that the interface-atom curvature-dependent formalism has a good prediction capability for protein–DNA interactions.

Availability: The code and data sets are available for download on http://www.hy8.com/bioinformatics.htm

Contact: kenandzhou@hotmail.com

Received on June 6, 2010; revised on August 5, 2010; accepted on August 14, 2010

1 INTRODUCTION

Protein–DNA interaction plays an important role in many biological processes, such as DNA replication, transcription and nucleosome remodeling (Cartharius *et al*., 2005; Johnson and McKnight, 1989; Kamei *et al*., 1996; Sancar *et al*., 2004; Stormo, 2000). In the beginning, scientists focused on the nucleic acid sequence information (Fickett, 1982; Schneider *et al*., 1986) and tried to explain protein–DNA interaction by genetic codes. However, later research on the geometric analysis of the protein–DNA interface (Pabo and Nekludova, 2000) showed that there is no simple code for protein–DNA recognition. In recent years, they paid more attention to the pairwise interatomic distance information provided by the 3D structure of the protein–DNA complex. Samudrala and Moult proposed an all-atom distance-dependent discriminatory function for the prediction of nucleic acid binding proteins (Samudrala and Moult, 1998). They applied the conditional probability theory to the analysis of the protein structures and got good results compared with the free energy theory. Later, Moont *et al*. (1999) applied an interface pairwise residue level potential to the screening of predicted docked complex. Recently, Robertson and Varani improved the method based on an interface-atom distance-dependent formalism and showed better prediction power than previous methods (Robertson and Varani, 2007). Gao and Skolnick developed a knowledge-based method, which can perform apo version of DNA-binding protein prediction (Gao and Skolnick, 2008). There are several other methods that have been devised to predict protein-related interactions: Liu *et al*. (2008) developed a structure-base method for the transcription factor binding site prediction, Ahmad proposed the usage of moment information in the prediction of DNA-binding proteins (Ahmad and Sarai, 2004). Ahmad *et al*. (2008) applied the clustering method in the analysis of protein–DNA structural data.

However, the underlying principle of protein–DNA interactions is still largely unknown. Previously, some scientists focused on the electrostatics features of the amino acid (Ahmad and Sarai, 2004) and some focused on the interatomic distance (Robertson and Varani, 2007; Samudrala and Moult, 1998). Although these attempts to predict protein–DNA interaction provided acceptable results, few paid enough attention to the 3D interface surface characteristics of the protein–DNA complex which are also direct factors influencing the binding process (Jones *et al*., 1999; Siggers *et al*., 2005). In protein–DNA interaction, the binding surface of the protein should provide certain conditions to adapt to particular DNA molecular surfaces. Such conditions can be the atom type, surface curvature, accessible surface area, net charge, etc. As more high-resolution 3D structures of biological molecules are becoming available, the surface characteristics of the molecules have become an important research topic (Bernauer *et al*., 2007; Nicola and Vakser, 2007; Ofran and Rost, 2003; Sael *et al*., 2008; Zhou and Yan, 2010). A useful tool for object surface analysis is the 3D alpha shape model. Alpha shape has been used for a long time in molecular volume computation, cavities detection and shape representation. Liang *et al*. (1998a, b) first proposed to use alpha shape modeling to compute the molecular area, volume and detect the inaccessible cavities in proteins. Li *et al*. (2003) used the edges in alpha shape modeling to represent the protein structure and atom contacts. Poupon used Voronoi tessellations to compute the protein

*To whom correspondence should be addressed.

volume and detect the pockets, cavities and voids on the protein surface (Poupon, 2004). Recently, alpha shape has been introduced into the study of molecular surface. Albou *et al*. (2009) applied alpha shape modeling to characterize the surface of the protein and defined the surface residue, and surface patches with some local features. However, most of the work has been done on protein surface analysis and little attention has been paid to the interface surface characteristics of the protein–DNA complex.

In our research, we propose to apply 3D alpha shape modeling to study the interface surface characteristics of the protein–DNA complex and develop a surface characteristic-based discriminatory function for the prediction of protein–DNA interaction.

2 METHODS

2.1 Experimental data selection

The correct protein–DNA complex data set contains 199 different types of DNA-binding proteins which expands the data created by Gao and Skolnick (2008). In order to compare the performance of the method proposed in this article and the distance-dependent method, 45 native complexes are selected as per Robertson and Varani (2007). The experimental data are generated using the Fourier Transform rigid-body docking package (FTDock) provided by Aloy *et al*. (1998). The protein and DNA structures are separated, the larger molecule is held fixed and the smaller molecule is allowed to move independently. FTDock is performed on 45 native structures with default scoring and search parameters and 10 000 top-scored decoys from each native structure are obtained to form the entire training set. After that, the 2000 highest surface-complementarity (SC, which can be determined using the FTDock program) (Gabb *et al*., 1997) scored decoy structures are retained for every complex. At the same time, the 2000 lowest RMSD (C_αRMSD to the native complex) decoys are also retained for every complex. Furthermore, another test set that contains 86 protein–RNA, 106 protein–ligand and 103 protein–protein structures is used to evaluate the specificity of our method. The protein–RNA complexes are found from the Bioinfo Bank (http://gibk26.bse.kyutech.ac.jp /jouhou/jouhoubank.html). The protein–protein complexes are the same as those used in (Murakami and Mizuguchi, 2010). A search on the PDB produces many entries of protein–ligand structures. We randomly choose 106 structures to make the number of samples comparable to those of the protein–RNA and protein–protein structures. For the purpose of testing the performance of the proposed method on the apo protein structures, another experimental data set is obtained which contains 104 DNA-binding proteins and 401 non-DNA-binding proteins.

2.2 Alpha shape modeling

We use the 3D alpha shape to represent the surface of the protein–DNA complex and extract features to characterize the interface of the complex from the alpha shape model.

