Analysis and classification of DNA-binding sites in single-stranded and double-stranded DNA-binding proteins using protein information

Wei Wang¹, Juan Liu¹, Yi Xiong², Lida Zhu¹, Xionghui Zhou¹

¹School of Computer, Wuhan University, Wuhan, Hubei, People’s Republic of China
²Department of Biological Sciences, Purdue University, West Lafayette, Indiana 47907, USA
E-mail: liujuan@whu.edu.cn

Abstract: Single-stranded DNA-binding proteins (SSBs) and double-stranded DNA-binding proteins (DSBs) play different roles in biological processes when they bind to single-stranded DNA (ssDNA) or double-stranded DNA (dsDNA). However, the underlying binding mechanisms of SSBs and DSBs have not yet been fully understood. Here, the authors firstly constructed two groups of ssDNA and dsDNA specific binding sites from two non-redundant sets of SSBs and DSBs. They further analysed the relationship between the two classes of binding sites and a newly proposed set of features (residue charge distribution, secondary structure and spatial shape). To assess and utilise the predictive power of these features, they trained a classification model using support vector machine to make predictions about the ssDNA and the dsDNA binding sites. The author’s analysis and prediction results indicated that the two classes of binding sites can be distinguishable by the three types of features, and the final classifier using all the features achieved satisfactory performance. In conclusion, the proposed features will deepen their understanding of the specificity of proteins which bind to ssDNA or dsDNA.

1 Introduction

DNA binding proteins play the crucial roles in a wide variety of biological processes, such as DNA replication, transcription and nucleosome remodelling. DNA binding proteins can be categorised into two types: single-stranded DNA-binding proteins (SSBs) and double-stranded DNA-binding proteins (DSBs). The former bind to single-stranded DNA (ssDNA) molecules, whereas the latter bind to double-stranded DNA (dsDNA). Some studies showed that the main functions of SSBs include stimulating DNA damage and activating transcription [1, 2] and DSBs are more ubiquitous and essential for many biological functions. It is known that SSBs bind specifically to ssDNA, while DSBs bind specifically to dsDNA. Since there exists the different physical and chemical properties between ssDNA and dsDNA molecules, it is necessary for us to explore the differences between SSBs and DSBs, which contribute to their specifically binding to ssDNA and dsDNA, respectively. With the progress of the structural genomics project, an increasing number of protein–DNA complex structures are available in Protein Data Bank [3]. Based on the structural information of these complexes, the computational analysis of the binding sites on SSBs and DSBs will aid in understanding the molecular mechanisms of the binding process.

Most of the existing bioinformatics methods focus on the interaction mechanisms of protein–RNA, protein–protein and protein–dsDNA by classification, clustering, network modelling and statistics modelling [4–11]. However, few studies have been conducted to identify the distinct information between the SSBs and DSBs in the protein–DNA complexes. Recently, several studies have discussed biological mechanisms of SSBs and DSBs from the view of molecular biology, such as structure, evolution and biophysical characterisation [12–14]. For example, Zasedateleva et al. [15] applied gel-based oligonucleotide microarray method to analyse the binding specificity and affinity of SSBs to ssDNA. Edsö et al. [16] had found two distinct types of Rap1 and Cdc13 DNA-binding protein, and both of them depend on different DNA-binding domains for specific binding to the double-stranded or the single-stranded telomeric DNA. However, to the best of our knowledge, there is no related work to analyse the discriminative characteristics of interfaces systematically between SSBs and DSBs using feature-based methods. Therefore, based on the biological characteristics, the computational methods are required to systematically compare the diverse characteristics of interfaces between SSBs and DSBs using feature-based methods. To address this question, our work proposed a variety of features, which can have important difference between SSBs and DSBs, with binding to the different types of DNA (ssDNA or dsDNA).

