Chapter 13

Survey of Computational Approaches for Prediction of DNA-Binding Residues on Protein Surfaces

Yi Xiong, Xiaolei Zhu, Hao Dai, and Dong-Qing Wei

Abstract

The increasing number of protein structures with uncharacterized function necessitates the development of in silico prediction methods for functional annotations on proteins. In this chapter, different kinds of computational approaches are briefly introduced to predict DNA-binding residues on surface of DNA-binding proteins, and the merits and limitations of these methods are mainly discussed. This chapter focuses on the structure-based approaches and mainly discusses the framework of machine learning methods in application to DNA-binding prediction task.

Key words Structure-based function prediction, Functional annotation, DNA-binding residue, Machine learning method

1 Introduction

Protein-DNA interactions play vital roles in various biological activities such as gene regulation, transcription, DNA repair, and DNA packaging. It has been estimated that DNA-binding proteins represent 2–3% and 6–7% of all proteins encoded in prokaryotic and eukaryotic genomes, respectively [1–3]. As of early May 2017, a total of 3915 protein-DNA complex structures have already been deposited in the PDB (http://www.rcsb.org/). Due to the important role of DNA-binding proteins, a variety of computational approaches have been proposed for prediction of DNA-binding function from protein sequences or structures in the past decades. The first category of methods utilizes a comparative approach to infer protein function by global/local sequence or structural similarity [4–7]. While sequence comparison methods are powerful and widely adopted for function inference, structure comparison methods are more sensitive to detect remote homologs with low or no sequence similarity. However, significant sequence or structural similarity does not necessarily dictate identical function, since many proteins have functional divergence during the course of

Tao Huang (ed.), *Computational Systems Biology: Methods and Protocols*, Methods in Molecular Biology, vol. 1754,
https://doi.org/10.1007/978-1-4939-7717-8_13,

evolution [8]. The second category of methods employs various machine learning techniques such as logistic regression, neural network, and support vector machines to train classification models with physicochemical, evolutionary, electrostatic, and structural features to distinguish DNA-binding proteins from other proteins [9–11]. If a protein is identified as DNA-binding protein, we will further develop the methods to detect which are DNA-binding residues on the given protein. The computational methods for prediction of DNA-binding residues can be categorized into two groups: (1) methods based on sequences and (2) methods based on structures. The first group of methods includes the sequence comparison methods and machine learning methods based on sequence-derived features. The sequence-based methods have the wide scope of application, since they require only sequence information as query input, rather than structures which have not been experimentally determined for most of the proteins encoded by genomes. However, these sequence-based predictors have at least two major limitations. One problem with sequence-based predictors is that amino acids that are sequential neighbors are not necessarily close in space to confer DNA-binding function. The other problem is that sequence information provides few clues to the interaction sites and is not sufficient for accurate prediction of DNA-binding residues. In fact, the information derived from protein structures is helpful for predicting DNA-binding function. In recent years, an increasing number of protein with unknown function are solved due to the efforts of structural genomics projects. Functional annotations of these targets are particularly challenging since many targets in structural genomics have low sequence identity to the proteins with known function. Therefore, it is urgent to develop computational approaches that utilize not only sequence but also structural information for function prediction. Since proteins always interact with other proteins or DNA/RNA molecules through their surface, we will focus on the review of computational approaches for prediction of DNA-binding residues on protein surface on the assumption that the given protein structure interacts with DNA.

2 Definition of Surface and DNA-Binding Residues

For any prediction model, the first step is to construct a reliable or benchmark data set of DNA-binding residues and nonbinding residues on the representative set of DNA-binding protein chains. It is relatively straightforward to determine DNA-binding residues if the three-dimensional (3D) structure of a protein-DNA complex is already solved. A residue is taken as a surface residue if its solvent-accessible surface area (SASA) is at least 10% of maximum values in a tripeptide state. The SASA of residues were calculated in each

protein multimer in the absence of DNA. A surface residue is labeled as a binding residue if it satisfies one of the three definition approaches as follows. The most frequently used method to assign DNA-binding residues is based on a minimum distance cutoff of atoms between amino acids in a protein and nucleotides in DNA. However, different distance cutoffs lead to accuracy variations, while a single cutoff biases certain prediction programs [12]. Most studies used a cutoff distance (i.e., 3.5–6 Å) between atoms of amino acids and nucleotides to assign DNA-binding residues on proteins. The second approach to assign binding residues is based on the difference of the solvent-accessible surface areas when the structure of DNA-binding protein transforms from the isolated (the protein without DNA present) to the complexed state (the protein with DNA present). The third definition is based on the scoring function using AMBER potential to calculate the interaction free-energy between atoms in protein and DNA molecules [13]. The residues with the energy score less than −1 kcal/mol are identified as DNA-binding residues. The scoring function-based approach can quantitatively measure the interaction strength, in comparison to the distance-based approach in which the residue-nucleotide pairs with different distances have been treated in the same manner.