Alpha shape modeling is very useful in reconstructing the surface of an object. It has found many applications in image processing and data visualization. It has also been used to study the molecular structures such as the detection of pockets in known structures, computation of the molecular volume and description of the protein surface (Albou *et al*., 2009; Edelsbrunner *et al*., 1998; Pontius *et al*., 1996; Poupon, 2004).

The 3D alpha shape can be formed based on the Delaunay triangulation (Delaunay, 1934), which is a unique partition of the 3D space in non-overlapping tetrahedrons. The edges of the Delaunay triangulation of a protein–DNA complex are shown in Figure 1A. Obviously, it cannot efficiently represent the surface of the complex. However, this structure contains all the edges we need to form the surface structure of the protein–DNA complex. Consequently, the alpha shape is developed by trimming the edges (Fig. 1B) of the Delaunay triangulation which is a subset of the tetrahedrons in the Delaunay triangulation complex. It is a generalization of

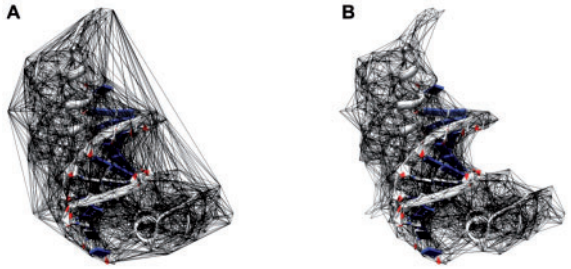

Fig. 1. (**A**) The edges of the Delaunay triangulation of a protein–DNA complex. (**B**) The edges of the alpha shape obtained from the Delaunay triangulation.

Fig. 2. Alpha shape model obtained with different alpha values. (**A**) The molecular surface cannot be obtained completely if the alpha value is too small. (**B**) Surface obtained perfectly with a proper alpha value. (**C**) The surface will lose some details if the alpha value is too large.

the convex hall of the point set (Edelsbrunner and Mucke, 1994) (the atoms in a molecule).

Based on the Delaunay triangulation, the alpha shape can be computed using the following procedure: First, define the alpha complex of the set of points $\{S\}$ which is a sub-complex of the Delaunay triangulation. For a given value of α, the alpha complex includes all the simplexes in the Delaunay triangulation which have an empty circumsphere with a squared radius equal to, or smaller than, α. Here 'empty' means that the open sphere does not include any points of $\{S\}$. The alpha shape is then simply the domain covered by the simplexes of the alpha complex. Notice that, the alpha value here actually controls the preciseness of the molecular surface obtained (Fig. 2). Smaller alpha values give us a more detailed representation of the molecular surface, but the molecular surface will become fragmentary if the alpha value is too small (Fig. 2A). However, if the alpha value is too large, the details of the molecular surface may be lost (Fig. 2C). In this work, we rely on the CGAL (CGAL) library to compute the alpha shape. Different alpha values are used to search for the optimal parameters. Alpha value selection is discussed in the 'Results and Discussion' section.

2.3 Features

In order to extract the features from the alpha shape model of the protein–DNA complex, we have to define the interface atoms. Because the vertices of the alpha shape model correspond to the surface atoms of the original structure, we can define the interface atoms of the protein–DNA structure using the following steps: first, we calculate the alpha shape of the protein–DNA complex (Fig. 3A) and obtain the complex surface atoms set $\{A_i\}$. Then we calculate the alpha shape of the protein (Fig. 3B) independently and obtain the protein surface atoms set $\{B_i\}$. After that, the interface atoms set can be obtained by observing the atoms which are in $\{B_i\}$ but not in $\{A_i\}$. The interface surface is shown in Figure 3B in red.

Three features including atom type, residue type and surface curvature are extracted from the interface atoms of the alpha shape model. All 20 amino acid residue types are taken into consideration in our work. According to the significance of the atom types in the protein–DNA structure, 36 special atom types as shown in Table 1 are considered (Samudrala and Moult, 1998).

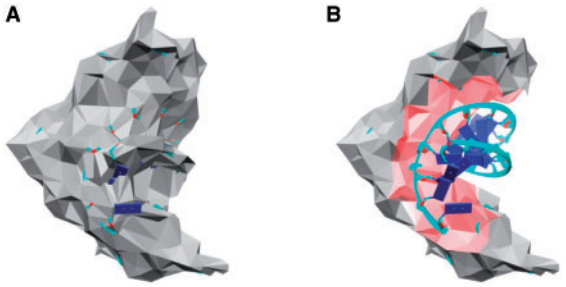

Fig. 3. Protein–DNA complex interface obtained from the alpha shape model. (**A**) The alpha shape model of the protein–DNA complex with the secondary structure inside. (**B**) The alpha shape model of the protein shown with the secondary structure of the DNA chain. The interface surface of the complex is shown in red.

Table 1. List of atom types used in the interface-atom curvature-dependent discriminatory function

C	C_α	C_β	C_δ	$C_{\delta1}$	$C_{\delta2}$
C_ε	$C_{\varepsilon1}$	$C_{\varepsilon2}$	$C_{\varepsilon3}$	C_γ	$C_{\gamma1}$
$C_{\gamma2}$	CH_2	C_ζ	$C_{\zeta1}$	$C_{\zeta2}$	N
$N_{\delta1}$	$N_{\delta2}$	N_ε	$N_{\varepsilon1}$	$N_{\varepsilon2}$	NH_1
NH_2	N_ζ	O	$O_{\delta1}$	$O_{\delta2}$	$O_{\varepsilon1}$
$O_{\varepsilon2}$	O_γ	$O_{\gamma1}$	OH	S_δ	S_γ

