Transcription Factor Binding Site Discovery by the Probabilistic Rules

Irina Khomicheva[1], Alexander Demin[2], and Evgeny Vityaev[3]

[1] Institute of Cytology and Genetics SB RAS, Lavrentyev aven., 10
630090 Novosibirsk, Russia
[2] A.P.Yershov Institute of Information Systems SD RAS, Lavrentyev aven., 6,
630090 Novosibirsk, Russia
[3] Sobolev Institute of Mathematics, Koptyug aven. 4,
630090 Novosibirsk, Russia
{khomicheva@bionet.nsc.ru, alexandredemin@yandex.ru, vityaev@math.nsc.ru}

Abstract. Control of gene expression at the level of transcription is achieved by nuclear factors that bind to regulatory elements, short DNA sequence motifs, called transcription factor binding sites. The development of reliable methods for binding site recognition is an important step in large-scale genome analysis. The Data Mining approaches adapted to bioinformatics tasks show high efficiency. Yet the specificity of the regulatory region analysis task consists in the high false-positive rates. In this paper the program system 'Discovery' was applied to tasks of binding site recognition. 'Discovery' makes a semantic probabilistic inference and finds the statistically significant probabilistic rules. The hypothesis class is defined by the expert in dialog mode. In this paper we demonstrate that 'Discovery' is consistently more accurate than the traditional weight matrices in binding site prediction task, as was established for three families of transcription factors.

Keywords: semantic probabilistic inference, probabilistic rule, transcription factor binding sites, prediction.

1 Introduction

Analysis of gene transcription regulatory regions is of great importance for understanding molecular mechanisms of transcription.

The task of transcription factor binding sites (TFBS) prediction is methodologically difficult due to a high variety of DNA binding proteins and the degeneracy of TFBSs conferred on them by the tissue- and stage-specific regulatory mechanisms. These sequences vary in length, position, redundancy, orientation in the DNA chain, and bases. As a direct consequence, the problem of a large number of false-positives necessarily takes place manifesting itself in the poor predictive performance of the corresponding software.

The problem of regulatory region analysis challenges the Data Mining and Machine Learning approaches. Machine Learning algorithms intended for addressing bioinformatics tasks are: hidden markov models, decision trees, neural network, genetic algorithms, etc. [1].

The traditional approach to predict TFBSs is the positional weight matrix, PWM, a very powerful tool, but still with some drawbacks and limitations [2]. PWM and consensus-based methods involve an explicit assumption that the contribution of each nucleotide position to the binding affinity is independent and the effect produced on the binding strength is cumulative. In PWM, elements simply correspond to the probabilities of observing each nucleotide at each position. Numerous works [3, 4, 5, 6] indicate that the nucleotides of TFBSs cannot be treated independently. This assumption is invalid and contradicts the processes underlying the biological model. Predictions can be further improved by taking into account the sequence context in which a predicted site is located.

Despite an evident importance of noncoding sequences to gene regulation, our ability to describe and properly localize them is extremely limited. The known approaches are rather restricted as they are confronted with the lack of sufficient training data and the degeneracy of the biological objects under analysis.

In this paper we applied 'Discovery' system to knowledge acquisition tasks on DNA sequences. Unlike PWM and consensus methods, 'Discovery' reveals mutual interdependences among the nucleotides which, in the general case, are rather distant from one another.

In this paper we demonstrate that 'Discovery' is consistently more accurate than the traditional weight matrices in TFBS prediction tasks, as was established for three families of binding factors.

2 'Discovery' as Implemented for Bioinformatics

As the training data for 'Discovery' system we used the samples of nucleotide sequences, putative TFBSs, that were organized in the data table. Each table row contained the binding site name and its nucleotide sequence. For example,

>S1916	gtccgtgggt
>S4809	ttgggggcga
>S6067	gaggggcgg
>S6069	gcgggggcgg
>S5824	acggaggcgg

The 'Discovery' system reveals the probabilistic rules of the form:

$$(\boldsymbol{Pos}_1 = \boldsymbol{C}_1)^{\varepsilon 1} \,\&\, (\boldsymbol{Pos}_2 = \boldsymbol{C}_2)^{\varepsilon 2} \,\&\, \ldots \,\&\, (\boldsymbol{Pos}_k = \boldsymbol{C}_k)^{\varepsilon k} \rightarrow (\boldsymbol{Class} = \boldsymbol{1}) \,. \qquad (1)$$

where $(\boldsymbol{Pos}_i = \boldsymbol{C})^{\boldsymbol{\varepsilon}}$ – is a predicate, denoting that the sequence position $\boldsymbol{i}$ contains (if $\boldsymbol{\varepsilon} = \boldsymbol{1}$) or doesn't contain (if $\boldsymbol{\varepsilon} = \boldsymbol{0}$) the symbol $C \in \{a, t, g, c\}$; $(\boldsymbol{Class} = \boldsymbol{1})$ – target predicate, which implies that the nucleotide sequence is one of the binding sites of a particular transcription factor.

