An approach to predict transcription factor DNA binding site specificity based upon gene and transcription factor functional categorization

Ziliang Qian2,3,†, Lingyi Lu2,3,†, XiaoJun Liu7, Yu-Dong Cai1,4,* and Yixue Li3,5,6
1CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 5Graduate School of the Chinese Academy of Sciences, 19 Yuquan Road, Beijing 100039, 3Bioinformatics Center, Key Lab of Molecular Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 320 Yueyang Road, Shanghai 200031, China, 2Department of Mathematics, University of Manchester, Institute of Science and Technology, P.O. Box 88, Manchester M60 1QD, UK, 3Shanghai Center for Bioinformation Technology, 100 Qinzhou Road, 200235 Shanghai, 6College of Life Science & Biotechnology, Shanghai Jiao Tong University, China and 7Molecular Physiology Laboratory, Centre for Cardiovascular Science Queen’s Medical Research Institute, 47 Little France Crescent, Edinburgh, EH16 4TJ, UK

Received on December 24, 2007; revised on June 4, 2007; accepted on June 27, 2007

ABSTRACT

Motivation: To understand transcription regulatory mechanisms, it is indispensable to investigate transcription factor (TF) DNA binding preferences. We noted that the generally acknowledged information of functional annotations of TFs as well as that of their target genes should provide useful hints in determining TF DNA binding preferences.

Results: In this contribution, we developed an integrative method based on the Nearest Neighbor Algorithm, to predict DNA binding preferences through integrating both the functional/structural information of TFs and the interaction between TFs and their targets. The accuracy of cross-validation tests on the dataset consisting of 3430 positive samples and 7000 negative samples reaches 87.0% for 10-fold cross-validation and 87.9% for jackknife cross-validation test, which is a much better result than that in our previous work. The prediction result indicates that the improved method we developed could be a powerful approach to infer the TF DNA preference in silico.

Contact: cyd@picb.ac.cn

Supplementary information: Supplementary data are available at Bioinformatics online

1 INTRODUCTION

Transcription factor (TF), designated as the major regulator of transcription, prefers to bind specific regions on DNA sequences to modulate the nearby genes’ expression. Given a newly identified TF, how do we know it has partiality for which kinds of DNA sequences? Typically in cases like this, we resort to a vast number of experiments such as DNA footprint technology (Fox, 1997) to try to find direct evidence of the interaction between the TF and its target genes or build statistical models (Stormo, 2000) to describe the TF’s DNA binding preferences. However, both methods are not an ideal approach when tackling the challenging problem due to its remarkable consumption of time and money. Nevertheless, with the advent of massive integrative biological databases and various data processing tools, like the InterPro (Apweiler et al., 2001), a database of protein families, domains and functional sites and the gene regulation database TRANSFAC (Matys et al., 2001; Wingender et al., 1996), we are able to provide an alternative view of methods to help us to investigate the problem. We have reached an age where we possess not only the experimental techniques but also the existing knowledge that is expected to be put to good use to give us at least some indications and clues concerning the TF of interest. Thus, a novel strategy based on integrating both: the existing transcription regulatory knowledge and the generally acknowledged information of functional annotations of TFs as well as that of their target genes will soon be within reach.

Actually, in our previous work (Qian et al., 2006), we tried to answer the question whether the querying pair of TF and potential transcription factor binding sites (TFBS) interact or not, so as to infer the TF’s DNA binding preferences. It was based on InterPro (Apweiler et al., 2001) annotations of TFs, which covers a large number of functional domains/sites found in known proteins, and TRANSFAC (Matys et al., 2006; Wingender et al., 1996) which provides well-understood interacting pairs of TF and TFBS. The success rate of cross-validation on the collected dataset reached 76.6% (Qian et al., 2006). More accurate predictions on DNA binding preferences are required. If we know the targets regulated by the querying TF, better predictions of DNA binding preferences can be expected. Right now, targets of TFs can be extracted from the
large experimental data generated by various high-throughput biotechniques, such as the ChIP-chip experiment. Therefore, in this contribution, by integrating both the InterPro (Apweiler et al., 2001) annotations of TF and the TF-Targets relationships, an improved strategy based on the Nearest Neighbor Algorithm (NNA) is introduced to predict DNA binding preferences of novel TFs.