The interface surface curvature is represented by the solid angle of the interface atoms in the alpha shape model. The solid angle (Van Oosterom and Strackee, 1983) is defined as follows: let OABC be the vertices of a tetrahedron with an origin at O subtended by the triangular face ABC. Let Φ_{ab} be the dihedral angle between the planes that contain the tetrahedral faces OAC and OBC and we define Φ_{bc} and Φ_{ac} similarly. The solid angle at O subtended by the triangular surface ABC is given by Equation (1). The solid angle of an interface atom is transformed to the range of -1 (cleft) to 1 (knob) using $\cos(\Omega/4)$.

$$\Omega = \Phi_{ab} + \Phi_{bc} + \Phi_{ac} - \pi \tag{1}$$

2.4 The conditional probability formulation

All possible protein–DNA structures can be divided into two sets: $\{C\}$ for the correct structures (native structures) and $\{I\}$ for the incorrect structures (decoy structures). Next, we consider a set of properties from the protein–DNA structure in which the correct structure and the incorrect structure are significantly different. Such properties can be molecular flexibility, electrostatics strength, interatomic distance, etc. In our study, we consider the interface surface curvature of the protein–DNA complex and use a set of features $\{(Sa_i, r_i, a_i)\}$ to characterize the protein–DNA structure. Here, Sa_i stands for the solid angle of the interface atom i, r_i stands for the residue type and a_i stands for the atom type.

We aim to calculate the probability that the structure is in the correct set when given that it has a set of features $\{(Sa_i, r_i, a_i)\}$ which can be expressed as:

$$P(C|\{(Sa_i, r_i, a_i)\}) \tag{2}$$

However, it is difficult to evaluate Equation (2) from the experimental data directly. Therefore, we assume that all the solid angles are independent of one another and the probability of the correctness of a structure can be expressed by the joint probability of the correctness of every interface atom curvature:

$$P(C|\{(Sa_i, r_i, a_i)\}) = \prod_i P(C|(Sa_i, r_i, a_i)) \tag{3}$$

Applying Bayesian theorem to Equation (2), we get:

$$P(C|(Sa_i, r_i, a_i)) \cdot P((Sa_i, r_i, a_i)) = P(C) \cdot P((Sa_i, r_i, a_i)|C) \tag{4}$$

Then we have:

$$P(C|\{(Sa_i, r_i, a_i)\}) = P(C) \cdot \prod_i \frac{P((Sa_i, r_i, a_i)|C)}{P((Sa_i, r_i, a_i))} \tag{5}$$

In this equation, $P(C)$ represents the priori probability of observing a correct protein–DNA structure which is constant. However, it is difficult to evaluate its value that it will not be considered in this article (note that the omission results in a normalized likelihood classification). $P((Sa_i, r_i, a_i)|C)$ stands for the probability of the correct structures that have a set of features (Sa_i, r_i, a_i) which can be calculated using a set of known native structures. We can make observations of the special atom characteristic in a particular solid angle value bin:

$$P((Sa_i, r_i, a_i)|C) = \frac{N_{obs}(Sa_i, r_i, a_i)}{\sum_{Sa_i} N_{obs}(Sa_i, r_i, a_i)} \tag{6}$$

where $N_{obs}(Sa_i, r_i, a_i)$ represents the number of interface atoms with atom type a_i, and residue type r_i in the specific solid angle bin Sa_i in the native structure set. For example, if the number of interface Glycine $C_\alpha(GC_\alpha)$ with solid angle range from 0.45 to 0.55 is found to be 20 in the experimental data, and the total number of interface GC_α with any solid angle is 100, the frequency of GC_α in the solid angle value bin 0.5 is 20/100.

$P((Sa_i, r_i, a_i))$ stands for the probability of any structure having a set of features (Sa_i, r_i, a_i) which can be estimated from the entire training data set including native structures and decoy structures:

$$P((Sa_i, r_i, a_i)) = \sum_{r_i} \sum_{a_i} \frac{N_{obs}(Sa_i, r_i, a_i)}{N_t} \tag{7}$$

where $N_{obs}(Sa_i, r_i, a_i)$ represents the number of interface atoms with atom type a_i, and residue type r_i in the a specific solid angle bin Sa_i in the entire training data. N_t stands for the total number of interface atoms observed in all atom types and all residue types in all solid angle value bins. The calculation procedure is similar to Equation (6).

Considering the limited number of correct structures, low count correction is performed using Sippl's method (Sippl, 1990). Equation (6) is modified accordingly:

$$P_c((Sa_i, r_i, a_i)|C) = \frac{P((Sa_i, r_i, a_i)) + \sigma N_{obs}(Sa_i, r_i, a_i) P((Sa_i, r_i, a_i)|C)}{1 + \sigma N_{obs}(Sa_i, r_i, a_i)} \tag{8}$$

where $P_c((Sa_i, r_i, a_i)|C)$ represents the low count corrected $P((Sa_i, r_i, a_i)|C)$ and the value of σ ensures that the terms have equal weights when $N_{obs}(Sa_i, r_i, a_i) = 1/\sigma$ [σ is set to 1/50 as per Sippl (1990)]. Then we use the negative log to scale the quantities into a small range and obtain the scoring function:

$$S = -\sum_i \ln \frac{P_c((Sa_i, r_i, a_i)|C)}{P((Sa_i, r_i, a_i))} \tag{9}$$

2.5 Apo structure prediction method

We further develop the method for the prediction of the apo version of DNA-binding proteins. A template library is set up which contains 199 different types of protein–DNA complexes as described in Section 2.1. First, we use structural alignment tool TM-align (Zhang and Skolnick, 2005) to compare the target and the structures in the library. The target structure is scanned against the 199 template structures for similar protein structure, and the largest TM-scored template structure is selected. Second, a new structure is created by replacing the protein sequence of the template structure with the aligned target structure. Third, the new structure is scored using the curvature-dependent method. Then we can obtain the prediction result by examining the score.

