TRANSCRIPTION FACTOR BINDING SITES RECOGNITION BY THE REGULARITIES MATRICES BASED ON THE NATURAL CLASSIFICATION METHOD

Vityaev E.E.*, 1, 2, Lapardin K.A. 2, Khomicheva I.V. 3, Levitsky V.G. 2, 3

1 Institute of Mathematics, SB RAS, Novosibirsk, 630090, Russia; 2 Novosibirsk State University, Novosibirsk, 630090, Russia; 3 Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia

* Corresponding author: e-mail : vityaev@bionet.nsc.ru

Key words: knowledge discovery and data mining, machine learning, eukaryotic promoter recognition, transcription factor binding sites

SUMMARY

Motivation: Numerous principles of constructing classifications are currently known. We propose the definition of “natural” classification and based on the definition a principally new approach to the classifications of nucleotide sequences.

Results: A method for constructing the “natural” classification, algorithm, and software system DNANatClass have been developed. As the application result we propose the regularities matrices describing SF1 and EGR1 transcription factor binding sites.


INTRODUCTION

Position weight matrix is the most common method for the transcription factor binding sites (TFBSs) recognition. In this paper we present the regularities matrices that arise from the concept of natural classification in its application to the nucleotide sequences. The concept of natural classification was investigated and developed in the previous papers (Vityaev, 1983; Vityaev, Kostin, 1992; Vityaev et al., 2002). The main property of the regularities matrices is that each of the nucleotides A, T, G, C in each position of the matrix is characterized by its regularities connecting it with nucleotides in other positions, whereas the weight matrices estimates the contribution of each nucleotide taken separately without any interconnectivity.

Numerous principles of constructing classifications are currently known. The classifications are based on the hypothesis of compactness and various measures of closeness in a feature space, on resemblance of standards, supertargets, various criteria of classification quality and quality functionals, separation of distribution mixtures, etc. (Classification and Clustering, 1977). In contrast to the above-listed classifications the objective of the “natural” classification is discovering the laws of nature. There are different definitions of the natural classification that were done by the naturalists in different times (see overview in Zabrodin, 1981). We propose the definition of the “natural” classification that is in accordance with the definitions of naturalists: “Objects should be divided into classes in accordance with the regularities satisfied by the objects. Objects of one class should obey one group of regularities, and objects of different classes should obey different groups of regularities. Objects of one
class should also possess some integrity which is understood as mutual prediction of object properties" (Vityaev et al., 1983).

**METHODS AND ALGORITHMS**

The following method realizes the above definition of the natural classification and includes three steps: regularities determining, classes formation and recognition (Vityaev et al., 2006a).

1. **Regularities discovery.**

**Definition 1.** The rule \( (P_{ij1}^1 \land \ldots \land P_{ijk}^k \Rightarrow P_{0i0}^0) \) is the *probabilistic law* if and only if:

1. \( \mu(P_{ij1}^1 \land \ldots \land P_{ijk}^k) > 0; \)
2. \( \mu(P_{ij1}^1 \land \ldots \land P_{ijk}^k) > \mu(P_{0i0}^0 | P_{ij1}^1 \land \ldots \land P_{ijk}^k), \) where \( \ldots \land \ldots \land \).

means the absence of one or more predicates in the premise of the rule, and conditional probability is defined as follows: \( \mu(P_{0i0}^0 | P_{ij1}^1 \land \ldots \land P_{ijk}^k) = \mu(P_{0i0}^0 | P_{ij1}^1 \land \ldots \land P_{ijk}^k) \times \mu(P_{ij1}^1 \land \ldots \land P_{ijk}^k) \).

For the predicate \( P_{ij1}^1 \) the index \( i1 \) means the position number, \( j1 \) means one of the nucleotide \( \{A,T,G,C\} \), \( \varepsilon = 0/1 \) means that the predicate has/hasn’t the negation. For example, the predicate \( P_{1ikA} \) means that in the position \( ik \) there is the nucleotide \( A \).