To examine the performance of our predictor, it was tested on a dataset consisting of 3430 true TF—transcription factor target (TFT) genes—TFBS triplets and 7000 artificial TF-TFT-TFBS triplets and achieved a success rate of 87.0%. This result indicates that the method we developed could be promising in TF DNA binding preference research.

2 MATERIALS AND METHODS

2.1 Positive dataset

For TFs as well as their targets and binding sites, the original dataset came from TRANSFAC v7.0 (Matys et al., 2006; Wingender et al., 1996). Then the original dataset was filtered in the following steps: (1) 327 TFs and 113 TFTs without SwissProt accessions were removed, and associated 407 TFBSs were filtered. (2) 743 TFBSs with length shorter than 5 bp or longer than 25 bp were removed, since most of the TFBSs’ length is within this range. (3) Finally, a positive dataset with 3430 TF-TFT-TFBS triplets which covered 143TF, 1416TFT and 571TFBS was built (c.f. Supplementary Material 1, Table 1).

2.2 Negative dataset

A negative dataset was randomly generated by shuffling the TFBS column in the collected positive dataset according to the following steps: (1) each TF-TFT-TFBS triplet is assigned a random number. (2) TFBSs are shuffled according to the random numbers while the TFs and TFTs are not changed. The details are illustrated in Figure 1. (3) The reduplicated record(s) that already existed in the positive dataset are removed. (4) Steps 1, 2 and 3 are repeated twice. (5) Finally, we achieve a negative dataset with about 7000 records, about two times as large as the positive dataset.

To minimize the sampling bias, we sampled 20 different negative datasets according to the above algorithm.

Table 1. Functional domain compositions of TFs T00333, T00511 and T00513

<table>
<thead>
<tr>
<th>Transcription Factor</th>
<th>Uniprot accession</th>
<th>InterPro annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>T00333 (<a href="http://www.gene-regulation.com/cgi-bin/pub/databases/transfac/getTF.cgi?AC=T00333">http://www.gene-regulation.com/cgi-bin/pub/databases/transfac/getTF.cgi?AC=T00333</a>)</td>
<td>P06536</td>
<td>IPR000536</td>
</tr>
<tr>
<td>T00511 (<a href="http://www.gene-regulation.com/cgi-bin/pub/databases/transfac/getTF.cgi?AC=T00511">http://www.gene-regulation.com/cgi-bin/pub/databases/transfac/getTF.cgi?AC=T00511</a>)</td>
<td>P22199</td>
<td>IPR000324</td>
</tr>
</tbody>
</table>

Each record is assigned a random number. Then the TFBS column is sorted according to the random number assigned, while the TF and TFT column remains unchanged.

2.3 Numeric representation system for TFs and TFTs

The prerequisite for building a predictor is to develop feasible numeric representation systems for proteins as well as for nucleotide sequences. In our contribution, InterPro (Apweiler et al., 2001) annotations were used to represent protein samples (TF, TFT) and 100D binary vectors were used to represent nucleotide sequences of DNA binding sites (TFBS). In the following section, we will discuss these two numeric representation systems in detail.

First, we extract the InterPro (Apweiler et al., 2001) annotations of each TF/TFT through using the Protein2Ipr mapping provided by InterPro (Apweiler et al., 2001) which altogether contains 8151 entries. Then, each TF is encoded with an 8151D (dimensional) vector. Each component of the 8151D vector, take the ith for example, is set to be 0 or 1, indicating whether this protein hit the ith InterPro (Apweiler et al., 2001) entry. This can be formulated as,

\[
T = \begin{pmatrix}
    t_1 \\
    t_2 \\
    \vdots \\
    t_i \\
    \vdots \\
    t_{8151}
\end{pmatrix}
\]

where

\[
t_i = \begin{cases}
1, & \text{if the } i\text{th InterPro entry} \\
0, & \text{otherwise}
\end{cases}
\]