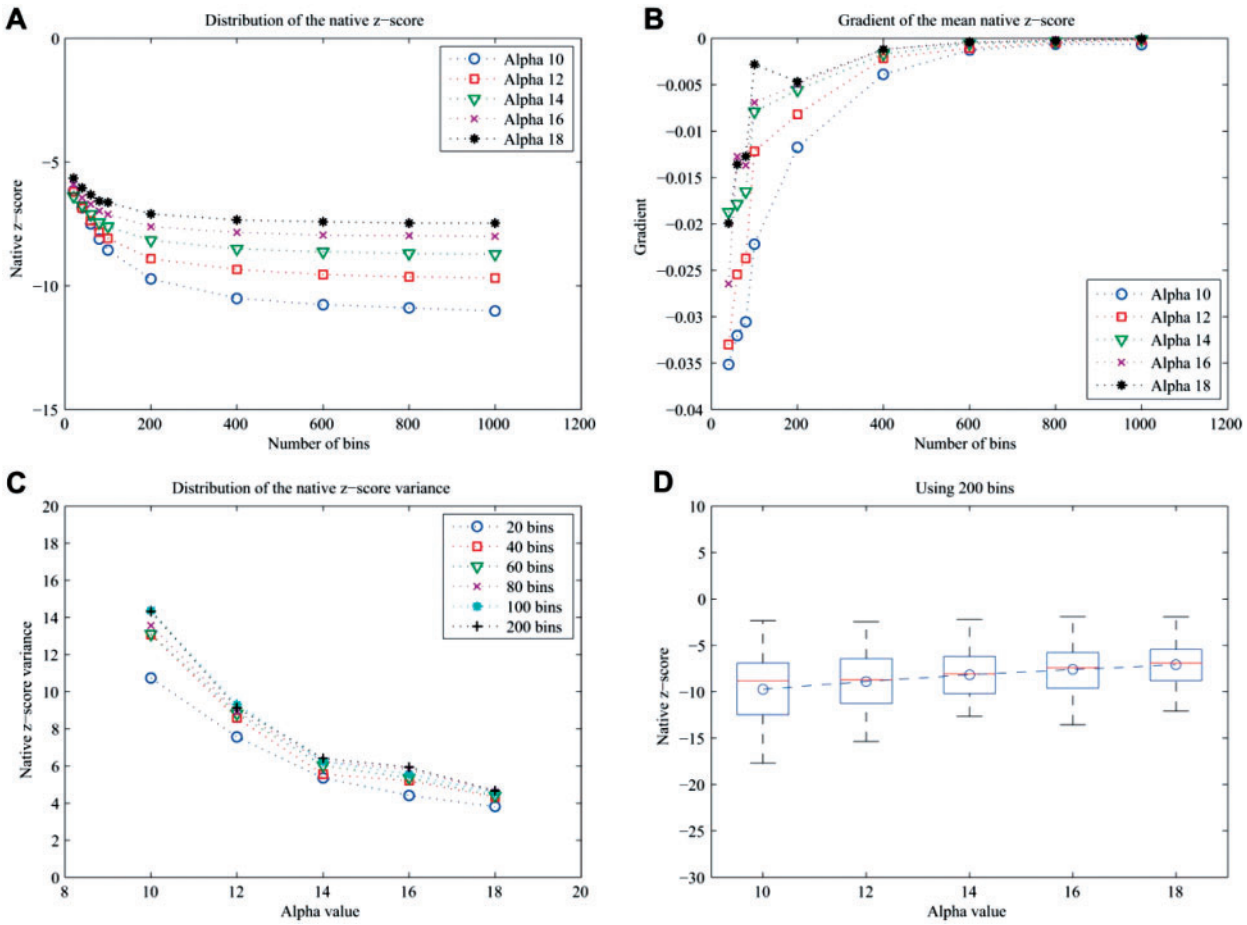

Fig. 4. (**A**) Relation between the mean native z-score and the number of bins with different alpha values. (**B**) The gradient shows that the performance of the local scoring function does not improve much with the number of bins larger than 200. (**C**) Distribution of the native z-score variance with different numbers of bins and different alpha values. (**D**) Performance of the local scoring function using 200 bins and different alpha values. The lower, middle and upper horizontal lines for each boxplot represent the 25th, 50th and 75th percentile native z-score, respectively. Whiskers extend to a distance of 1.5 times the interquartile range. The mean native z-score for each set are represented as points.

3 RESULTS AND DISCUSSION

The experiments are divided into two types. The first type is carried out to evaluate the performance of the interface-atom curvature-dependent conditional probability discriminatory function for discriminating native structures from decoy structures. It is determined by computing the z-score of the native structure relative to all scored decoys (native z-score). The second type is used to test the performance of the proposed method to discriminate the apo version of DNA-binding proteins and non-DNA-binding proteins.

3.1 Optimal parameter selection

We use the 45 selected native structure and 90 000 lowest RMSD decoys as the training set and the 90 000 highest surface-complementarity scored decoys as the scoring set in an experiment for the purpose of searching for the optimal parameters.

One of the parameters affecting the discriminatory result is the alpha value. As mentioned in Section 2.2, the preciseness of the molecular surface captured by the alpha shape model is controlled by the alpha value. Alpha shape with smaller alpha values provides better details of the molecular surface, but the molecular surface will be fragmented if the alpha value is too small. In order to obtain optimal alpha value, we have conducted experiments for a wide range of alpha values and observed that the optimal value ranges from 10 to 20 within the native structure set. Accordingly, we set five different alpha values of 10, 12, 14, 16, 18 and then select the best result. The other parameter influencing the final result is the number of bins needed to separate solid angle values. This value actually controls the resolution of the conditional probability function. With a very large number of bins, most of the structures may be considered to be completely different from each other. In this case, it is meaningless to use the conditional probability formalism. In other words, the number of bins affects the generality and the sensitivity of the discriminatory function. Therefore, different numbers: 20, 40, 60, 80, 100, 200, 400, 800 and 1000 are used to search for the best value.

