Design of accurate predictors for DNA-binding sites in proteins using hybrid SVM–PSSM method

Shinn-Ying Ho a,b,∗, Fu-Chieh Yu b, Chia-Yun Chang b, Hui-Ling Huang c

a Department of Biological Science and Technology, National Chiao Tung University, Hsinchu 300, Taiwan
b Institute of Bioinformatics, National Chiao Tung University, Hsinchu 300, Taiwan
c Department of Information Management, Jin Wen Institute of Technology, Hsin-Tien, Taipei, Taiwan

Received 1 May 2006; received in revised form 29 July 2006; accepted 18 August 2006

Abstract

In this paper, we investigate the design of accurate predictors for DNA-binding sites in proteins from amino acid sequences. As a result, we propose a hybrid method using support vector machine (SVM) in conjunction with evolutionary information of amino acid sequences in terms of their position-specific scoring matrices (PSSMs) for prediction of DNA-binding sites. Considering the numbers of binding and non-binding residues in proteins are significantly unequal, two additional weights as well as SVM parameters are analyzed and adopted to maximize net prediction (NP, an average of sensitivity and specificity) accuracy. To evaluate the generalization ability of the proposed method SVM–PSSM, a DNA-binding dataset PDC-59 consisting of 59 protein chains with low sequence identity on each other is additionally established. The SVM-based method using the same six-fold cross-validation procedure and PSSM features has NP = 80.15% for the training dataset PDNA-62 and NP = 69.54% for the test dataset PDC-59, which are much better than the existing neural network-based method by increasing the NP values for training and test accuracies up to 13.45% and 16.53%, respectively. Simulation results reveal that SVM–PSSM performs well in predicting DNA-binding sites of novel proteins from amino acid sequences.

© 2006 Elsevier Ireland Ltd. All rights reserved.

Keywords: Amino acid sequence; DNA-binding prediction; Position-specific scoring matrices (PSSM); Protein; Support vector machine (SVM)

1. Introduction

The regulation of gene expression plays an important role within an organism. It is mainly controlled via binding of transcription factors to DNA for promoting or repressing gene expression levels. These transcription factors are mainly DNA-binding proteins coded by 2–3% of the genome in prokaryotes and 6–7% in eukaryotes (Frishman and Mewes, 1997; Luscombe et al., 2000; Lejeune et al., 2005). The malfunction of genetic activities may affect normal physiological functions or lead to disease in organisms. Thus we cannot neglect their decisive role in maintaining cells’ normal metabolism.

A variety of atomic contacts involved electrostatic, hydrogen bonds, hydrophobic and other van der Waals interactions between nucleic acids and amino acids have been studied for years (Luscombe et al., 2000; Lejeune et al., 2005; Nadassy et al., 1999; Luscombe and Thornton, 2002; Stawiski et al., 2003; Cheng et al., 2003). This research reveals that the DNA–protein recognition mechanism is complicated and there is no simple rule for this recognition problem (Pabo and Nekludova, 2000; O’Flanagan et al., 2005; Sarai and Kono, 2005). Previous research focused mainly on prediction and analysis of protein binding sites in DNA (Wingender et al., 2000; Kel et al., 2003; Pudimat et al.,...
of novel proteins. It is better to consider the following characteristics in designing classifiers: (1) the numbers of binding and non-binding residues in proteins are significantly unequal such that the unbalanced distribution should be considered in enhancing the NP accuracy; (2) the size of the given training dataset is relatively small compared to the number of used features such that concerns should be given to overfitting; (3) it is essential to design proper datasets for evaluating the generalization ability of the designed classifier in predicting potentially novel DNA-binding proteins.

Support vector machines (SVMs) were commonly used to analyze biological problems with satisfying results, such as classification of cancers in microarray (Paul and Iba, 2006), protein relative solvent accessibility prediction (Nguyen and Rajapakse, 2005), protein secondary structure prediction (Guo et al., 2004), protein transmembrane region prediction (Natt et al., 2004) and protein disulfide connectivity prediction (Chen and Hwang, 2005). SVM is a machine learning method with complete statistical learning theory basis (Vapnik, 1995). Furthermore, SVM has several advantages, such as (1) SVM can employ kernel functions that operate in extremely high-dimensional feature spaces, and the different class of samples are separated by the set of support vectors; (2) SVM can avoid falling into the local optimum solution in training phase (Burges, 1998); (3) SVM has a strong generalization ability when the size of given training dataset is relatively small, compared with the number of used features.