In the same manner as the TF, each TFT is also encoded with 8151D binary vector,

\[
G = \begin{pmatrix}
    g_1 \\
    g_2 \\
    \vdots \\
    g_i \\
    \vdots \\
    g_{8151}
\end{pmatrix}
\]

where

\[
g_i = \begin{cases}
1, & \text{if the } i\text{th InterPro entry} \\
0, & \text{otherwise}
\end{cases}
\]

By doing so, various kinds of knowledge covered by InterPro (Apweiler et al., 2001), such as Pfam (Finn et al., 2006) annotations, PRINT (Attwood, 2002) annotations are now extracted for building the predictor. It is anticipated that the predictor will achieve better performance, since this functional related information, other than a raw primary sequence level information, is taken into consideration.

2.4 Numeric representation system for nucleotide sequences

Without information loss, the TFBS was encoded with a 0-1 system (Bhasin et al., 2005; Jia et al., 2006) according to the following steps: (1) At first, TFBSs with a length <25bp are extended to exactly 25bp through adding ‘N’ suffixes, e.g., 21bp nucleotide sequence ‘TTCGATCGATCGATCGATCGT’ will be extended to ‘TTCGATCGATCGATCGATCGTNNNNN’, meanwhile TFBSs with length of exact 25bp remain unchanged.
(2) Then, these TFBSs can be represented in a $25 \times 5 = 125D$ dimensional vector, e.g. 'TTCGATCGATCGATCGATCGNNN' will be represented in a 125D binary vector as,

\[
\begin{array}{cccccc}
0 & 1 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 \\
\end{array}
\]

since four different nucleotides were encoded with four orthogonal 5D binary vectors as

- 'A' := 00001
- 'C' := 00010
- 'G' := 00100
- 'T' := 01000
- 'N' := 10000

(3) Finally, each TFBS can be formulated as

\[
D = \begin{pmatrix}
d_1 \\
d_2 \\
\vdots \\
d_j \\
\vdots \\
d_{125}
\end{pmatrix},
\]

where, $d_i$ can be either 0 or 1.

2.5 The hybridization space

Obviously, DNA binding preferences can be inferred by predicting the interactions among TF, TFT and TFBS as mentioned above. To facilitate predicting interactions among TF, TFT and TFBS, a numeric representation was developed to cover TF-TFT-TFBS triplets. This can be done as follows: Suppose $T_x$, $G_z$, and $D_y$ are the $x$th TF, $y$th TFT, and $z$th TFBS, respectively.

Fig. 1. Shuffle the TFBSs to generate negative dataset.
Correctly Predicted true TF - TFT - TFBS triplets
Correctly Predicted artificial TF - TFT - TFBS triplets

\[ R = TGD(u, v, z) = T_{x} \oplus G_{z} \oplus (k \cdot D_{y}) \]

\[
= \begin{pmatrix}
I_{x,1} & I_{x,2} & \cdots & I_{x,8151} \\
I_{z,1} & I_{z,2} & \cdots & I_{z,125} \\
\vdots & \vdots & \ddots & \vdots \\
I_{t,1} & I_{t,2} & \cdots & I_{t,125}
\end{pmatrix}
\oplus
\begin{pmatrix}
G_{x,1} & G_{x,2} & \cdots & G_{x,8151} \\
G_{z,1} & G_{z,2} & \cdots & G_{z,125} \\
\vdots & \vdots & \ddots & \vdots \\
G_{t,1} & G_{t,2} & \cdots & G_{t,125}
\end{pmatrix}
\oplus
\begin{pmatrix}
k \cdot d_{y,1} \\
k \cdot d_{y,2} \\
\vdots \\
k \cdot d_{y,125}
\end{pmatrix}
\]

(5)

where the operator \( \oplus \) represents the orthogonal sum, \( k \) is the weight to facilitate removing the bias caused by the contribution difference between two encoding systems, \( T \) is the transpose operator and \( T_{x}, I_{x, \alpha}, I_{z, \beta}, D_{y} \), and \( d_{y, \gamma} \) have the same meanings as listed in formulae 1–3.