3 Structure-Based Methods for Prediction of DNA-Binding Residues

For prediction of DNA-binding residues, the structure-based methods can be categorized into three main types. The first type is the template-based methods based on the structural alignment [4, 14] or dynamic alignment [15]. The second type is based on the physical principles that ultimately govern protein-DNA interactions, such as knowledge-based [5] and docking-based methods [16]. The third type is feature-based methods using various machine learning technologies, which are elaborated in detail in the next section.

4 Machine Learning Methods for Prediction of DNA-Binding Residues Using Structure-Based Features

4.1 Representation of Environment of DNA-Binding Residues

As an input vector for training or testing by machine learning technologies, the sample of DNA-binding residue is commonly represented by the properties of the target residue and its neighbor residues to include the environmental information of the target residue. Similar to the sequence window used by sequence-based methods, structure-based methods utilize different types of structural windows or patches to incorporate the neighbor information of the target residue in 3D space. The common type of spatial

window or patch is constructed as follows: for each surface residue, its distances by their alpha C atoms with all other surface residues in the same protein chain are calculated and sorted in ascending order, and then the L spatially nearest surface residues constitute a surface patch/window for including the environmental information. The size of the surface patch is a parameter to be optimized in the training stage. A topological patch or window is similarly defined by the n vertices with the smallest geodesic distances (shortest paths) to the center vertex. In this case, protein structures are recast as topological graphs based on protein residue contact maps, where each vertex of the graph represents the alpha C atom of an amino acid and edges connect vertices within a distance cutoff of 8 Å [17, 18].

4.2 Structure-Based Features of DNA-Binding Residues

For an effective classification model, the selected features should be highly related to the class of DNA-binding residues and have discrimination power to distinguish DNA-binding residues from nonbinding residues on surface of DNA-binding residues. A large number of studies have identified various sequence features [19–34], such as amino acid composition, physiochemical properties, and predicted structure features, and evolutionary features based on position-specific score matrix (PSSM) generated by PSI--BLAST [35]. However, sequence-based features and evolutionary features are not sufficient for prediction of DNA-binding residues, since the functions of proteins are more directly affected by their structural features. Thus, an increasing number of prediction methods have incorporated structural features, such as secondary structure, solvent-accessible surface area, spatial neighbors, *B*-factor, the empirical preference of electrostatic potential [36–39], and the shape of molecular surfaces [40–42].

4.2.1 Relative Solvent Accessibility (RSA)

Relative solvent accessibility of a residue was calculated as the ratio of its SASA to the nominal maximum area of its residue type in a tripeptide state. The results in previous studies show that positively charged residues Arg and Lys were more exposed in the binding group than in the nonbinding group, giving resultant more binding propensity, whereas for negatively charged residues Asp and Glu, it was opposite [43, 44].

4.2.2 B-Factor

B-factors are highly related to the flexibility of atoms and residues in a protein and are determined by X-ray crystallographic experiments. *B*-factor of alpha C atom was used to represent its residue flexibility and obtained from its PDB file. For each protein chain, the *B*-factor of each alpha C atom was normalized as follows:

$$\mathrm{NB} = \frac{B - \mu(B)}{\sigma(B)} \tag{1}$$

where B is the *B*-factor value of a given residue and $\mu(B)$ and $\sigma(B)$ are the average value and the standard deviation of the

B-factors for the selected chain, respectively. In our previous studies, the average values of the *B*-factors of 20 types of residues in DNA-binding group were also significantly lower than the non-binding group. This result can be explained by the fact that the atoms of DNA-binding residues are less exposed to solvent and experience less fluctuation, resulting in relatively lower *B*-factors. Generally, the *B*-factor is a distinguished feature to characterize the binding residues of biological macromolecules in their bound states. However, the *B*-factors of binding residues in apoproteins (without DNA present) were ranging from relatively lower values (rigidity) to higher values (flexibility), which suggests that the binding sites have dual character about mobility.