This paper proposes a hybrid method using SVM in conjunction with PSSMs for predicting DNA-binding sites in proteins from amino acid sequences. To advance the proposed method SVM–PSSM, the control parameters of SVM and two weight parameters for the unbalanced distribution of samples are analyzed and adopted to maximize NP accuracy. Furthermore, to enhance the accuracy of predicting novel proteins, an additional DNA-binding dataset PDC-59 consisting of 59 protein chains with low sequence identity on each other is established for evaluating generalization abilities of predictors. The SVM-based method using the same 6-CV procedure and PSSM features has accuracy NP = 80.15% for the training dataset PDNA-62 and NP = 69.54% for the independent test on the dataset PDC-59, which are much better than the NN-based method (Ahmad and Sarai, 2005) by increasing the NP values for training and test accuracies up to 13.45% and 16.53%, respectively. Simulation results reveal that SVM–PSSM performs well in predicting DNA-binding sites of novel proteins from amino acid sequences.

2. Methods

2.1. Datasets

We use three datasets (PDNA-62, PDNA-48, PDC-59) to evaluate our SVM–PSSM method which aims to have accurate prediction ability when giving a novel protein with low sequence identity compared with existing samples. Therefore, a filtering tool PISCES with much more rigorous sequence identity (Wang and Dunbrack, 2003) is used to filter out highly homologous sequences. Sequence identities for PDB (Protein Data Bank) sequences in PISCES are
determined by the combination of CE structural alignment and PSI-BLAST alignment, which is more sophisticated than the traditional local and global alignment method. The sequence identity in PDNA-48 and PDC-59 is confirmed by PISCES.

The missed hydrogen of the obtained PDB structures is added by MolProbity (Davis et al., 2004), and it optimizes all hydrogen atoms, both polar and non-polar, on amino acids and nucleic acids. We define the amino acid as a binding residue if its side chain or backbone atoms fell within a cut-off distance 3.5 Å, which is the same as a previous study (Ahmad et al., 2004; Ahmad and Sarai, 2005) from any atom in DNA sequences. Otherwise, the sample is a non-binding residue. Our calculation result of DNA-protein binding positions is highly consistent with that of the PDBsum database.

2.1.1. PDNA-62

For comparisons, the same dataset PDNA-62 containing 62 proteins in previous studies (Ahmad et al., 2004; Ahmad and Sarai, 2005) is used to predict DNA-binding sites in proteins. This dataset consisting of 7967 non-binding and 1792 binding residues has representative protein–DNA complexes from PDB and the protein structure resolution is 2.5 Å or better.

2.1.2. PDNA-48

The low sequence identity of each protein chain within a dataset would assist the samples in the uniform distribution within the sample space and thus can help the design of classifiers with strong generalization ability. Therefore, PDNA-62 was further filtered by PISCES using an identity threshold 25%. The obtained dataset PDNA-48 contains 48 protein chains (total 6431 residues; 1030 binding residues), listed in Table 1.

2.1.3. PDC-59

For further evaluating performance of SVM–PSSM in predicting novel proteins, we established a dataset PDC-59 for independent tests in this study. To obtain a discriminating dataset from PDNA-62, these proteins of PDC-59 are extracted from the PDB database with released dates after year 2000 by utilizing keywords: transcription factor, repressor, regulator, transposase, endonuclease and DNA-binding. These proteins were also filtered with mutual sequence identity less than 25% compared to each other and to PDNA-48 by PISCES. PDC-59 contains 59 protein chains (total 13041 residues; 1454 binding residues), listed in Table 2.

Note that the numbers of binding and non-binding residues in proteins are significantly unequal that the unbalanced distribution should be taken into account in designing accurate predictors.