Our previous work, Zeng et al. [17] proposed an interfacial biclique pattern between DSBs and SSBs, and we found that DSBs and SSBs have their own right residues at the right places for the binding preference and association with
nucleotides. In this study, our key idea is to examine properties of conserved interface residues, secondary structure and topological spatial distance of interface regions. It is known that the physicochemical properties of interface residues play key roles in their binding to DNA, such as the accessible surface area, electrostatic potentials, hydrogen-bonding potential, van der Waals contacts, water-mediated bonds and so on [18–21]. Previous studies have already been reported that the DNA-binding sites are usually positively charged, compensating for the negative charges on the DNA backbone [22–24]. Our study will further explore whether the residue charges of interfaces differ between SSBs and DSBs. The other important features are the distribution of secondary structure in the interfaces, and the features were used in the prediction of binding residues of protein–RNA [24] and protein–protein [25] interactions. Our experimental results suggest that DSBs prefers α-helix in the interfaces, but SSBs prefers the β-strand, and DSBs and SSBs have no significant difference in the coil structure. Finally, we calculated the geometrical feature of interfaces based on the 3D structures of the SSBs and DSBs. As a result, the interfaces of DSBs usually present wave shape, while the interfaces of SSBs are usually smaller and irregular, which adapt themselves to binding to the swinging flexible ssDNA.

To assess and utilise the effectiveness of these features, a variety of classification models were trained by different types of machine learning methods using leave-one-out validation and 10-fold cross-validation on all data or non-redundant data, and the classification models results are shown in the four supplementary tables. Our final classification model employed the support vector machine (SVM) to classify the binding residues of SSBs and DSBs using these proposed features, with an accuracy of 85.4%, F-measure of 0.86, MCC of 0.72 and AUC of 0.86. Our experimental results have also shown the prediction power of the individual and the combined features, which are promising for biologist to further study the mechanism of the binding specificity.

## 2 Materials and methods

### 2.1 Data sets

In this study, the protein–DNA complexes were downloaded from the PDB (http://www.rcsb.org/pdb/). We collected a total of 889 DSBs and 74 SSBs complexes from the PDB. Then, the non-redundant set was obtained by using the PISCES program (http://www.dunbrack.fccc.edu/Guoli/PISCES.php) [26] with sequence identity no more than 30%, and a resolution better than 3 Å if the structures are solved by X-ray crystallography, and a minimum chain length of 40 amino acid residues. The non-redundant set consists of 272 non-redundant protein–DNA complexes, 235 of which are protein–dsDNA complexes (the set is referred as DSBs-235) and the rest of which are protein– ssDNA complexes (the set is referred as SSBs-37).

We focused on identifying and analysing the DNA binding regions on the protein surfaces which interact with dsDNA or ssDNA. Following the same criterion suggested in previous studies [19, 20, 27], a residue is regarded as interacting with DNA when the Euclidean distance between any heavy atom of the residue and any heavy atom of the DNA molecule is smaller than 4.5 Å. Under this criterion, the training dataset contains 6454 and 1113 interface residues in DSBs and SSBs, respectively.

### 2.2 Descriptors for interface propensity

For each DNA binding protein, the interface residues are directly involved in protein–DNA interaction. DNA-binding residues provide clues to understanding the mechanism of protein–DNA recognition on the SSBs and DSBs. The processes of recognition and interactions are governed by two major factors: geometrical and biochemical characteristics. The geometrical characteristic signifies the complementarity of the protein–DNA binding surfaces, whereas the biochemical characteristic embodies the existence of preponderant chemical interactions, particularly the electrostatic interactions, that is implied in the formation of salt bridges and of hydrogen bonds. Therefore the binding regions of the SSBs/DSBs provide certain features or attributes to adapt themselves to the different types of DNA molecular surfaces. In order to discriminate SSBs from DSBs, we extracted the features based on conserved sequence, residue charge, secondary structure and spatial shape features in protein–DNA interface.

#### 2.2.1 Weighting properties of conserved residues:

Evolutionarily restrained amino acid positions, which are often referred to as evolutionary conserved, are often important to maintain the structure and function of proteins. For a given protein sequence, the values of conservation were generated by multiple sequence alignment of all homologous proteins obtained from HSSP database [28]. The conservation of residues was estimated by the Shannon entropy of the aligned sequences at position $i$ [29] where $E(i) = - \sum_{a=1,7} p_n \ln(p_n)$