4.2.3 Betweenness Centrality

Protein structures are recast as topological graphs based on protein residue contact maps, where each vertex of the graph represents the alpha C atom of an amino acid and edges connect vertices within a distance cutoff of 8 Å. Once the graph is constructed, a variety of topological metrics can be used to describe functional residues.

Betweenness centrality (BC) measures how frequently a vertex occurs on the shortest path between all other vertex pairs within the contact graph (undirected graph) of a protein chain of length n. Since the chains vary in length, the measure is normalized by dividing through the number of pairs of vertices not including v, which is $(n-1)(n-2)/2$.

$$\mathrm{BC}(v) = \frac{2}{(n-1)(n-2)} \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{2}$$

where V is the set of vertices, σ_{st} is the number of shortest paths from s to t, and $\sigma_{st}(v)$ is the number of shortest paths from s to t that pass through vertex v.

The weighted average betweenness centrality for a surface patch was calculated from the betweenness centrality values of component residues, weighted by the weighting factor w_i of residue i:

$$\mathrm{WBC} = \sum_{i=1}^{L} w_i \cdot \mathrm{BC}(i) \tag{3}$$

Our previous study found that residues located at protein-DNA interfaces exhibit the central role in the protein network with high betweenness centrality [18]. The predictive power of betweenness centrality was low (PR-AUC 0.208) for individual residues but was high (PR-AUC 0.228) when averaged over a patch of neighboring residues. This may suggest that a set of residues with higher betweenness centralities form a community so as to play an important role in protein-DNA interaction. It was

indicated that DNA-binding residues are distinguishable from the nonbinding residues on protein surfaces by their higher weighted average betweenness centrality.

4.2.4 Electrostatic Potential

Electrostatic complementarity is shown to be important for protein-DNA interaction. DNA-binding sites have a large overlap with the surface patches which have the largest positive electrostatic potential [39, 45]. PBEQ-Solver can be used to calculate electrostatic potential of all atoms in a protein [36].

Protein surface was placed on a cubic grid. For each atom, the electrostatic potential of nearby grid points was averaged to construct an electrostatic feature at an atom-scale. The electrostatic potential values of grid points between the van der Waals and solvent-accessible surfaces were averaged, using a solvent probe at the radius of 1.4 Å. For each atom, values for the electrostatic potential at grid points were averaged at grid points outside the van der Waals surface of the given protein but within a distance that is the sum of the atom radius and the solvent radius. Three additional groups of features were derived by moving the shell slightly outward, by radius offsets 0.1 Å, 0.3 Å, and 0.5 Å. Note that the region of the shell maintains a width of 1.4 Å regardless of the size of the offset, but the regions move farther away from the van der Waals surface as the offset is varied. Mathematically, this is equivalent to adding the offset value to the radius of all the atoms and repeating the previous calculation described for the van der Waals and solvent-accessible surfaces. Figure 1 illustrates the details of this calculation of electrostatic feature at the atom-level.

Next, the local sums and averages of the residue-level electrostatic features were derived in the neighborhood of the target residue. More details can be found at [36].

4.3 Classification Algorithms

To our best knowledge, classification algorithms are mainly categorized into three classes: decision tree-based, artificial intelligence-based, and statistics-based methods. Decision tree algorithms provide an intuitive way for classifying a new sample based on a set of simple and easily interpretable rules. The artificial intelligence-based classification methods include the artificial neural network [46], deep learning algorithms, and evolutionary algorithms [47], which can be further classified into the genetic algorithm and swarm algorithm. The statistics-based methods include various algorithms such as support vector machine [48, 49], random forest [50], stochastic gradient boosting algorithm [51], and Bayesian classifier [52]. These diverse classification methods have already been explored to the prediction of DNA-binding residues on DNA-binding proteins. The comprehensive overview of the characteristics and specific application of these classification algorithms is out of scope of this review.