2.2. SVM–PSSM

2.2.1. PSSM and feature vector representation

We use multiple sequence alignment profiles generated from PSI-BLAST (Altschul et al., 1997) for each protein chain. We obtain the non-redundant protein sequence database from NCBI (National Center for Biotechnology Information). We set parameters of PSI-BLAST using BLOSUM62 substitution matrix, three iteration runs and exception value 0.001. The other parameters are set using default values. The PSI-BLAST program by querying each protein against the NR (non-redundant) database is used to generate PSSM profiles which are in the form of 20N matrix, where N is the length of queried protein sequence. Let the residue \(i\) be represented by \(a_i = (a_{i,1}, \ldots, a_{i,20})\) where \(1 \leq i \leq N\). Each query residue is represented by a vector of 20 attributes. These profiles are normalized into the range \([0, 1]\) for speeding up the SVM training phase.

In the previous study, PSSMs were generated from reference databases with different sizes. Although it was observed that computational time can be saved by replacing the reference database with a much smaller size without loss of much prediction ability (about 2% of NP), we still take the NR database from NCBI as our reference database to make sure that PSI-BLAST can have better multiple sequence alignment results and generate representative PSSMs.

The input pattern to SVM using the PSSM features for the residue \(i\) is \(x_i = (a_{i-k}, \ldots, a_i, \ldots, a_{i+k})\) where \(k\) is the number of neighborhood residues on either side. We construct a matrix with window size \(s = 2k + 1\) centered on the target residue \(i\). The used profile \(x_i\) is the form of a \(20 \times s\) matrix.

2.2.2. SVM

SVM is a very popular and powerful method to deal with classification, prediction and regression problems (Cortes and Vapnik, 1995). The original idea of SVM is to use a linear separating hyperplane which maximizes the distance between two classes to create a classifier. It relies on preprocessing the
data to represent patterns in a high-dimensional space with an appropriate mapping function \( \phi \). For the binary SVM, the training data consist of \( N \) pairs \((x_i, y_i)\), \( i = 1, \ldots, N \) where instance vectors \( x_i \in \mathbb{R}^m \) and class labels \( y_i \in \{0,1\} \). The main task in the training phase is to replace \( w \) with a constant. SVM allows sample classification, including various variants of SVM. The used software tool LibSVM (Library for SVM) for support vector

In this work, we used \( K \) and \( \gamma \) are determined by maximizing the ratio of mean to variance of sensitivity, specificity, NP and accuracy. Numerous candidate values of the pair \((w_0, w_1)\) are obtained in \( [0, 0.5, 1.0] \) and \( [1.0, 50] \). Performances of the best SVM classifiers with \( C, \gamma, w_0 \) and \( w_1 \) for some specified values of window size \( s \) using 6-CV on PDNA-62 are listed in Table 3. Finally, we choose the classifier with parameters \( s = 7 \), \( C = 0.73 \), \( \gamma = 0.27 \), \( w_0 = 1.0 \) and \( w_1 = 7.0 \) which has the best performance in terms

![Image](image-url)

**Fig. 1.** The distribution plot of the NP accuracy for PDNA-62 with window size 7 and various values of \( C \) and \( \gamma \), where gray bar represents the value of NP in percentage.
Table 3
Performances of the best SVM classifiers with \( C \), \( \gamma \), \( w_0 \) and \( w_1 \) for some specified values of window size \( s \) using 6-CV on PDNA-62

<table>
<thead>
<tr>
<th>( s )</th>
<th>( C )</th>
<th>( \gamma )</th>
<th>( w_0 )</th>
<th>( w_1 )</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
<th>NP (%)</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.70</td>
<td>4.44</td>
<td>1.0</td>
<td>6.0</td>
<td>73.55</td>
<td>73.73</td>
<td>73.64</td>
<td>73.70</td>
</tr>
<tr>
<td>3</td>
<td>0.50</td>
<td>1.52</td>
<td>1.0</td>
<td>9.0</td>
<td>78.35</td>
<td>78.44</td>
<td>78.39</td>
<td>78.42</td>
</tr>
<tr>
<td>5</td>
<td>0.74</td>
<td>0.60</td>
<td>0.5</td>
<td>3.0</td>
<td>79.30</td>
<td>79.33</td>
<td>79.31</td>
<td>79.32</td>
</tr>
<tr>
<td>7</td>
<td>0.73</td>
<td>0.27</td>
<td>1.0</td>
<td>7.0</td>
<td>80.08</td>
<td>80.23</td>
<td>80.15</td>
<td>80.20</td>
</tr>
<tr>
<td>9</td>
<td>0.60</td>
<td>0.30</td>
<td>0.5</td>
<td>3.2</td>
<td>80.08</td>
<td>80.11</td>
<td>80.09</td>
<td>80.10</td>
</tr>
<tr>
<td>13</td>
<td>1.30</td>
<td>0.10</td>
<td>0.5</td>
<td>3.1</td>
<td>80.02</td>
<td>79.97</td>
<td>79.99</td>
<td>79.98</td>
</tr>
</tbody>
</table>