In general, 'Discovery' system makes a semantic probabilistic inference [7, 8], which allows the user to find all statistically significant rules $\boldsymbol{P_1} \& ... \& \boldsymbol{P_m} \to \boldsymbol{P_0}$ that predict the predicate $\boldsymbol{P_0}$ with the highest probability.

The semantic probabilistic inference is based on the definition of the probabilistic rule, which is as follows. The rule $\boldsymbol{P_1} \& ... \& \boldsymbol{P_m} \to \boldsymbol{P_0}$ is a probabilistic rule, if for any rule $\boldsymbol{P_{i1}} \& ... \& \boldsymbol{P_{ik}} \to \boldsymbol{P_0}$ such that $\{\boldsymbol{P_{i1}}, ..., \boldsymbol{P_{ik}}\} \subset \{\boldsymbol{P_1}, ..., \boldsymbol{P_m}\}$ the conditional probability satisfies the inequality $\boldsymbol{p}(\boldsymbol{P_0} \mid \boldsymbol{P_{i1}} \& ... \& \boldsymbol{P_{ik}}) < \boldsymbol{p}(\boldsymbol{P_0} \mid \boldsymbol{P_1} \& ... \& \boldsymbol{P_m})$.

Let us introduce the *correction* to the rule. The rule $\boldsymbol{P_1} \& ... \& \boldsymbol{P_m} \& \boldsymbol{P_{m+1}} \to \boldsymbol{P_0}$ is a *correction* to the rule $\boldsymbol{P_1} \& ... \& \boldsymbol{P_m} \to \boldsymbol{P_0}$, if the former rule was created by adding an any predicate $\boldsymbol{P_{m+1}}$ to the statement of the rule $\boldsymbol{P_1} \& ... \& \boldsymbol{P_m} \to \boldsymbol{P_0}$.

Consider the algorithm of making semantic probabilistic inference.

The algorithm provides a successive introduction of corrections to the rules and a consistent check for conformity to the criterion of being probabilistic rule. Checking all possible rules is a difficult, time-consuming computational task, and for that reason semantic probabilistic inference in practice involves two successive checks: the *basic check* and the *advanced check.*

The algorithm of making semantic probabilistic inference.

1. At the first step, a set of probabilistic rules $\boldsymbol{REG_1}$ is generated by the exhaustive search of all the rules $\boldsymbol{P_1} \& ... \& \boldsymbol{P_m} \to \boldsymbol{P_0}$ for $\boldsymbol{1} \le \boldsymbol{m} \le \boldsymbol{d}$ and checking for conformity to the criterion of being probabilistic. That is, $\boldsymbol{REG_1} = \{\boldsymbol{R_i}\}$, where $\boldsymbol{i} \in \boldsymbol{I_1}$, $\boldsymbol{R_i} = \boldsymbol{P_1} \& ... \& \boldsymbol{P_m} \to \boldsymbol{P_0}$, $\boldsymbol{1} \le \boldsymbol{m} \le \boldsymbol{d}$, $\boldsymbol{R_i}$ is the probabilistic rule. This step is referred to as a *basic check*, and the value *d* is referred to as the *depth of the basic check.*
2. At the *k*-th (*k* > 1) step, a set of probabilistic rules $\boldsymbol{REG_k}$ is generated by correcting all the rules that were found at the previous step and checking for conformity to the criterion of being probabilistic rule. That is, $\boldsymbol{REG_k} = \{\boldsymbol{R_i}\}$, $\boldsymbol{i} \in \boldsymbol{I_k}$, $\boldsymbol{R_i} = \boldsymbol{P_1} \& ... \& \boldsymbol{P_m} \to \boldsymbol{P_0}$, $\boldsymbol{m} = \boldsymbol{d} + (\boldsymbol{k} - \boldsymbol{1})$, $\boldsymbol{R_i}$ is a probabilistic rule, $\boldsymbol{R_i} \in \boldsymbol{Spec}(\boldsymbol{REG_{k-1}})$, where $\boldsymbol{Spec}(\boldsymbol{REG_{k-1}})$ is the set of all corrections to the rules in $\boldsymbol{REG_{k-1}}$.
3. The algorithm stops when no further correction to any rule is possible. The ultimate set of all rules $\boldsymbol{REG}$ is equal to the union of all $\boldsymbol{REG_k}$: $\boldsymbol{REG} = \bigcup_{\boldsymbol{k}} \boldsymbol{REG_k}$.