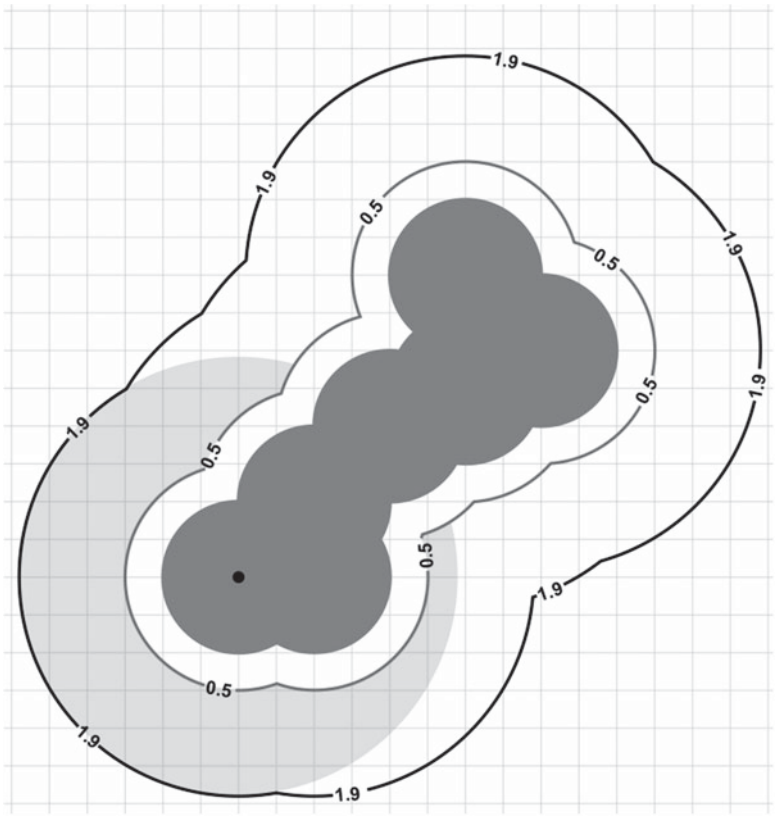

Fig. 1 The illustration of calculation of the atom-level electrostatic feature within a shell offset 0.5 Å from the van der Waals surface. The grid on which the electrostatic potential is calculated is shown relative to the molecule, shown in dark gray, and the atom at which the feature is calculated is marked using a black dot. Electrostatic potential values at grid points within the light gray annular region are those averaged to generate the feature for this atom. Grid points inside the 0.5 Å offset surface are excluded from the calculation. The light gray annular region is 1.4 Å in width regardless of the offset used to define the shell. The figure is extracted from [36]

4.4 Model Validation and Evaluation

When a prediction model is constructed by one type of classification algorithms, it requires a benchmark data set to validate and evaluate how the model works. The benchmark data set consists of well-labeled samples, which are divided into a training set and testing set. To evaluate the performance of classification models, the validation methods are mainly consisting of k-fold cross-validation, leave-one-out cross-validation, and independent tests. In k-fold cross-validation, the sample set is randomly partitioned into k subsets with equal size. Of the k subsets, one subset is selected as the validation data for testing the model, and the remaining $k - 1$ subsets are used as training data. The cross-validation process is then repeated k times (the folds), with each of the k subsets used exactly once as the validation data. The results from k folds are finally averaged to generate a single estimation metric. Leave-one-

Table 1
A list of common metrics and their equations

Metric	Equation
ACC	$(TP + TN)/(TP + TN + FP + FN)$
SN	$TP/(TP + FN)$
SP	$TN/(TN + FP)$
PR	$TP/(TP + FP)$
MCC	$(TP \times TN - FP \times FN)/\sqrt{(TP + FN) \times (TP + FP) \times (TN + FP) \times (TN + FN)}$
F_1	$2 \times SN \times PR/(SN + PR)$

out cross-validation (LOOCV) involves using a single instance (or sample) from the sample set as the validation data and the remaining instances as the training data. This is repeated such that each instance in the sample set is used once as the validation data. This is the same as a *k*-fold cross-validation with *k* being equal to the number of instances in the original sample set. Leave-one-out cross-validation is computationally expensive when the number of samples in the training set is too large. In the task of prediction of DNA-binding residues, the instance or sample can be a residue or protein chain in the leave-one-out cross-validation. In order to test the model in an unseen sample set, an independent test is usually adopted and conducted on a separate set which is independent of the training set. This type of test resembles a true prediction and reflects the generalization ability of a prediction model.