Table 4
Performances of the SVM classifier with \( s = 7 \), \( C = 0.73 \) and \( \gamma = 0.27 \) for some values of \( w_0 \) and \( w_1 \) on the dataset PDNA-62

<table>
<thead>
<tr>
<th>( w_0 )</th>
<th>( w_1 )</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
<th>NP (%)</th>
<th>Accuracy (%)</th>
<th>Mean (%)</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>2.0</td>
<td>55.86</td>
<td>93.79</td>
<td>74.82</td>
<td>86.82</td>
<td>77.82</td>
<td>827.27</td>
</tr>
<tr>
<td>1.0</td>
<td>5.0</td>
<td>76.95</td>
<td>83.24</td>
<td>80.10</td>
<td>82.09</td>
<td>80.60</td>
<td>22.76</td>
</tr>
<tr>
<td>1.0</td>
<td>6.0</td>
<td>78.79</td>
<td>81.51</td>
<td>80.15</td>
<td>81.01</td>
<td>80.37</td>
<td>4.25</td>
</tr>
<tr>
<td>1.0</td>
<td>6.5</td>
<td>79.41</td>
<td>80.81</td>
<td>80.11</td>
<td>80.55</td>
<td>80.22</td>
<td>1.13</td>
</tr>
<tr>
<td>1.0</td>
<td>6.7</td>
<td>79.52</td>
<td>80.62</td>
<td>80.07</td>
<td>80.42</td>
<td>80.16</td>
<td>0.70</td>
</tr>
<tr>
<td>1.0</td>
<td>7.0</td>
<td>80.08</td>
<td>80.23</td>
<td>80.15</td>
<td>80.20</td>
<td>80.17</td>
<td>0.01</td>
</tr>
<tr>
<td>1.0</td>
<td>8.0</td>
<td>81.31</td>
<td>79.20</td>
<td>80.25</td>
<td>79.59</td>
<td>80.09</td>
<td>2.54</td>
</tr>
<tr>
<td>1.0</td>
<td>50</td>
<td>83.98</td>
<td>74.97</td>
<td>79.48</td>
<td>76.63</td>
<td>78.76</td>
<td>46.69</td>
</tr>
</tbody>
</table>

3. Results

3.1. Performance comparison of training datasets

To evaluate performance of the proposed method SVM–PSSM, the existing NN-based method is conveniently compared using the same 6-CV on PDNA-62. The comparison results are given in Table 5. The NP values of the NN-based methods using sequence information only (Ahmad et al., 2004) and the PSSM feature with window size \( s = 3 \) (Ahmad and Sarai, 2005) are 58.4% and 66.7%, respectively. The SVM–PSSM classifiers with \( s = 3 \) and 7 have NP = 78.39% and 80.15%, respectively. The SVM–PSSM classifier is much better than the NN-PSSM classifier by increasing the value of NP up to 13.45% for the training dataset PDNA-62.

Besides the NP performance, the receiver operating characteristic (ROC) curve is commonly used to evaluate the discrimination ability of a classifier. The larger the area under the ROC curve, the better discrimination ability a classifier has. Fig. 2 gives the performance comparison using the ROC curves on PDNA-62. The ROC curve of the SVM classifier is obtained from Table 4. The ROC curve of the NN-based classifier is obtained from the DBS-PSSM website as mentioned in (Ahmad and Sarai, 2005). It shows that the area under the ROC curve of SVM is much larger than that of the NN-based method obviously. It also shows that the SVM-based method has better classification ability than the NN-based method in classifying binding and non-binding residues in proteins.