The depth of the basic check, that is, the maximum length of the rules that will be subject to the basic check, is specified by the user. In practice, the depth of the basic check is normally equal to 2 or 3.

As the algorithm proceeds, the conditional probability of the rules is assessed using a training set. To prevent the rules that fail to achieve significance, additional criteria are normally applied to assess statistical significance among the rules. We used the exact Fisher criterion applied to contingency tables [9]. The rules that fail to meet criteria are to be discarded even if they prove highly accurate on the training set.

The calculation of object score is critical for decision making if the nucleotide sequence is one of the binding sites of a particular transcription factor. 'Discovery' supports several procedures for calculating the score. For TFBS recognition purposes, we used the following procedure:

$$\boldsymbol{score} = \frac{\sum_{\boldsymbol{R} \in \boldsymbol{TR}} \boldsymbol{p}(\boldsymbol{R})}{\sum_{\boldsymbol{R} \in \boldsymbol{AR}} \boldsymbol{p}(\boldsymbol{R})}. \tag{2}$$

where $\boldsymbol{p(R)}$ – conditional probability of the rule $\boldsymbol{R}$, $\boldsymbol{AR}$ – all the rules discovered by the system, $\boldsymbol{TR}$ – all the rules that are applicable to the object.

Further the sequence score is compared to the threshold $\boldsymbol{\delta} \in [\boldsymbol{0}, \boldsymbol{1}]$. If the object score is higher than the threshold, then the object is one of the TFBSs. We defined the false positive (FP) rates based on false negative (FN) rate using the standard jackknife procedure.

3 Experimental Data Analysis

We analyzed the DNA targets of three protein families: sterol regulatory element binding protein (SREBP), early growth response factor 1 (EGR1), and Hepatocyte nuclear factor 4 (HNF4). The training data sets (sequences of TFBSs with flanks) were retrieved from the TRRD database [10].

We performed the accuracy comparison of 'Discovery' and PWM according to the standard jackknife procedure [11]. Totally, the data sets contained 38 sequences of SREBP binding sites, 22 (EGR1), 30 (HNF4) (Table 1). First of all, we tried PWM on different sequence lengths to reach the highest PWM recognition accuracy. When the optimal sequence lengths for PWM were found to be 18 nucleotides for SREBP data, 10 for EGR1 and 13 for HNF4, we prepared positive training sets containing sequences of binding sites 18, 10, and 13 nucleotides in length. The negative training set consisted of randomly generated sequences with the same frequencies as in the positive set.

Table 1. Samples of nucleotide sequences, recognition accuracy of the 'Discovery' system and PWM for TFBS SREBP, EGR1 and HNF4. False positive rates at the stringent threshold are defined by the false negative rate equal to 50%.

TFBS	Power of TFBSs Set	Sequence length, nt	FP rate PWM	FP rate 'Discovery'

SREBP	38	18	4.70E-04	3.90E-04
EGR1	22	10	4.06E-03	2.39E-03
HNF4	30	13	2.14E-04	7.00E-05

During each jackknife iteration methods were trained on the data set of sequences leaving exactly one for the control. The trained methods were applied to the rest sequence, estimating the FP error; the control negative sample was randomly generated with the nucleotide frequencies as in the positive samples and contained 100 000 sequences. Totally the number of jackknife iterations in each case was equal to the number of sequences in the data sets. We arranged the control sites according to the FP rates. Figure 1 depicts the correlations between the true positive (TP) and FP rates for the HNF4 binding sites data. 'Discovery' outperforms PWM at any error cutoff.

The obtained result for the rest families of transcription factors (EGR1 and SREBP) is analogous to the HNF4, 'Discovery' favorably competes with the PWM. We produce the FP rates for both algorithms at the stringent threshold defined by the FN rate equal to 50% (table 1).