The result is shown in Figure 4A in terms of mean native z-score with different pairs of alpha values and number of bins. The gradient of the five curves (Fig. 4B) shows that the mean native z-score decreases dramatically with the number of bins at fewer than 200 and

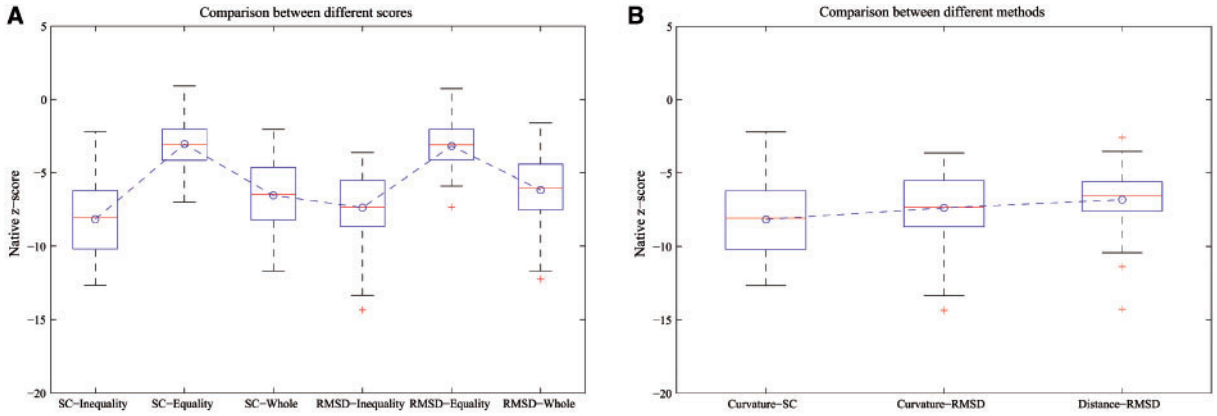

Fig. 5. (**A**) Performance comparison of three different scores. SC represents the experiment of discriminating the native structure from the highest surface-complementarity scored decoys and RMSD represents the experiment of discriminating the native structure from the lowest RMSD decoys. (**B**) Performance comparison between the curvature-dependent function and the distance-dependent function in a decoy discriminating test. Curvature represents the result using curvature-dependent function in the experiment. Distance represents the result using distance dependent function in the experiment. In the boxplot, the outliners are marked as crosses.

the decrement is not obvious when the number of bins is more than 200. Notice that, the larger the number of bins is, the less generality we can achieve. Considering the balance between the generality and the performance of the discriminatory function, 200 is chosen as the optimal number of bins in the following experiments. In order to obtain the optimal alpha value, we calculate the variance of the 45 native z-scores with different number of bins. The variances of native z-score with alpha value 14, 16 and 18 are observably smaller than those of alpha value 10 and 12 as shown in Figure 4C. Checking the result in Figure 4A, the mean native z-score with alpha value 10, 12 and 14 are smaller than those with alpha value 16 and 18. We choose 14 as the optimal alpha value since it produces the overall small z-score and small variance as shown in Figure 4A and C. The performance achieved by using 200 bins and different alpha values is shown in Figure 4D.

3.2 Prediction of protein–DNA complex

In order to evaluate the performance of the curvature-dependent discriminatory function, we conduct an experiment to discriminate the 45 native structures from the 90 000 highest surface-complementarity scored decoys and 90 000 lowest RMSD decoys, respectively. The optimal parameter setting with an alpha value equal to 14 and the number of bins equal to 200 is used in this experiment. We compute three scores in the following experiment. In order to compare with the performance of the distance-dependent method, the correct structure training set contains only 45 native structures to compute the first two scores: (i) leave-one-out cross-validation is applied here which results in an 'equality' score that omits one native structures from the training set and use the other 44 complexes as training set only, and (ii) self-consistent test is used to obtain an 'inequality' score which includes all 45 complexes as the training set. We consider all 199 correct structures as the correct structure training set in the third score which is called the 'whole' score. The 2000 highest surface-complementarity scored decoys and the 2000 lowest RMSD decoys for each native structure are then scored, respectively.

The performance of the discriminatory function is evaluated using the native z-score. The resulting inequality mean native z-score for discriminating the native structures from the highest surface-complementarity scored decoys is -8.17, and -7.38 for the lowest RMSD decoys. The equality mean native z-score for discriminating the native structures from the highest surface-complementarity scored decoys is -3.05, and -3.18 for the lowest RMSD decoys. The whole mean native z-score for the two scoring set is -6.18 and -6.55, respectively. Figure 5A shows the comparison of three different scoring results. We can see that the whole score is not as good as the inequality score. The reason is that the enlargement of the correct training set would reduce the probability of the interface atom from the 45 native structures to appear in the correct atom set. The inequality score shows a better result than the equality score. It is obvious that the performance of the curvature-dependent function is better with the native structure in the training. The reasons are that we use conditional probability in the function and that we can get a more accurate $P\left(\left(Sa_i,\ r_i,\ a_i\right)|C\right)$ with the native structure information. Although the equality z-score is not as good as the inequality z-score, it is acceptable and it still indicates how good the performance of the curvature dependent function is in the discrimination test. Figure 5B shows the comparison between the performance of the curvature-dependent function and the distance-dependent function developed by Robertson and Varani (2007) (best mean native z-score: -6.8).We can see that the curvature-dependent function has a better performance in discriminating the native structures from the highest surface-complementarity scored decoys than its performance in discriminating the native structures from the lowest RMSD decoys. This comparison of performance between two different scoring sets shows that the curvature-dependent function works well in discriminating both the protein–DNA complexes which have resemblant surface condition and complexes which have similar space structure. The comparison between the two different formalisms shows that the curvature-dependent function has better performance than the distance-dependent function in terms of mean