$$E(i) = \sum_{a=1,7} p_n \ln(p_n)$$

$E(i)$ is the Shannon entropy for residue $i$ in a protein–DNA interface, the score is also a measure of sequence conservation at each position. The lower value of $E(i)$ represents the higher degree of weighting in the $i$th position. $p_n$ is the frequency of amino acids of class $n$ at the $i$th position. The 20 types of amino acids are grouped into the following seven classes based on the similarity of environment in protein structure: (i) Ala, Val, Leu, Ile, Met and Cys; (ii) Gly, Ser and Thr; (iii) Asp and Glu; (iv) Asn and Gin; (v) Arg and Lys; (vi) Pro, Phe, Tyr and Trp; and (vii) His [30] where $W_i = 1 - \frac{E_i - \min(E)}{\max(E) - \min(E)}$ (2)

where $W_i$ is the weighting factor of amino acid in position $i$, $\min(E)$ and $\max(E)$ denote the minimum and maximum of entropy of all residues in the entire sequence. For each interface residue, the $W_i$ was calculated to normalise the evolutionary conservation.

**Residue charge distribution:** Many studies have demonstrated the importance of electrostatic interactions in protein–DNA [31, 32]. The DNA-binding sites generally show the most remarkable property of positive electrostatic charge, in complement to the backbone phosphates with negative charges on the surface of DNA [22]. Therefore electrostatic interaction mainly adjusts the attraction between the positively charged protein interface and the negatively charged DNA surface. Here, we verify whether or not the interface charge should be the underlying difference between SSBs and DSBs. In the experiment, the positively charged of residues such as arginine and lysine
were assumed as $+1$ in terms of their charge scores, and histidine residue was assumed as $+0.5$ since its positive charge depends on their local environments. The negatively charged residues, such as asparagine and glutamine, were assumed as $-1$. Therefore we calculated the average charge score of conservation interface residues, and used the score to represent the overall charge distribution of the interface

$$E = \frac{1}{n} \sum_{i=1}^{n} P_i W_i$$

where $E$ is the average of weighted charge scores in the total of $n$ interface residues, and $P_i$ is the charge score of amino acid for position $i$.

2.2.2 Secondary structure composition of interfaces: Recently, the secondary structure of protein–DNA interaction has been considered as an important descriptor of their interaction sites [18]. Some studies had applied secondary structure to characterise the surface residue, and defined the surface patches with some local features [33]. Most of SSBs contain a common structural domain called the OB-fold (oligosaccharide or oligonucleotide) that usually enables them to bind ssDNA [34], which has a five-stranded $\beta$-sheet coiled to form a closed $\beta$-barrel, and an $\alpha$-helix as the cap on the barrel. OB-fold of SSBs has been shown to be important in specific protein–ssDNA interactions, and the $\alpha$-helix of the interface surface is also an important characteristic in the protein–dsDNA interaction. Therefore the secondary structure provides a strong hint to the role for specific recognition of their unique partner (dsDNA against ssDNA).

In our study, the secondary structure was introduced to describe the interfaces of DNA-binding proteins. The secondary structure information was obtained from the S2C database (http://www.dunbrack.fccc.edu/Guoli/s2c/index.php). We classified them into three structural classes: H: $\alpha$-helix, E: $\beta$-strand and C: coil. Type G (310Helix) and H are considered as $\alpha$-helix; type E and B (bridge) are considered as strand; type T (turn) and C are considered as coil. Here, in order to extract the interface features, we employed the probability of the secondary structure to different classes in the interface

$$S_i = \frac{n_i}{n}, \quad i \in \{H, E, C\}$$

where $S_i$ represents the frequency of residues assigned as $\alpha$-helix, stand or coil structures, and $n_i$ is the number of residues in the secondary structure ($\alpha$-helix, stand or coil), and $n$ is the total number of residues in the interface.

2.2.3 Geometrical characteristic of protein–DNA recognition: We assumed that the binding interfaces on SSBs and DSBs are expected to differ in their geometrical features, taking into consideration the different natural properties of ssDNA and dsDNA. The geometrical features of the protein–DNA interface should explain the fundamental principles of molecular recognition. As shown in Fig. 1a, the interfaces of DSBs are dented and protruded wave shape, because of the adaptability of binding to the major or minor groove of dsDNA. It is generally known that SSBs bind to distorted, melted or unwound DNA duplexes [35], whereas the ssDNA presents swinging flexible structure (Fig. 1b). The dsDNA molecules are more inflexible than ssDNA. According to the protein–ligand interactions which follow the ‘lock and key’ paradigm, the DSBs are significantly different with SSBs in the interface structure. Figs. 1c, d shows the geometry of the interface on the DSB and SSB, respectively. It can be noted that DSBs have striking fluctuations of residues, however, the interface
of SSBs is relatively flat. Hence we proposed a novel spatial shape feature to characterise the distribution of interfaces. The spatial distances of binding residues were used to characterise the structure of the protein–DNA interface.