By using this hybridization approach, all the TF-TFT-TFBS triplets can be represented in the simple form of a \((16427D + 8151D + 125D)\) vector. Similar approaches have also been used in several previous works, such as predicting protein–protein interaction (Chou and Cai, 2006) and achieving a very good performance, which indicates that the hybridization approach provides an efficient encoding system for numeric representation of interacting pairs/triplets.

### 2.6 The nearest neighbor algorithm

The NNA is one of the most widely used classifiers which finds the closest (according to some distance metric) training point to the unknown point and predicts the category of that training point. Put more formally, given a collection of points \( S \) and a query point \( q \), what is the point \( x \) closest to \( q \) in \( S \)? The nearest-neighbor classifier predicts that the category of \( q \) is the same as the category of \( x \). It is particularly useful when the underlying distribution is unknown and has been widely used in previous works, such as predicting protein–protein interaction (Chou and Cai, 2006), protein quanyrdary structure classification (Yu et al., 2006) and TF classification (Qian et al., 2006), and achieved good performance.

### 3 RESULTS AND DISCUSSION

To some extend, TFs with similar biological functions may share similar DNA binding sites. For example, data from TRANSFAC V7.0 show that TFs T00333 (SwissProt Accession: P06536) and T00511 (SwissProt Accession: P22199) belong to the family ‘Steroid hormone receptors’.

The prediction will be better if more biological information can be taken into consideration. In this contribution, not only functional annotations of TFs but also that of TF targets are integrated into our prediction system. In the next paragraph, we will illustrate our method (c.f. see Sections 2.1 and 2.2).

To test the performance of our predictor, Jackknife cross-validation test (Bhasin et al., 2005; Cai and Chou, 2006; Chou and Cai, 2004, 2006; Jia et al., 2006; Qian et al., 2006a; Qian et al., 2006b) and 10-fold cross-validation test were adopted. In our implementations, jackknife cross-validation tests were operated as follows: for each TF-TFT-TFBS triplets in the dataset, we applied our predictor to predict its property. Predictor successes occur when the predicted property agrees with the truth. Finally, the success rate can be calculated as

\[
\text{success rate for on positive dataset} = \frac{\text{Correctly Predicted true TF - TFT - TFBS triplets}}{\text{true TF - TFT - TFBS triplets}}
\]

\[
\text{success rate for on negative dataset} = \frac{\text{Correctly Predicted artificial TF - TFT - TFBS triplets}}{\text{artificial TF - TFT - TFBS triplets}}
\]

(6)

In order to minimize the sampling bias, we randomly generated 20 negative datasets by using the described method in Figure 1 and did our computational experiments on these datasets 20 times. Table 2 shows the average performance.

In our dataset consisting of 3430 true TF-TFT-TFBS triplets and 7000 artificial TF-TFTS-TFT triplets, when \( k \) is set to 0.5 (formula 5), the success rate on positive and negative

<table>
<thead>
<tr>
<th>Methods</th>
<th>Our previous work (Qian et al., 2006)</th>
<th>NNA</th>
<th>SVM (polynomial kernel)*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Jackknife cross-validation test</td>
<td>10-fold cross-validation test</td>
<td>Jackknife cross-validation test</td>
</tr>
<tr>
<td>Success rate on positive dataset</td>
<td>71.90% NA</td>
<td>84.70% 83%</td>
<td>NA* 76.40%</td>
</tr>
<tr>
<td>Success rate on negative dataset</td>
<td>78.90% NA</td>
<td>89.30% 89.10%</td>
<td>NA* 88.00%</td>
</tr>
<tr>
<td>Overall success rate</td>
<td>76.60% NA</td>
<td>87.90% 87%</td>
<td>NA* 84.10%</td>
</tr>
</tbody>
</table>

* Jackknife tests for SVMs on our dataset take too much time to be executed.
datasets is 84.7 and 89.3%, respectively, and on overall datasets reaches 87.9% (Table 2).