In order to assess the classification performance, various threshold-dependent metrics are utilized and defined as follows. They are accuracy (ACC), sensitivity (SN, also called recall), specificity (SP), precision (PR), Matthew's correlation coefficient (MCC), and F-measure (F_1). These metrics are calculated using the numbers of true positives (TP), false positives (FP), true negatives (TN) , and false negatives (FN) for each classifier. Their equations are defined in Table 1.

TP is the number of correctly predicted DNA-binding residues, TN is the number of correctly predicted nonbinding residues, FP is the number of nonbinding residues predicted as binding residues, and FN is the number of binding residues wrongly predicted as nonbinding.

The receiver operating characteristic (ROC) curve is a plot of the sensitivity versus (1-specificity) for a binary classifier at varying thresholds. The area under the curve (AUC) can be used as a threshold-independent measure of classification performance. It is a nontrivial task to assess the quality of prediction for heavily unbalanced data sets. On the unbalanced data sets, the accuracy and AUC of ROC curve can present overly optimistic assessments

Table 2
Web servers for prediction of DNA-binding residues using structure-based methods

Server	Website	Description/structural feature	Reference
DBS-Pred	http://www.netasa.org/dbs-pred	Solvent accessibility and secondary structure	[33]
PreDs	http://pre-s.protein.osaka-u.ac.jp/~preds	Electrostatic potential, the local curvature and the global curvature	[53]
DISPLAR	http://pipe.scs.fsu.edu/displar.html	Solvent accessibility	[44]
DBD-Hunter	http://cssb.biology.gatech.edu/skolnick/webservice/DBD-Hunter	Structural comparison and the evaluation of a statistical potential	[5]
DNABINDPROT	http://www.prc.boun.edu.tr/appserv/prc/dnabindprot	The fluctuations of residues in high-frequency modes	[54]
DNABind	http://mleg.cse.sc.edu/DNABind/	Combining machine learning- and template-based approaches	[14]
DBSI	http://dbsi.mitchell-lab.org	A banded electrostatic feature	[55]

of performance of an algorithm. Instead, the precision-recall (PR) curve is a plot of the recall versus precision for a binary classifier at varying thresholds.

4.5 Web Servers

To establish a useful structure-based predictor for a biological system, it is usual to develop a user-friendly web server for the predictor that is accessible to the public. Table 2 briefly presents a summary of publicly available methods that predict DNA-binding residues using structure-based features of DNA-binding proteins.

Acknowledgments

This work was supported by the grants from National Natural Science Foundation of China for Young Scholars (Grant No. 31601074 and 21403002), the funding from National Key Research Program (Contract No. 2016YFA0501703), and the Open Fund of Shanghai Key Laboratory of Intelligent Information Processing (Contract No. IIPL-2016-005).

References

1. Luscombe NM, Austin SE, Berman HM, Thornton JM (2000) An overview of the structures of protein-DNA complexes. Genome Biol 1(1):REVIEWS001
2. Biswas S, Guharoy M, Chakrabarti P (2009) Dissection, residue conservation, and structural classification of protein-DNA interfaces. Protein Struct Funct Bioinformatics 74 (3):643–654