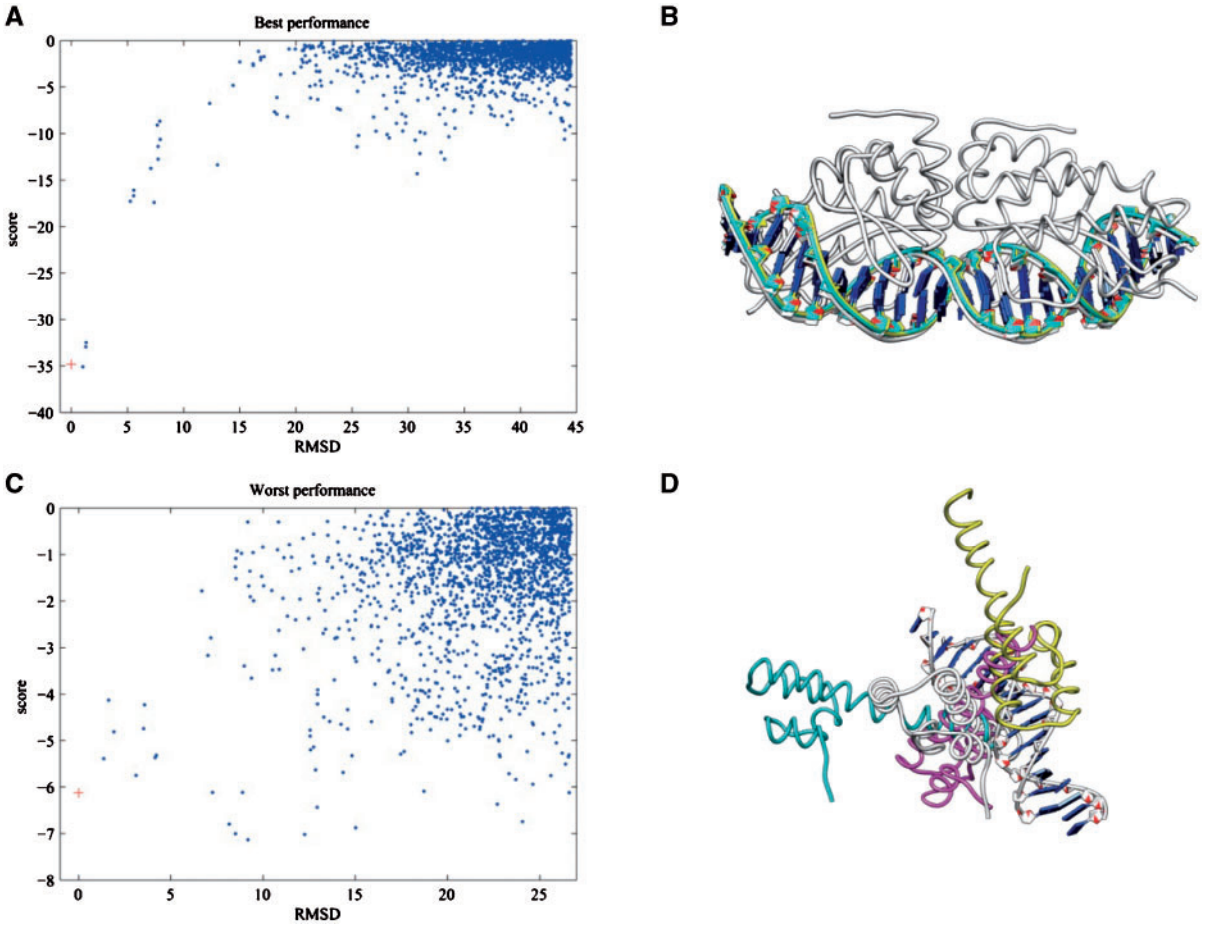

Fig. 6. (**A**) The best performance of the curvature-dependent function in the decoy discriminating test. The local scores for the docking decoys of '1g9z' are plotted relative to their RMSD from the native structure. The native structure is represented as a cross. (**B**) Comparison among the native structure '1g9z' (in white) and the highly similar decoy structures (in different colors). (**C**) The worst performance in the test. Scores for the decoys of '1skn' are plotted relative to their RMSD from the native structure. (**D**) Comparison among the native structure '1skn' (in white) and the highly scored decoy structures (in different colors).

native z-score. The comparison result demonstrates that protein–DNA interface surface curvature also plays an important role in the protein–DNA interaction. The results in this article reaffirm the fact that the surface condition will influence the binding site of the nucleic acid binding protein and proper surface condition should be provided for the protein–DNA interaction to take place.

3.3 Example of representative discrimination results

It is not sufficient to show the discriminatory performance using z-score alone. Therefore, we plot the local score relative to the native RMSD in a diagram. In order to provide a clear picture of the correlation between the local score and the native RMSD, we show the best and the worst performance of discriminating the native structures from the decoy sets (Fig. 6). The best performance shows the relation between the local score and the native RMSD in discriminating the homing endonuclease I-CreI (PDB id: 1g9z) from the decoy set with native z-score of -14.36. Figure 6A shows that those decoys with larger RMSD would have a larger local score and only few decoys with very low RMSD would have a relative low local score which makes it easy to discriminate the correct structure from the decoy structures. We notice that there are three decoys with very small RMSD having a close local score with the native structure. These similar structures are shown with the native structure in Figure 6B. These decoys have a highly relative 3D structure compared to the native structure. The worst performance occurs in discriminating the DNA-binding domain of Skn-1 (PDB id: 1skn) from the decoy set with native z-score of -3.63. Figure 6C shows that some decoys with large RMSD have large local score comparable to the native structure. The three highest local scored structures are shown with the native structure in Figure 6D. We notice that '1skn' is a smaller molecule compared to other native structures in our data set. This is the reason for the poor performance of the algorithm in this case. Because it is a small molecule, the interface area between the protein and DNA is relatively small and not much information can be extracted from it. However, from Figure 6D, we can see that the binding sites of the protein to the DNA in these structures are almost at the same position. This fact shows that even in the worst performance, the curvature-dependent formalism still has precise prediction for the protein binding sites.