3.2. Performance comparison of independent test

In order to evaluate the generalization abilities of SVM–PSSM and the NN-based method (Ahmad and
Table 6
Independent test results of the NN-based method and SVM–PSSM (using either PDNA-62 or PDNA-48 as the training dataset) on PDC-59

<table>
<thead>
<tr>
<th>Method</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
<th>NP (%)</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM–PSSM (PDNA-48)</td>
<td>65.41</td>
<td>75.48</td>
<td>70.44</td>
<td>74.36</td>
</tr>
<tr>
<td>SVM–PSSM (PDNA-62)</td>
<td>59.35</td>
<td>79.72</td>
<td>69.54</td>
<td>77.45</td>
</tr>
<tr>
<td>DBS–PSSM (Ahmad and Sarai, 2005)</td>
<td>46.36</td>
<td>59.65</td>
<td>53.01</td>
<td>58.19</td>
</tr>
</tbody>
</table>

Sarai, 2005) in predicting novel proteins, PDC-59 is used for independent tests. The SVM classifier is obtained from the best one of 6-CV which has the best NP performance on the training dataset PDNA-62 mentioned above. The results of the NN-based method are obtained through the DBS-PSSM website (Ahmad and Sarai, 2005). Table 6 gives independent test results of the two compared methods. The results of the SVM and NN-based methods are NP = 69.54% and 53.01%, respectively. The SVM classifier is much better than the NN-based method by increasing the NP values for test accuracy up to 16.53%. It also reveals that the SVM classifier has better generalization ability to predict novel proteins.

To further improve the generalization ability of the SVM classifier, we filter out the proteins with identity greater than 25% in the dataset PDNA-62 by PISCES. The obtained dataset PDNA-48 is used as the training dataset. Consequently, we construct an SVM classifier using the same procedure as that on PDNA-62. The parameters of the obtained classifier from the best one of six classifiers using 6-CV are \( s = 7 \), \( C = 0.58 \), \( \gamma = 0.23 \), \( w_0 = 1.0 \) and \( w_1 = 7.2 \). Table 6 shows that the sensitivity performance is improved from 59.35% to 65.41% and NP performance is improved slightly from 69.54% to 70.44%. Therefore, when we used low identity proteins as training data, it is helpful to obtain a classifier with high generalization ability for correctly predicting binding residues of novel proteins.

The process of DNA–protein recognition is flexible and continuous (Gunther et al., 2006; Sarai and Kono, 2005), and the crystals of protein–DNA complex just catch a moment of this whole process. Therefore, the amino acid defined as a non-binding residue with a distance slightly larger than the cut-off distance 3.5 Å may assist or take part in protein–DNA recognition. We analyze the distance distribution of non-binding residues in PDC-59 which are classified incorrectly as binding ones by the SVM classifier using the training dataset PDNA-48. The result given in Fig. 3 reveals that there are 39% of non-binding residues with the distance in the range 3.5–8.5 Å close to the nearest DNA nucleotide. The percentage of misclassified non-binding residues decreases gradually when their distance increases. The logical result reveals that SVM–PSSM is a good predictor for biologists to analyze the protein-DNA binding mechanism.

Fig. 2. The performance comparison between SVM–PSSM and the NN-based method using the ROC curve on PDNA-62.

Fig. 3. Distribution of misclassified non-binding residues. X-axis represents the distance between the residues to the nearest atom on DNA. Y-axis represents the percentage of misclassified non-binding residues to total non-binding residues in the specified distance range.
3.3. Analysis and discussion

The class of a query residue is determined by the discrimination function of SVM. When the function value of a residue is greater than zero, it would be classified into the non-binding class. Otherwise, it would be classified into the binding class. Fig. 4 shows the relationship of discrimination function values and distances between the residues to the nearest atom on DNA using all misclassified non-binding residues. It reveals that these misclassified non-binding residues which are closer to DNA would get smaller values of the SVM discrimination function. This scenario indicates that the SVM classifier has good screening abilities to select potential binding residues. These amino acids in the vague region (with distances near the cut-off distance 3.5 Å to DNA) may potentially take part in or assist protein–DNA recognition and can be further verified by biologists.

