REGULATORY GENOMIC SEQUENCES

RECOGNITION OF BINDING SITES FOR THE TRANSCRIPTION FACTORS SREBP, PPAR, HNF4, COUP-TF, AND SF-1 BY A GENETIC ALGORITHM BASED ON ITERATIVE DISCRIMINANT ANALYSIS

*Levitsky V.G., Ignatieva E.V., Proscura A.L., Pozdnyakov M.A., Busygina T.V.*

Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia, e-mail: levitsky@bionet.nsc.ru

*Corresponding author

Key words: transcription factor binding sites recognition, lipid metabolism, endocrine system, genetic algorithm, discriminant analysis

Summary

Motivation: Development of methods for search for transcription factor binding sites (TFBSs) is important for investigation of regulatory regions in eukaryotic genes. The accuracy of currently used methods is insufficient for correct recognition of binding sites of a given transcription factor (TF) or a group of factors of a single family, class, etc.

Results: We propose a new approach to search for TFBSs by the example of five site types: SRE, PPRE, HNF4, COUP-TF, and SF-1. The approach involves partitioning the regions including TFBSs into local regions and selection of the most suitable dinucleotides for each of the regions. Cross-validation tests of the functions of TFBS recognition allowed division of the sites into two groups which agree with the structure-functional subdivision of these TFs into two superclasses: (1) Basic Domains and (2) Zinc-coordinating DNA-binding domains.

Availability: The program for site recognition is included into the GeneExpress system; section “RegScan”, http://wwwmgs.bionet.nsc.ru/mgs/programs/sitega/.

Introduction

The 5'-regulatory regions of genes are characterized by extremely complex structure and abundance of regulatory elements (Kolchanov et al., 2002). The regulatory regions of eukaryotic genes are following a block-hierarchical model of organization. Transcription-factor binding sites are the most important units of the structure. A single set of regulatory elements (blocks) present in a certain regulatory region can give rise to a great variety of transcription complexes and, as a result, implementation of various expression patterns. Core domains are recognized within a TFBS. There can be one or several of them, and they are separated with variable regions (spacers). The presence of cores is related to the fact that transcription factors have a block structure and may contain several domains or subunits, performing specific functions (Wingender, 1997). Taking into account the importance of the block structure of TFBSs, we used this principle as the basis for the algorithm for their recognition. To reveal the block structure of a site and its flanks, we followed an approach based on the determination and analysis of context-homogenous DNA regions. We analyzed samples of binding sites of five types: SRE, PPRE, HNF4, SF-1 and COUP-TF. Formerly, we showed that the transcription factors SREBP and PPAR, interacting with the sites SRE and PPRE, respectively, and the factor HNF4 are involved in the regulation of transcription of lipid metabolism genes (Ignatieva et al., 2000). The SF-1 factor of the nuclear receptor family is known to be a regulator of development and function of the hypothalamic-pituitary-gonadal complex and adrenals (Luo, 1999). This factor plays a crucial role in regulation of the transcription of genes controlling steroid production (Busygina et al., 2000). The transcription factor COUP-TF is a negative transcription regulator (Tsai, Tsai, 1997), and its binding site frequently occurs in the regulatory regions of genes of the lipid metabolism system, as well as in the regulatory regions of genes of the endocrine system (Ignatieva et al., 2000; Busygina et al., 2000).

Methods and Algorithms

The method we developed for recognition of TFBSs consists of two stages: (1) search for a partition of the region including the site and adjacent parts into local regions and (2) selection of major dinucleotides within each region. Both stages were implemented with the use of a genetic algorithm, which utilizes a linear discriminant function of dinucleotide frequencies characteristic of the local regions as a functional. Here we present analysis of samples of 120 bp long TFBS regions bearing the binding site of a certain TF.