We calculated the differential height of interface residues, as the height is related to the shape characteristic of the interface, which can be represented as the following formulas

\[ C_{\text{protein}}(x, y, z) = \frac{1}{n} \sum_{i=1}^{n} C(x_i, y_i, z_i) \]  
(5)

where \( C_{\text{protein}}(x, y, z) \) represents the geometric centre of all \( n \) amino acids in the protein chain, which can be as the base point for measuring the concave or convex point in the surface, and the centre \( C(x_i, y_i, z_i) \) is the coordinate of \( i \)th atom in protein structures

\[ D(x_i) = \sqrt{(x_i - x)^2 + (y_i - y)^2 + (z_i - z)^2} \]  
(6)

where \( D(x_i) \) represents the Euclidean distance from the centre \( C(x_i, y_i, z_i) \) of residue \( i \) to \( C_{\text{protein}}(x, y, z) \) of the protein

\[ \text{AvgD}_{\text{interface}} = \frac{1}{k} \sum_{i=1}^{k} D(x_i) \]  
(7)

\[ \text{SpD} = \frac{1}{k} \sum_{i=1}^{k} |D(x_i) - \text{AvgD}_{\text{interface}}| \]  
(8)

where \( \text{AvgD}_{\text{interface}} \) is defined as the average distance of the interface of the proteins, \( k \) is the number of binding sites and \( \text{SpD} \) represents the spatial distance of the interface, which is also the fluctuation index of interface residues.

### 2.2.4 SVM for training and prediction

SVM has been widely used for classification in bioinformatic applications. It is the supervised learning method, which requires training data with known category labels. In SVM classifiers, the input data are non-linearly mapped from low-dimensional space to a high-dimensional space and the best hyper plane separate the original data objects into two classes [36]. In our classification tasks, SVM classifiers were constructed using the aforementioned features of interfaces as input.

In this study, SVM classifiers were implemented using Matlab 2012a SVM package with the radial basis function as the kernel. The model was evaluated by 10-fold cross-validation experiments. The overall performance was obtained by averaging the performance of the ten subsets (at the fold level). Owing to the heavily unbalanced datasets, the samples of the majority class were selected by down sampling strategies in the training phase. The performance of our method was assessed by the accuracy (ACC), sensitivity (SE), specificity (SP), F-measure \( F_1 \), AUC and Matthew’s correlation coefficient (MCC). These metrics can be calculated by the number of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) for each classifier. The performance measures are defined as

\[ \text{ACC} = \frac{TP + TN}{TP + FN + TN + FP} \]  
(9)

\[ \text{SE} = \frac{TP}{TP + FN} \]  
(10)

\[ \text{SP} = \frac{TN}{TN + FP} \]  
(11)

\[ F_1 = \frac{2 \times TP}{2 \times TP + FP + FN} \]  
(12)

\[ \text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \]  
(13)

### 3 Results and discussion

In this section, in order to understand the specificity of DNA binding between SSBs and DSBs, we firstly analysed three types of features in the interfaces, and employed SVM algorithms to construct a series of classifiers for evaluating the classification performance of the different features.
3.1 Propensity of residue types in interfaces

The goal of this section is to analyse the propensity of 20 types of amino acid residues in interfaces between SSBs and DSBs. In Fig. 2, the distribution of the residue composition in the interfaces presents the difference between SSBs and DSBs (Fig. S1 annotates the calculation in supplementary information). The types of the Lys and Arg residues in interfaces of DSBs have a higher propensity than that of SSBs in the interface. Based on the fact, we hope that the charge distribution of interfaces may be used to distinguish the binding sites on DSBs and SSBs.