Generally, the cut-off value of SVM discrimination function is set to zero for normal classification. Fig. 5 shows the distribution of binding and non-binding residues in PDC-59 using the SVM classifier with a cut-off value equal to zero. Once the SVM classifier with the best setting of parameters \((s, C, \gamma, w_0, w_1)\) is developed, we may adjust the cut-off value in using the SVM classifier according to the preference such as higher NP, higher sensitivity, higher specificity, etc. Table 7 gives the result of the SVM classifier \((s = 7, C = 0.58, \gamma = 0.23, w_0 = 1.0 \text{ and } w_1 = 7.2)\) on the dataset PDC-59 for various cut-off values. For example, if the cut-off value is set to 0.29, the highest value of NP equals to 71.15%. If the higher Sensitivity performance is desirable, the cut-off value of the SVM classifier can be properly increased.

### Table 7

<table>
<thead>
<tr>
<th>Cut-off value</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
<th>NP (%)</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>−3.00</td>
<td>0.00</td>
<td>100.00</td>
<td>50.00</td>
<td>88.84</td>
</tr>
<tr>
<td>−2.00</td>
<td>1.93</td>
<td>99.84</td>
<td>50.88</td>
<td>88.91</td>
</tr>
<tr>
<td>−1.00</td>
<td>24.28</td>
<td>96.05</td>
<td>60.16</td>
<td>88.04</td>
</tr>
<tr>
<td>−0.50</td>
<td>44.02</td>
<td>88.17</td>
<td>66.09</td>
<td>83.24</td>
</tr>
<tr>
<td>−0.30</td>
<td>53.09</td>
<td>83.85</td>
<td>68.47</td>
<td>80.42</td>
</tr>
<tr>
<td>−0.10</td>
<td>61.97</td>
<td>78.69</td>
<td>70.33</td>
<td>76.82</td>
</tr>
<tr>
<td>0.00</td>
<td>65.41</td>
<td>75.48</td>
<td>70.44</td>
<td>74.36</td>
</tr>
<tr>
<td>0.10</td>
<td>69.33</td>
<td>72.00</td>
<td>70.66</td>
<td>71.70</td>
</tr>
<tr>
<td>0.13</td>
<td>70.70</td>
<td>70.92</td>
<td>70.81</td>
<td>70.89</td>
</tr>
<tr>
<td>0.29</td>
<td>77.10</td>
<td>65.19</td>
<td>71.15</td>
<td>66.52</td>
</tr>
<tr>
<td>0.30</td>
<td>77.44</td>
<td>64.75</td>
<td>71.10</td>
<td>66.16</td>
</tr>
<tr>
<td>0.50</td>
<td>83.77</td>
<td>56.80</td>
<td>70.28</td>
<td>59.80</td>
</tr>
<tr>
<td>1.00</td>
<td>94.09</td>
<td>35.20</td>
<td>64.64</td>
<td>41.77</td>
</tr>
<tr>
<td>2.00</td>
<td>99.31</td>
<td>6.32</td>
<td>52.81</td>
<td>16.68</td>
</tr>
<tr>
<td>3.00</td>
<td>100.00</td>
<td>0.20</td>
<td>50.10</td>
<td>11.32</td>
</tr>
</tbody>
</table>

4. Conclusions

In this paper, we have proposed a hybrid method using SVM in conjunction with the PSSM features for prediction of DNA-binding sites in proteins from amino acid sequences by achieving high accuracy for novel proteins. Using the same PSSM features, simulation results show that our method SVM–PSSM is much better than the existing neural network-based method in terms of net prediction (NP) accuracy by increasing the NP values for training and test accuracies up to 13.45% and 16.53%, respectively. To best of our knowledge, up to now, the proposed method is the most effective method for recognizing mechanism of binding residues in proteins based...
on protein sequence without using 3D structural information, such as hydrogen bond, hydrophobic, hydrophilic, ion interaction, etc. By adjusting the cut-off value of the SVM classifier, the proposed prediction method would be helpful to biologists for filtering novel proteins without significant homology with known protein to find out the potential binding regions in proteins.

References


Ahmad, S., Sarai, A., 2005. PSSM-based prediction of DNA-binding sites in proteins. BMC Bioinform. 6, 33.