3. Ahmad S, Keskin O, Sarai A, Nussinov R (2008) Protein-DNA interactions: structural, thermodynamic and clustering patterns of conserved residues in DNA-binding proteins. Nucleic Acids Res 36(18):5922–5932
4. Zhao H, Yang Y, Zhou Y (2010) Structure-based prediction of DNA-binding proteins by structural alignment and a volume-fraction corrected DFIRE-based energy function. Bioinformatics 26(15):1857–1863
5. Gao M, Skolnick J (2008) DBD-Hunter: a knowledge-based method for the prediction of DNA-protein interactions. Nucleic Acids Res 36(12):3978–3992
6. Jones S, Barker JA, Nobeli I, Thornton JM (2003) Using structural motif templates to identify proteins with DNA binding function. Nucleic Acids Res 31(11):2811–2823
7. Gao M, Skolnick J (2009) A threading-based method for the prediction of DNA-binding proteins with application to the human genome. PLoS Comput Biol 5(11):e1000567
8. Gherardini PF, Helmer-Citterich M (2008) Structure-based function prediction: approaches and applications. Brief Funct Genomic Proteomic 7(4):291–302
9. Nimrod G, Szilagyi A, Leslie C, Ben-Tal N (2009) Identification of DNA-binding proteins using structural, electrostatic and evolutionary features. J Mol Biol 387(4):1040–1053
10. Ahmad S, Sarai A (2004) Moment-based prediction of DNA-binding proteins. J Mol Biol 341(1):65–71
11. Liu B, Wang S, Wang X (2015) DNA binding protein identification by combining pseudo amino acid composition and profile-based protein representation. Sci Rep 5:15479
12. Miao Z, Westhof E (2015) A large-scale assessment of nucleic acids binding site prediction programs. PLoS Comput Biol 11(12): e1004639
13. Gromiha MM, Fukui K (2011) Scoring function based approach for locating binding sites and understanding recognition mechanism of protein-DNA complexes. J Chem Inf Model 51(3):721–729
14. Liu R, Hu J (2013) DNABind: a hybrid algorithm for structure-based prediction of DNA-binding residues by combining machine learning- and template-based approaches. Proteins 81(11):1885–1899
15. Zen A, de Chiara C, Pastore A, Micheletti C (2009) Using dynamics-based comparisons to predict nucleic acid binding sites in proteins: an application to OB-fold domains. Bioinformatics 25(15):1876–1883
16. Gao M, Skolnick J (2009) From nonspecific DNA-protein encounter complexes to the prediction of DNA-protein interactions. PLoS Comput Biol 5(3):e1000341
17. Maetschke SR, Yuan Z (2009) Exploiting structural and topological information to improve prediction of RNA-protein binding sites. BMC Bioinformatics 10:341
18. Xiong Y, Xia J, Zhang W, Liu J (2011) Exploiting a reduced set of weighted average features to improve prediction of DNA-binding residues from 3D structures. PLoS One 6(12): e28440
19. Zhou J, Xu R, He Y, Lu Q, Wang H, Kong B (2016) PDNAsite: identification of DNA-binding site from protein sequence by incorporating spatial and sequence context. Sci Rep 6:27653
20. Yan J, Friedrich S, Kurgan L (2016) A comprehensive comparative review of sequence-based predictors of DNA- and RNA-binding residues. Brief Bioinform 17(1):88–105
21. Peng Z, Kurgan L (2015) High-throughput prediction of RNA, DNA and protein binding regions mediated by intrinsic disorder. Nucleic Acids Res 43(18):e121
22. Si J, Zhang Z, Lin B, Schroeder M, Huang B (2011) MetaDBSite: a meta approach to improve protein DNA-binding sites prediction. BMC Syst Biol 5(Suppl 1):S7
23. Wang L, Huang C, Yang MQ, Yang JY (2010) BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features. BMC Syst Biol 4(Suppl 1):S3
24. Cai Y, He Z, Shi X, Kong X, Gu L, Xie L (2010) A novel sequence-based method of predicting protein DNA-binding residues, using a machine learning approach. Mol Cells 30 (2):99–105
25. JS W, Liu HD, Duan XY, Ding Y, HT W, Bai YF, Sun X (2009) Prediction of DNA-binding residues in proteins from amino acid sequences using a random forest model with a hybrid feature. Bioinformatics 25(1):30–35
26. Wang L, Yang MQ, Yang JY (2009) Prediction of DNA-binding residues from protein sequence information using random forests. BMC Genomics 10(Suppl 1):S1
27. Badis G, Berger MF, Philippakis AA, Talukder S, Gehrke AR, Jaeger SA, Chan ET, Metzler G, Vedenko A, Chen X, Kuznetsov H, Wang CF, Coburn D, Newburger DE, Morris Q, Hughes TR, Bulyk ML (2009) Diversity and complexity in DNA recognition by transcription factors. Science 324 (5935):1720–1723