To examine whether the average residue charge in the interface and the DNA base pairing properties are important for the intermolecular interactions in the different types of proteins, we measured the charge distribution of conservative residues in the interfaces. In Fig. 3, the histogram shows the charge distribution of conservative residues in interfaces. From the histogram, we can observe the significant difference of distributions on two types of proteins. For example, 28% of interfaces from DSBs have average charge values larger than 0.3, whereas only 7% of the SSBs are in the range. Although the average charge values are smaller than 0.2, there are 40% of the DSBs proteins and 60% of the SSBs. It is concluded that DSBs tend to have a higher positive charge in their interfaces than that of SSBs.

3.2 Distribution of secondary structure

The interface usually interacts with DNA via the secondary structure motif in protein–DNA interaction. The residues of interfaces were assigned as the one of three classes (α-helix, β-strand and coil). In the interfaces of DSBs and SSBs, the propensities of the secondary structure show the difference (shown in Fig. 4), which can have an important effect of specific binding to the different DNA stranded.

In Fig. 4, it shows the frequency of α-helix, β-strand and coil in the interface residues. It is indicated that more than 40% residues in DSBs are α-helix, significantly higher than the 28% in SSBs. The frequency of β-strand is lower when compared to helical residues in both SSBs and DSBs (about 25 and 15%, respectively). The coil (including turns, coils...
and loops) appears in high frequency in both, but the coil of SSBs has only slightly higher frequency than DSBs (shown in Fig. 4). These results suggest that DSBs prefer α-helix in the interfaces, but SSBs prefer the β-strand. However, there is no significant difference in the interfaces of DSBs and SSBs in terms of coil structures, and the frequency of coil structures in the interfaces of SSBs is slightly higher than that on DSBs. The frequency of different types of secondary structures may be a potential feature to differentiate the DSBs from SSBs in the interfaces.

As shown in Fig. 5a, about 15% of DSBs proteins have a higher abundance (75%) of α-helix residues, and about 45% of DSBs proteins have more than 50% α-helix residues, which indicates that α-helix plays a greater role in DSBs than SSBs. In the group of DSBs, there are three major families of structural DNA-binding domains: helix-turn-helix, zinc finger and leucine zippers. Their mainly characteristic structures are α-helix motif in the interfaces, such as the HTH structure consisting of two α-helices. In Fig. 5b, about 20% of SSBs proteins have more than 50% residues in β-strand structures, which are higher than that of DSBs. This result is an agreement with the previous finding that many SSBs contain a common structural domain called the OB-fold that mainly consists of a five-stranded β-sheet. In Fig. 5c, about 15% of SSBs proteins have more than 75% residues in non-regular coil structures.

3.3 Geometrical shape of interfaces

In previous sections, the features of residue charge distribution and secondary structure show some difference between SSBs and DSBs. In this section, we aim to test whether or not the geometrical shape of SSBs interface is different from that of DSBs.

Fig. 6 presents that the spatial distance of the interfaces shows a small value (<5) in more than 60% of SSBs proteins, which indicates that the interfaces are usually flat. On the other hand, more than 30% of DSBs proteins have large value (>15) of the spatial distance in the interfaces, whereas a small amount of SSBs proteins have large spatial distance in their interfaces. According to the shape complementarity of the receptors and ligands, if the interface of two molecules binds together closely, the interface of SSBs should be usually tighter and smaller cleft than DSBs. This result is also supported by the fact that the interfaces of DSBs commonly penetrate into the major and minor grooves of dsDNA so that the convex and concave faces of DSBs have striking fluctuations of spatial positions for bound dsDNA closely. Therefore this feature should be an important factor to distinguish the interfaces of DSBs from SSBs.

3.4 Prediction performance of the SVM classifier

In this section, we attempt to evaluate the predictive power of the individual features and their combination using an SVM model. Table 1 shows the predicted performance on the