Sample sizes are listed in Table. Random sequences obtained by mixing positive samples were used as negative samples. The parameter of the method is p, the number of local regions into which the regions is partitioned. In our study, p=10. The method was described in more detail in (Levitski, Katokhin, 2001; Levitsky et al., 2001; Levitsky, Katokhin, this issue).
For recognition of a TFBS in a nucleotide sequence, a 120 bp long sliding window are considered. Denote this fragment as X. For each position of the window, find the recognition function value according to the following equation:

\[
\varphi(X) = \frac{1}{R^2} \sum_{n=1}^{N} \sum_{k=1}^{N} \left[ \left( f_n^{(1)}(X) - f_n^{(2)} \right) \times \left( f_n^{(1)} + f_n^{(2)} \times S_n^{-1} \times \left( f_k^{(2)} - f_k^{(1)} \right) \right) \right]
\]  

(1)

Here, \( f_n^{(1)} = f_n^{(1)}(i, p) \) is the frequency of the \( i \)-th dinucleotide for the \( p \)-th region averaged over the site sample; and \( f_n^{(2)} \) is the corresponding frequency for the negative sample; and \( f_n(X) \) is the dinucleotide frequency vector constructed with regard to the partition of the fragment X under study into local regions. The \( n(i, p) \) dependence is defined during the selection of the most significant dinucleotide frequencies for the partition regions. Denote the number of dinucleotides used for \( p \)-th region as \( n_p \leq 16 \). Then the total number of variables in (1) is determined as

\[
N = \sum_{p=1}^{P} n_p
\]  

(2)

In equation (1) \( S^{-1} \) denotes the reverse matrix for the combined covariation matrix; \( S = S^{(1)} + S^{(2)} \); \( S^{(1)} \) and \( S^{(2)} \) are covariation matrices for the positive and negative sequence samples for the dinucleotide frequency vectors \( f_n^{(1)} \) and \( f_n^{(2)} \); and \( R^2 \) is the Mahalanobis distance (Mahalanobis, 1936) between the samples of sites and random sequences:

\[
R^2 = \sum_{k=1}^{N} \sum_{n=1}^{N} \left[ \left( f_n^{(2)} - f_n^{(1)} \right) \times S_n^{-1} \times \left( f_k^{(2)} - f_k^{(1)} \right) \right]
\]  

(3)

Values of function \( \varphi(X) \) close to 1 correspond to higher probabilities of site recognition. For predicting TFBSs, the program uses the significance level index \( \alpha \). The TFBS recognition function \( \varphi(X) \) (obtained according to (1)) was converted as follows:

\[
\varphi_\alpha(X) = \begin{cases} 
\frac{1 - \varphi(X)}{P_\alpha \times \sigma_\varphi}, & \text{if } |1 - \varphi(X)| < P_\alpha \times \sigma_\varphi \\
0, & \text{otherwise}
\end{cases}
\]  

(4)

Here, \( P_\alpha \) is the \( \alpha \) quantile of the standard normal distribution (for example, \( P_{0.95} = 1.96 \)), and \( \sigma_\varphi \) is the standard deviation of the values of the recognition function \( \varphi(X) \) over the site sequence sample. Nucleotide sequence regions with \( \varphi_\alpha(X) > 0 \) were considered to be potential TFBSs.

**Implementation and Results**

We compiled samples of TFBSs of five types from the data banks T RR D and EMBL (Table). The site samples are 120 bp long fragments of the regulatory regions of genes containing the sequence of a particular type in the center.

**Table.** TFBS samples used in the analysis.

<table>
<thead>
<tr>
<th>Site name</th>
<th>Transcription factor interacting with the site</th>
<th>Number of sequences</th>
<th>Number of sequences for cross-validation test</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRE*</td>
<td>SREBP (Sterol Regulatory Element Binding Protein)</td>
<td>27</td>
<td>27</td>
</tr>
<tr>
<td>PPRE**</td>
<td>PPAR (Peroxisome Proliferator-Activated Receptor)</td>
<td>16</td>
<td>11</td>
</tr>
<tr>
<td>HNF4</td>
<td>HNF4 (Hepatic Nuclear Factor 4)</td>
<td>35</td>
<td>33</td>
</tr>
<tr>
<td>COUP-TF</td>
<td>COUP-TF (Chicken Ovalbumin Upstream Promoter Transcription Factor)</td>
<td>33</td>
<td>12</td>
</tr>
<tr>
<td>SF-1</td>
<td>SF-1 (Steroidogenic Factor 1)</td>
<td>39</td>
<td>30</td>
</tr>
</tbody>
</table>

SRE, Sterol Regulatory Element

**PPRE - Peroxisome Proliferator-Responsive Element**

By the above-described method, we constructed recognition functions for sites of five types. The best partitions of site DNA sequences with adjacent flanks into local regions calculated for the five site types are shown in Fig. 1. The recognition functions were tested by cross-validation tests of DNA sequences from the samples collected by us. For these tests, we excluded sequences from different samples occurring in same genes and partly overlapping (Column 4 in Table) from the site samples used as learning samples for the recognition programs (Column 3 in Table).
Fig. 1. Optimum partitions of the regions [–60; +60] of sites with adjacent regions into local regions calculated for five types of sites: SRE, PPRE, HNF4, COUP-TF, and SF-1.