28. Ofran Y, Mysore V, Rost B (2007) Prediction of DNA-binding residues from sequence. Bioinformatics 23(13):I347–I353
29. Hwang S, Gou ZK, Kuznetsov IB (2007) DP-Bind: a Web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins. Bioinformatics 23 (5):634–636
30. Ho SY, FC Y, Chang CY, Huang HL (2007) Design of accurate predictors for DNA-binding sites in proteins using hybrid SVM-PSSM method. Biosystems 90 (1):234–241
31. Wang LJ, Brown SJ (2006) BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. Nucleic Acids Res 34:W243–W248
32. Wang L, Brown SJ (2006) Prediction of DNA-binding residues from sequence features. J Bioinform Comput Biol 4(6):1141–1158
33. Ahmad S, Gromiha MM, Sarai A (2004) Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. Bioinformatics 20(4):477–486
34. Yan J, Kurgan L (2017) DRNApred, fast sequence-based method that accurately predicts and discriminates DNA- and RNA-binding residues. Nucleic Acids Res 45 (10):e84
35. Ahmad S, Sarai A (2005) PSSM-based prediction of DNA binding sites in proteins. BMC Bioinformatics 6:33
36. Zhu X, Ericksen SS, Mitchell JC (2013) DBSI: DNA-binding site identifier. Nucleic Acids Res 41(16):e160
37. Tsuchiya Y, Kinoshita K, Nakamura H (2004) Structure-based prediction of DNA-binding sites on proteins using the empirical preference of electrostatic potential and the shape of molecular surfaces. Protein Struct Funct Bioinformatics 55(4):885–894
38. Chen YC, CY W, Lim C (2007) Predicting DNA-binding amino acid residues from electrostatic stabilization upon mutation to Asp/Glu and evolutionary conservation. Proteins 67(3):671–680
39. Bhardwaj N, Lu H (2007) Residue-level prediction of DNA-binding sites and its application on DNA-binding protein predictions. FEBS Lett 581(5):1058–1066
40. Zhou W, Yan H (2010) A discriminatory function for prediction of protein-DNA interactions based on alpha shape modeling. Bioinformatics 26(20):2541–2548
41. Zhou P, Tian F, Ren Y, Shang Z (2010) Systematic classification and analysis of themes in protein-DNA recognition. J Chem Inf Model 50(8):1476–1488
42. Sonavane S, Chakrabarti P (2009) Cavities in protein-DNA and protein-RNA interfaces. Nucleic Acids Res 37(14):4613–4620
43. Xiong Y, Liu J, Wei DQ (2011) An accurate feature-based method for identifying DNA-binding residues on protein surfaces. Proteins 79(2):509–517
44. Tjong H, Zhou HX (2007) DISPLAR: an accurate method for predicting DNA-binding sites on protein surfaces. Nucleic Acids Res 35 (5):1465–1477
45. Jones S, Shanahan HP, Berman HM, Thornton JM (2003) Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins. Nucleic Acids Res 31 (24):7189–7198
46. Dai H, Xu Q, Xiong Y, Liu WL, Wei DQ (2016) Improved prediction of michaelis constants in CYP450-mediated reactions by resilient back propagation algorithm. Curr Drug Metab 17(7):673–680
47. Yao Y, Zhang T, Xiong Y, Li L, Huo J, Wei DQ (2011) Mutation probability of cytochrome P450 based on a genetic algorithm and support vector machine. Biotechnol J 6 (11):1367–1376
48. Xiong Y, Liu J, Zhang W, Zeng T (2012) Prediction of heme binding residues from protein sequences with integrative sequence profiles. Proteome Sci 10(Suppl 1):S20
49. Li L, Xiong Y, Zhang ZY, Guo Q, Xu Q, Liow HH, Zhang YH, Wei DQ (2015) Improved feature-based prediction of SNPs in human cytochrome P450 enzymes. Interdiscip Sci Comput Life Sci 7(1):65–77
50. Zhang W, Xiong Y, Zhao M, Zou H, Ye X, Liu J (2011) Prediction of conformational B-cell epitopes from 3D structures by random forests with a distance-based feature. BMC Bioinformatics 12:341
51. Xu Q, Xiong Y, Dai H, Kumari KM, Xu Q, HY O, Wei DQ (2017) PDC-SGB: prediction of effective drug combinations using a stochastic gradient boosting algorithm. J Theor Biol 417:1–7
52. Sun Y, Xiong Y, Xu Q, Wei D (2014) A hadoop-based method to predict potential effective drug combination. Biomed Res Int 2014:196858
53. Tsuchiya Y, Kinoshita K, Nakamura H (2005) PreDs: a server for predicting dsDNA-binding

site on protein molecular surfaces. Bioinformatics 21(8):1721–1723

54. Ozbek P, Soner S, Erman B, Haliloglu T (2010) DNABINDPROT: fluctuation-based predictor of DNA-binding residues within a network of interacting residues. Nucleic Acids Res 38(Web Server issue):W417–W423
55. Sukumar S, Zhu X, Ericksen SS, Mitchell JC (2016) DBSI server: DNA binding site identifier. Bioinformatics 32(18):2853–2855