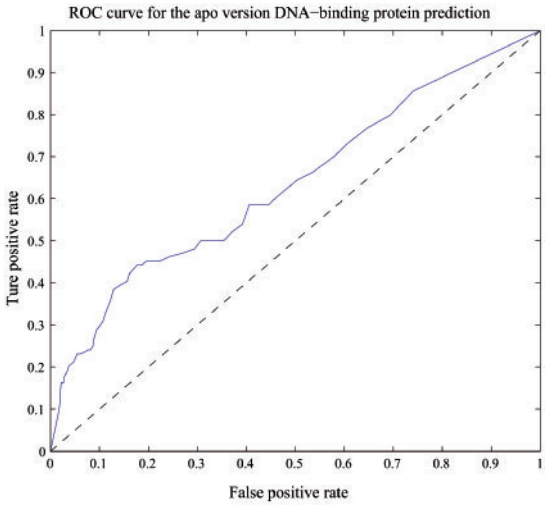

Fig. 7. The ROC curve for the apo version of DNA-binding protein prediction.

Although the worst example does not show a good correlation between the local score and the RMSD as the best example, most of the decoys with high RMSD still have large local scores and the percentage of the decoys being recognized as incorrect structures is 99.6%. It means we can still discriminate a majority of the incorrect structures even in the worst case, which provides the robustness of the curvature-dependent formalism.

3.4 Prediction of apo DNA-binding proteins

We have demonstrated the ability of the curvature-dependent function to discriminate the native structures from the low RMSD and high surface-complementarity decoy structures. In the following experiment, we aim to test the performance of our method for the prediction of apo version of DNA-binding proteins. In this experiment, we consider 104 apo version of DNA-binding proteins and 401 non-DNA-binding proteins. As mentioned in Section 2.1, a library containing 199 typical protein-DNA interaction complexes serve as a template library for the input target. For a given target, any template with sequence identity >35% is excluded from the template library. A search through the library is applied using TM-align, and the largest TM-scored structure is selected as the complex template for the target. We make a new complex by replacing the protein chains in the native structure with the target protein chains. Then the new structures are scored by the curvature-dependent formalism. Different thresholds are set to obtain the receiver operating characteristic (ROC) curve shown in Figure 7. We observe that the thresholds in the range from −1.7 to −2.3 produce the best result. The resulting sensitivity ranges from 48.08% to 44.23% and specificity ranges from 73.82% to 84.29%, which is comparable to the resulting sensitivity 47% from the DBD-Hunter develop by Gao and Skolnick (2008).

3.5 Discrimination of non-DNA-binding proteins

In order to evaluate the specificity of our method, we apply the algorithm to the discrimination of non-DNA-binding proteins from DNA-binding ones. The test data set contains 86 protein–RNA complexes, 106 protein–ligand complexes and 103 protein–protein complexes, as described in Section 2.1. We divide the native protein–DNA set into two parts: the first 100 structures are used in the training data set, and the remaining 99 structures are used as the control set to determine the threshold value. From the result, we observe that our method produces the best result when the threshold is set between −2.8 to −3.3. The resulting specificity ranges from 86.44% to 91.19% and sensitivity ranges from 47.47% to 42.42%. When the threshold is set to −3.3, 12 RNA-binding sites, 2 ligand-binding sites and 12 protein-binding sites are recognized as DNA-binding sites, which results in an accuracy of 86.04%, 98.11% and 88.35%, respectively. We can see that the worst result comes from the RNA-binding sites, and the reason is that protein–DNA complexes and protein–RNA complexes have similar 3D interface characteristics. This shows a limitation of our method, so more robust features in addition to the solid angle need to be investigated to improve the prediction performance. Nevertheless, the result is still acceptable from this test with an accuracy around 86%, which demonstrates the capacity of our method to discriminate non-DNA-binding proteins from DNA-binding ones.

4 CONCLUSION

In this study, we have constructed an interface-atom curvature-dependent discriminatory function for the prediction of protein–DNA interaction. A 3D alpha shape model is introduced to represent the surface of the protein–DNA complex. In this model, solid angle, atom type and residue type are extracted to characterize the interface surface of the protein–DNA complex. We use the conditional probability to form the discriminatory function. The performance of function is tested by discriminating the native structures from a set of docking decoy structures and the near native decoy structures. The interface-atom curvature-dependent formulation shows better performance than the previous pairwise potential method in terms of native z-scores in the same decoy discrimination test. We reaffirm the importance of the geometric complementarity in determining the structure of a complex and show that interface surface curvature plays an important role in protein–DNA interaction. We show that our method is also applicable to the prediction of apo version of DNA-binding proteins.

Our work can be extended in several ways. The alpha shape model should also be useful for the analysis of other types of molecular interactions, such as protein–RNA, protein–ligand and protein–protein complexes, and for the study of multiple proteins, multiple binding sites or a specific family of proteins. These problems would require modeling interface surfaces of different characteristics such as different sizes and the compatibility and cooperativity between these surfaces, thus new surface features in addition to the solid angle may be needed. Recently, it has been shown that interface cluster patterns found based on multiple sequence alignment (Ahmad *et al*., 2008) and graph models (Sathyapriya *et al*., 2008) play an important role in molecular interactions. However, currently these cluster patterns can only incorporate limited 3D information and steric compatibility. The alpha shape model discussed in this article is true 3D in nature, thus it can be used to extract 3D interface patterns. This would be an interesting future research direction.

Funding: Hong Kong Research Grant Council (Projects CITYU 123408 and 123809).

Conflict of Interest: none declared.