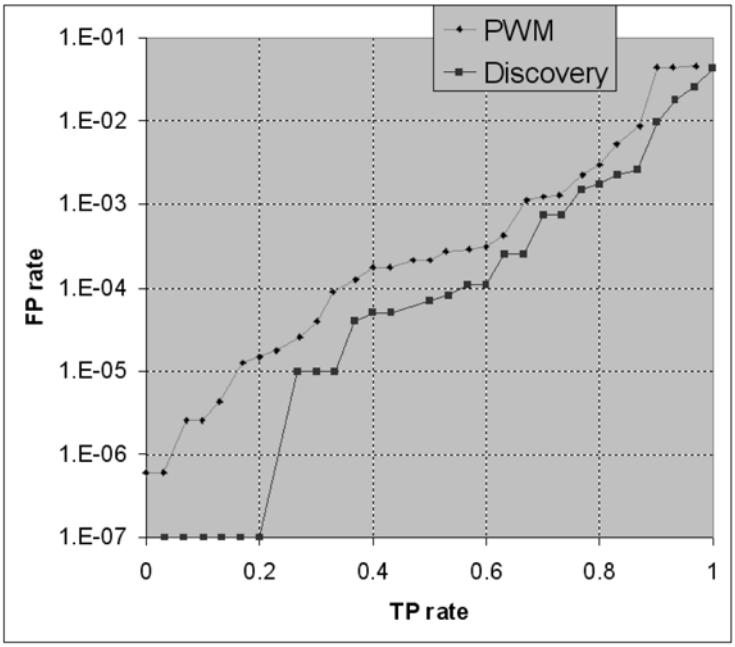

Fig. 1. Recognition performance of 'Discovery' system and PWM for HNF4 binding sites. The best six HNF4 binding sites are recognized at the FP rate lower then 1.E-07.

Acknowledgments. Siberian Branch of the Russian Academy of Sciences (integration project no. 115). The work was in part supported by the President of the Russian Federation, Scientific Schools grant 4413.2006.1.

References

1. Tan, A.C., Gilbert, D.: An empirical comparison of supervised machine learning techniques in bioinformatics. Proceedings of First Asia Pacific Bioinformatics Conference (APBC) (2003)
2. Stormo, G.D.: DNA binding sites: representation and discovery. Bioinformatics, 16, 16--23 (2000)
3. Benos, P.V., Bulyk, M.L., Stormo, G.D.: Additivity in protein-DNA interactions: how good an approximation is it? Nucleic Acids Res, 30, pp. 4442--4451 (2002)
4. Man, T.K., Stormo, G.D.: Non-independence of Mnt repressoroperator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay, Nucleic Acids Res., 29, pp. 2471--2478 (2001)
5. Barash, Y., Elidan, G., Friedman, F., Kaplan, T.: Modeling dependencies in protein-DNA binding sites. RECOMB, pp. 28--37 (2003)
6. Udalova, I.A., Mott, R., Field, D., Kwiatkowski, D.: Quantitative prediction of NF-kB DNA–protein interactions. Proc. Natl. Acad. Sci. USA, 99, pp. 8167--8172 (2002)
7. Vityaev, E., Data mining. Computational intelligence. Cognitive models. Novosibirsk State University, Novosibirsk, p. 293 (2006)
8. Vityaev, E.E.: Semantic approach to knowledge base creating. Semantic probabilistic inference of the best for prediction PROLOG-programs by a probability model of data Logic and Semantic Programming. Computational Systems, 146, Novosibirsk, pp.19--49 (1992)
9. Kendall, M.G., Stuart A.: The advanced theory of statistics, 4th ed., v.3. Charles Griffin & Co LTD, London (1977)
10. Kolchanov, N.A., Ignatieva, E.V., Ananko, E.A., Podkolodnaya, O.A., Stepanenko, I.L., Merkulova, T.I., Pozdnyakov, M.A., Podkolodny, N.L., Naumochkin, A.N., Romashchenko, A.G.: Transcription Regulatory Regions Database, (TRRD): its status in 2002. Nucleic Acid Res., 30, pp. 312--317 (2002)
11. Efron, B., Gong, G.: A leisurely look at the bootstrap the jackknife and resampling. American Statistician, 37, pp. 36--48 (1983)