For estimation of the accuracy of recognition, we calculated correlation coefficients (CC), which describe the accuracy of site recognition. An increase in CC points to an increase in site recognition accuracy. We concluded that the sites HNF4 and PPRE (CC are equal to 0.76 and 0.72, respectively) demonstrated a greater accuracy of recognition, and the accuracy of COUP-TF, SF-1, and SRE is lower (CC 0.63, 0.54, and 0.39, respectively).

The results of cross-validation tests of the recognition function performed for the five types of binding sites are shown in Fig. 2. A site was considered recognized if the value \( \phi_s(\bar{X}) \) (equation 4) was positive at the significance level \( \alpha = 0.95 \).

Consider the results of the cross-validation tests in more detail. It is seen from Fig. 2 that four sites (COUP-TF, PPRE, HNF4 and SF-1) are more efficiently recognized with the recognition functions of one another. On the contrary, the site SRE is less recognizable with the recognition functions of COUP-TF, PPRE, HNF4 and SF-1. According to Fig. 2, recognized sites constitute 7.4, 3.7, 7.4, and 11.1%, respectively.

Fig. 2. Cross-validation tests of the recognition functions constructed for the samples of the binding sites COUP-TF, PPRE, SRE, HNF4, and SF-1. Y axis, the percentage of recognized sites. Equally filled bars denote recognition programs corresponding to learning samples of a particular site. Types of site samples for which the analysis was performed are indicated on the X-axis.
Thus, we found that classification of TFBSs on the base of cross-validation tests of recognition functions into two groups—
(1) COUP-TF, PPRE, HNF4, and SF-1 and (2) SRE—matches the structure-functional subdivision of the corresponding
TFs into two classes: (1) Basic helix-loop-helix/leucine zipper factors (bHLH/ZIP), superclass Basic domains, and (2) Cys4
zinc finger of nuclear receptor type, superclass Zinc-coordinating DNA-binding domains. This appears to be related to the
fact that the attribution of a factor to a particular class determines a context uniformity, which is revealed by cross-validation
tests of recognition functions.

It should also be mentioned that the sites COUP-TF, PPRE, SRE and HNF4 include two cores, and the site SF-1 includes
only one. This is reflected in the structure of partitions used for construction of site-recognition functions (Fig. 1). It is seen
that the central part of the partition of the one-core site SF-1 contains a 5 bp long region bordered by longer regions,
whereas the central parts of the two-core sites COUP-TF and HNF4 contain two partition regions. The partitions of the sites
PPRE and SRE do not show any clear dependence of the number of cores on the partition structure in the central parts of
the sites. Most probably, this illustrates the weak context signal of the site cores in comparison with the neighboring flanking
regions.

Acknowledgements

The authors are grateful to Prof. N.A. Kolchanov for fruitful discussion and Dr. A.V. Osadchuk for supporting the work on
the construction of the sample SF-1. The study was supported in part by the Russian Foundation for Basic Research (grants
№ 01-07-90376, 02-07-90355, and 00-04-49229); Russian Ministry of Industry, Science, and Technologies (grant
№ 43.073.1.1.1501); Siberian Branch of the Russian Academy of Sciences (Integration Project № 65); US National
Institute of Health (grant № 2 R01-HG-01539-04A2); and US Department of Energy (grant № 535228 CFDA 81.049).

References

1, 41–44.
2. Ignatieva E.V., Likhoshvai V.A., Ratushny A.V., Kosarev P.S. Knowledge base on molecular-genetical foundations of lipid metabolism
(BGRS'2002), this issue.
7. Luo X., Ikeda Y., Lala D., Rice D., Wong M., Parker K.L. Steroidogenic factor 1 (SF-1) is essential for endocrine development and
18, 229–240.