<table>
<thead>
<tr>
<th>Feature</th>
<th>ACC</th>
<th>SE</th>
<th>SP</th>
<th>AUC</th>
<th>MCC</th>
<th>F₁</th>
</tr>
</thead>
<tbody>
<tr>
<td>RC</td>
<td>0.6239</td>
<td>0.5150</td>
<td>0.7358</td>
<td>0.6427</td>
<td>0.2681</td>
<td>0.5832</td>
</tr>
<tr>
<td>H</td>
<td>0.6671</td>
<td>0.7100</td>
<td>0.6233</td>
<td>0.6717</td>
<td>0.3644</td>
<td>0.6947</td>
</tr>
<tr>
<td>S</td>
<td>0.6470</td>
<td>0.5925</td>
<td>0.7008</td>
<td>0.6489</td>
<td>0.3152</td>
<td>0.6424</td>
</tr>
<tr>
<td>C</td>
<td>0.6141</td>
<td>0.9150</td>
<td>0.3058</td>
<td>0.5953</td>
<td>0.2964</td>
<td>0.7082</td>
</tr>
<tr>
<td>SpD</td>
<td>0.7668</td>
<td>0.6517</td>
<td>0.8842</td>
<td>0.7786</td>
<td>0.5660</td>
<td>0.7487</td>
</tr>
<tr>
<td>RC+H</td>
<td>0.6880</td>
<td>0.6592</td>
<td>0.7133</td>
<td>0.6901</td>
<td>0.3888</td>
<td>0.6856</td>
</tr>
<tr>
<td>H+SpD</td>
<td>0.8389</td>
<td>0.7950</td>
<td>0.8825</td>
<td>0.8439</td>
<td>0.6953</td>
<td>0.8401</td>
</tr>
<tr>
<td>H+S+SpD</td>
<td>0.8112</td>
<td>0.7083</td>
<td>0.9175</td>
<td>0.8289</td>
<td>0.6561</td>
<td>0.7992</td>
</tr>
<tr>
<td>all features</td>
<td>0.8542</td>
<td>0.8483</td>
<td>0.8725</td>
<td>0.8416</td>
<td>0.6871</td>
<td>0.8386</td>
</tr>
</tbody>
</table>

training dataset with various combinations of features using 10-fold cross validation. In total, there are five classification models constructed with individual features. As shown in Table 1, the highest prediction accuracy is 76.68% using SpD (spatial distance of the interface) feature, and its AUC, MCC and $F_1$ were 0.7768, 0.5660 and 0.7487, respectively. The result suggests that the spatial structural feature of the interfaces is the important feature to distinguish SSBs from DSBs. In addition, the residue charge distribution and secondary structure also show good classification ability (inferior to the predictive power of the spatial structural feature) to predict the two types of interfaces. In the experiment, the performance was remarkably improved, when two or more features were combined as input vectors for classification. With all features as input, our SVM classifier achieves high test performance, with an ACC of 85.42%, AUC of 0.8623, $F_1$ of 0.8615 and MCC of 0.7175. Therefore, in our work, the final classification model was constructed by combining all the features (ROC curve of the final classifier is shown in Fig. 7).

4 Conclusions

In this study, we have performed a feature-based analysis and constructed a classification model for the classification of binding residues in the interfaces of SSBs and DSBs. In this model, the residue charge distribution, secondary structure and spatial distance features have been verified to effectively characterise the interface surfaces of the protein–ssDNA and protein–dsDNA complexes. We employed the SVM to construct the discriminatory function, and the features are verified by leave-one-out validation with the different classifiers in the datasets (Results in supplementary information). The electrostatic interaction of protein–DNA has been useful for the analysis of many types of molecular interactions, such as protein–ligand, protein–RNA and protein–protein complexes, but paying less attention to the specific types of SSBs and DSBs. Our study indicates that the residue charge distribution has an obvious difference in the interfaces of SSBs and DSBs. The positively charged residue lysine and arginine occur more frequently in DSBs than SSBs. In addition, both the interface shapes and the secondary structures of interfaces are also playing an important role in determining the binding specificity of ssDNA and dsDNA.

We believed that the features may effectively reveal the implied difference between SSBs and DSBs to facilitate future discovery of the biological mechanisms. The proposed features will deepen our understanding of the specificity of proteins which bind to ssDNA or dsDNA. Meanwhile, the interface features should also be useful for the analysis of other types of molecular interactions, and can potentially be expanded to predict other protein functions, such as protein–protein and protein–RNA interaction residues.

5 Acknowledgments

This work is supported by grants from the National Science Foundation of China (61272274); Program for New Century Excellent Talents in Universities (NCET-10-0644), the Open Research Fund of State Key Laboratory of Hybrid Rice (Wuhan University) (KF201301) and the Fundamental Research Funds for the Central Universities (no. 2012211020204).

6 References