Let \( \mu(\phi) = \mu(P_{0i0}^0 | P_{ij1}^1 \land \ldots \land P_{ijk}^k) \) be the conditional probability of the rule \( \phi \). Given the sample of the sequences we discover the set of regularities \( F \). By the estimation of regularity we mean the value \( \mu^p(\phi) = -\ln(1-\mu(\phi)) \) calculated with confidence level \( \beta \).

2. **Classes discovery.** Let us define the criterion of regularities interconnection. By the tuple of properties values \( x_{s1}, \ldots, x_{sm} \) we call the set \( \{Y_{s1}, \ldots, Y_{sm}\} \), \( Y_{st} \subset I_{st}, Y_{st} \neq \emptyset, t = 1, \ldots, m, I_{st} - \) the set of all values of the feature \( st \). We designate that the regularity \( (P_{ij1}^1 \land \ldots \land P_{ijk}^k \Rightarrow P_{0i0}^0) \) is *applied* to the set \( \{Y_{s1}, \ldots, Y_{sm}\} \), if \( \{i0,i1,\ldots,i_k\} \subset \{s1,\ldots,sm\} \) and also \( x_{itj} \in Y_{it} \) if \( \varepsilon = 1 \) and \( (x_{itj} \notin Y_{it}) \) if \( \varepsilon = 0 \), \( t = 1, \ldots, k \). If the regularity is applied to the set \( \{Y_{s1}, \ldots, Y_{sm}\} \) and the conclusion of the rule \( P_{0i0}^0 \) is fulfilled for that set \( (x_{i0j} \in Y_{i0} \text{ if } \varepsilon = 1 \text{ and } x_{i0j} \notin Y_{i0} \text{ if } \varepsilon = 0) \), then we say that the regularity is *satisfied* for that set, but if conclusion is not fulfilled, then we say that the regularity is *falsified* for that set. By the criterion of regularities interconnection on the set \( \{Y_{s1}, \ldots, Y_{sm}\} \) we designate the value:

\[
\Lambda(\{Y_{s1}, \ldots, Y_{sm}\}) = \sum_{\phi \in F} \mu^p(\phi) - \sum_{\phi \in O} \mu^p(\phi)
\]

reaches the local maximum.

The set \( \{Y_{s1}, \ldots, Y_{sm}\} \) can be presented as the matrix. For example the sequence [A][A][C][A][G][C][T][T][A][C][G][G][T][A][A][G][G][G][C][T] can be presented as matrix \( M(Y_{s1}, \ldots, Y_{sm}) \):

<table>
<thead>
<tr>
<th>( A )</th>
<th>( T )</th>
<th>( G )</th>
<th>( C )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

In addition to the matrix \( M(Y_{s1}, \ldots, Y_{sm}) \) we define the regularity matrix \( R(Y_{s1}, \ldots, Y_{sm}) \) as the matrix of predictions of the cells of the matrix \( M(Y_{s1}, \ldots, Y_{sm}) \) by regularities. The sum of values of regularity matrix \( R(Y_{s1}, \ldots, Y_{sm}) \) is equal to the criterion \( \Gamma(\{Y_{s1}, \ldots, Y_{sm}\}) \).

Also we use the involvement matrix \( I(Y_{s1}, \ldots, Y_{sm}) \) to show the involvement of all predicates of the regularities in there interconnection, which have estimation \( \mu^p(\phi) \) for each predicate of the regularity.
3. Recognition. Given the control set $B$ of sequences, class $O = \{Y_1, ..., Y_n\}$, and the set of regularities $F$ we can recognize the positive and control samples by calculating the score $I(\{Y_1, ..., Y_n\})$ for every training and control sequence. When we define some threshold of the score, we can calculate the true/false positive rates for the training and control sets.