Meanwhile 10-fold cross-validation tests were operated as shown in the following steps: (1) we randomly split the dataset including both positive data and negative data into 10 portions. (2) Then, for each portion, we tried to predict the category of each sample. The predictor succeeds if it can correctly predict the category of one sample. (3) Finally, the success rate of the 10-fold cross-validation test on positive and negative datasets is 83.0 and 89.1%, respectively, and on overall datasets reaches 87.0% (Table 2).

Support Vector Machine (SVM) algorithm with polynomial kernel (Joachims, 1999) as a classification method was also adopted to perform the prediction.

The total result shows that DNA binding preferences are closely correlated to both their function and their target’s function. Through comparison with our previous work which uses only the TF and TFBS information (Qian et al., 2006), we can see that the prediction performance with TF-TFT-TFBS triplets increases as expected when we take TFTs into consideration.

Here, we also show the improved proportion of the correct prediction compared to our previous work (Qian et al., 2006). Because the negative dataset is randomly generated, only the positive dataset was compared. The comparison is illustrated in the following Table 3.

In Table 3, we can see that there are 211 duplexes which are correctly predicted in TF-TFT-DTF triplets, but failed in TF-DFT duplexes. However, only 41 duplexes were correctly predicted in TF-DFT duplexes, but failed in TF-TFT-DTF triplets. This means our current method does effectively improve the prediction performance. The detailed duplex list concerned in Table 3 can be found in Supplementary Material 2.

However, there are still several difficulties related to this information integrative approach. First, it is difficult to determine which TFBS candidates should be chosen. When the length of TFBS is L, the count of TFBS candidates would reach 4^L, which is a huge number and needs lots of computation time. Nowadays more and more statistical methods have been developed to identify sequence motifs (D’Haeseleer, 2006).

In our future works, these statistical methods will be integrated, when motifs generated by them, instead of enumerated 4^L sequence fragments, are considered as the inputs of this approach. The second bottleneck is the limited number of known TF-TFBS duplexes from TRANSFAC (Matys et al., 2006; Wingender et al., 1996). More prior information would improve the predictions.

4 CONCLUSIONS

Specific DNA binding is a fundamental issue in understanding transcription regulatory mechanisms. Therefore, predicting DNA binding preferences is a critical problem in this research area. In this contribution, we applied a familywise approach to predict DNA binding preferences by using functional related TFs through integrating functional domain compositions. The performance of our predictor reached 87.0%. This contribution provides a novel way to investigate TFs lacking DNA binding preferences and provides a useful tool to develop novel TF engineering which can have fascinating implications in the discovery of new drugs (Jamieson et al., 2003).

ACKNOWLEDGEMENTS

This work is supported by National Basic Research Program of China No. 2006CB910700, No. 2003CB715900, No. 2004CB518660 and National High-Tech Research and Development Program of China (863 Program) No. 2006AA02Z320.

Conflict of Interest: none declared.

REFERENCES


Table 3. Comparison of the prediction performance of TF-DFT duplex and TF-TFT-DTF triplets

<table>
<thead>
<tr>
<th></th>
<th>Correctly predicted in TF-DFT duplex</th>
<th>Failed in TF-DFT duplex</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correctly predicted in TF-TFT-DTF triplets</td>
<td>252</td>
<td>211</td>
</tr>
<tr>
<td>Failed in TF-TFT-DTF triplets</td>
<td>41</td>
<td>48</td>
</tr>
</tbody>
</table>

Note: The sum 552(211 + 252 + 48 + 41) is less than 3430 (the number of positive TF-TFT-DTF) because TFs can have several targets. And the sum 552 is less than 3456 [the number of positive TF-DFT in previous work (Qian et al., 2006)] with only a part of previous datasets covered for not all TFs have target information in TRANSFAC v7.0 (Matys et al., 2006).