REFERENCES

Ahmad,S. *et al*. (2008) Protein-DNA interactions: structural, thermodynamic and clustering patterns of conserved residues in DNA-binding proteins. *Nucleic Acids Res*., **36**, 5922–5932.

Ahmad,S. and Sarai,A. (2004) Moment-based prediction of DNA-binding proteins. *J. Mol. Biol*., **341**, 65–71.

Albou,L.P. *et al*. (2009) Defining and characterizing protein surface using alpha shapes. *Proteins*, **76**, 1–12.

Aloy,P. *et al*. (1998) Modelling repressor proteins docking to DNA. *Proteins*, **33**, 535–549.

Bernauer,J. *et al*. (2007) A new protein-protein docking scoring function based on interface residue properties. *Bioinformatics*, **23**, 555–562.

Cartharius,K. *et al*. (2005) Matlnspector and beyond: promoter analysis based on transcription factor binding sites. *Bioinformatics*, **21**, 2933–2942.

CGAL Computational Geometry Algorithms Library. http://www.cgal.org.

Delaunay,B. (1934) Sur la sphère vide. *Izvestia Akademii Nauk SSSR*, 7, 793–800.

Edelsbrunner,H. *et al*. (1998) On the definition and the construction of pockets in macromolecules. *Discr. Appl. Math.*, **88**, 83–102.

Edelsbrunner,H. and Mucke,E.P. (1994) Three-dimensional alpha-shapes. *ACM T Graphic*, **13**, 43–72.

Fickett,J.W. (1982) Recognition of protein coding regions in DNA sequences. *Nucleic Acids Res*., **10**, 5303–5318.

Gabb,H.A. *et al*. (1997) Modelling protein docking using shape complementarity, electrostatics and biochemical information. *J. Mol. Biol.*, **272**, 106–120.

Gao,M. and Skolnick,J. (2008) DBD-Hunter: a knowledge-based method for the prediction of DNA-protein interactions. *Nucleic Acids Res*., **36**, 3978–3992.

Johnson,P.F. and McKnight,S.L. (1989) Eukaryotic transcriptional regulatory proteins. *Annu. Rev. Biochem.*, **58**, 799–839.

Jones,S. *et al*. (1999) Protein-DNA interactions: a structural analysis. *J. Mol. Biol.*, **287**, 877–896.

Kamei,Y. *et al*. (1996) A CBP integrator complex mediates transcriptional activation and AP-1 inhibition by nuclear receptors. *Cell*, **85**, 403–414.

Li,X. *et al*. (2003) Simplicial edge representation of protein structures and alpha contact potential with confidence measure. *Proteins*, **53**, 792–805.

Liang,J. *et al*. (1998a) Analytical shape computation of macromolecules: I. molecular area and volume through alpha shape. *Proteins Struct. Funct. Genet.*, **33**, 1–17.

Liang,J. *et al*. (1998b) Analytical shape computation of macromolecules: II. Inaccessible cavities in proteins. *Proteins Struct. Funct. Genet.*, **33**, 18–29.

Liu,Z.J. *et al*. (2008) Structure-based prediction of transcription factor binding sites using a protein-DNA docking approach. *Proteins*, **72**, 1114–1124.

Moont,G. *et al*. (1999) Use of pair potentials across protein interfaces in screening predicted docked complexes. *Proteins Struct. Funct. Genet.*, **35**, 364–373.

Murakami,Y. and Mizuguchi,K. (2010) Applying the Naive Bayes classifier with kernel density estimation to the prediction of protein-protein interaction sites. *Bioinformatics*, **26**, 1841–1848.

Nicola,G. and Vakser,I.A. (2007) A simple shape characteristic of protein-protein recognition. *Bioinformatics*, **23**, 2201–2201.

Ofran,Y. and Rost,B. (2003) Analysing six types of protein-protein interfaces. *J. Mol. Biol.*, **325**, 377–387.

Pabo,C.O. and Nekludova,L. (2000) Geometric analysis and comparison of protein-DNA interfaces: why is there no simple code for recognition? *J. Mol. Biol*., **301**, 597–624.

Pontius,J. *et al*. (1996) Deviations from standard atomic volumes as a quality measure for protein crystal structures. *J. Mol. Biol.*, **264**, 121–136.

Poupon,A. (2004) Voronoi and Voronoi-related tessellations in studies of protein structure and interaction. *Curr. Opin. Struct. Biol.*, **14**, 233–241.

Robertson,T.A. and Varani,G. (2007) An all-atom, distance-dependent scoring function for the prediction of protein-DNA interactions from structure. *Proteins*, **66**, 359–374.

Sael,L. *et al*. (2008) Rapid comparison of properties on protein surface. *Proteins*, **73**, 1–10.

Samudrala,R. and Moult,J. (1998) An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *J. Mol. Biol*., **275**, 895–916.

Sancar,A. *et al*. (2004) Molecular mechanisms of mammalian DNA repair and the DNA damage checkpoints. *Annu. Rev. Biochem*., **73**, 39–85.

Schneider,T.D. *et al*. (1986) Information content of binding sites on nucleotide sequences. *J. Mol. Biol.*, **188**, 415–431.

Siggers,T.W. *et al*. (2005) Structural alignment of protein-DNA interfaces: Insights into the determinants of binding specificity. *J. Mol. Biol.*, **345**, 1027–1045.

Sippl,M.J. (1990) 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.

Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.

Van Oosterom,A. and Strackee,J. (1983) The solid angle of a plane triangle. *IEEE Trans. Biomed. Eng.*, **30**, 125–126.

Zhang,Y. and Skolnick,J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res*., **33**, 2302–2309.

Zhou,W. and Yan,H. (2010) Relationship between periodic dinucleotides and the nucleosome structure revealed by alpha shape modeling. *Chem. Phys. Lett.*, **489**, 225–228.