**IMPLEMENTATION AND RESULTS**

For the TFBSs recognition we have chosen the samples of sites SF1, EGR1. The train data sets were extracted from the TRRD database (Kolchanov et al., 2002). We added to the positive samples the sets of randomly generated sequences, which were generated with the same frequencies as for the positive samples. The number of randomly generated sequences was ten times more then the number of positive sequences. Then using that mixed sample we performed the classification of the whole data and discovered the class(es) for the positive samples. Exactly the one class was discovered for the SF1, EGR1 sites samples.

The negative control sample was randomly generated with the frequencies as in the positive samples. For the recognition of the positive and control sequences we first performed the classification of that samples. Then the score $I'$ was calculated for the positive and control sequences that were classified as belonging to the class. We defined the threshold for which the 50 % of positive sequences were recognized as belonging to the class. With this threshold we calculated the false positive rate. The more detailed description of results is depicted in the following table.

<table>
<thead>
<tr>
<th>Site</th>
<th># positive sequences</th>
<th># negative control sequences</th>
<th># of regularities belonging to class</th>
<th># of classified positive</th>
<th># of classified negative</th>
<th># of recognized positive</th>
<th># of recognized negative</th>
<th>False positive rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>SF1</td>
<td>54</td>
<td>100000</td>
<td>1670</td>
<td>54</td>
<td>81940</td>
<td>3900</td>
<td>25</td>
<td>2/100 000</td>
</tr>
<tr>
<td>EGR1</td>
<td>22</td>
<td>110000</td>
<td>789</td>
<td>16</td>
<td>25502</td>
<td>900</td>
<td>7</td>
<td>0/110 000</td>
</tr>
</tbody>
</table>

The class $[T/C][C][A][A][G][G][T/C][C][A][G]$ was discovered for the SF1 site, where $[T/C]$ means that on the first place there can be one of two nucleotides T or C. The class $[G][C][G][G][G][G][G][C][A][G]$ was discovered for the EGR1 site.

The regularity matrix $R([T/C][C][A][A][G][G][T/C][C][A][G])$:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>0.00</th>
<th>0.00</th>
<th>0.00</th>
<th>0.00</th>
<th>0.00</th>
<th>0.00</th>
<th>0.00</th>
<th>0.00</th>
<th>0.00</th>
<th>0.00</th>
<th>0.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>0.00</td>
<td>484.95</td>
<td>643.73</td>
<td>481.73</td>
<td>421.14</td>
<td>872.68</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>G</td>
<td>154.92</td>
<td>0.00</td>
<td>2.31</td>
<td>61.84</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>C</td>
<td>0.00</td>
<td>0.00</td>
<td>4.13</td>
<td>9.89</td>
<td>103.06</td>
<td>634.36</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

The regularity matrix $R([G][C][G][G][G][G][G][C][A][G][G])$:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>0.00</th>
<th>0.00</th>
<th>0.00</th>
<th>0.00</th>
<th>0.00</th>
<th>0.00</th>
<th>0.00</th>
<th>0.00</th>
<th>0.00</th>
<th>0.00</th>
<th>0.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>G</td>
<td>0.00</td>
<td>447.36</td>
<td>0.00</td>
<td>-47.98</td>
<td>-15.30</td>
<td>-2.69</td>
<td>0.00</td>
<td>515.56</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>C</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

BGRS’2006
DISCUSSION

Further we plan to improve the method and use it in cooperation with the ExpertDiscovery method (Vityaev et al., 2006b). We can discover the complex signals by the ExpertDiscovery system and use them as ordinary properties in the classification system DNANatClass.

ACKNOWLEDGEMENTS

The work is partially supported by the Russian Foundation for Basic Research No. 05-07-90185-v, Scientific Schools grant at the President of the Russian Federation No. 4413.2006.1, Innovation project IT-CP.5/001 “Development of software for computer modeling and design in postgenomic system biology (system biology in silico)”.

REFERENCES