SYSTEM OF STATISTICAL COMPARISON OF METHODS OF SEARCH OF CIS-ELEMENTS

[1,2]Cheremushkin E., [2]Dunaev A., [2]Murzin F.

[1] **Institute of Informatics Systems SB RAS, Novosibirsk**

[2] **Institute of Cytology and Genetics SB RAS, Novosibirsk**

There is a huge variety of theoretical and experimental researches for disclosing mechanisms of gene expression regulation. One of the basic concepts are transcription factors which represent ability of recognition of specific short fragments of DNA called binding sites. Therefore much attention is paid to binding sites recognition methods. We tested 2 popular sites search methods: MATCH and FOOTPRINT and shown the distribution of their positive prediction value and sensitivity.

The introduction

There is a huge variety of theoretical and experimental researches for disclosing mechanisms of gene expression regulation. One of the basic concepts playing a key role during a transcription, are transcription factors which represent regulatory ability of recognition of specific short fragments of DNA called binding sites. Therefore much attention is paid to binding sites recognition methods. Despite of a variety of approaches [1, 2, and 3], the problem of construction of exact methods of recognition of TF binding sites now cannot be considered completely solved. The reason is big variety of contextual, physical and chemical and conformational features of binding sites, mechanisms of interactions between sites and transcription factors, specificity of the surrounding context of binding sites, degrees of conservatism of nucleotide context in evolution. The recognition quality of binding sites search methods depends on type of sites. For example some types of sites are well distinguished by matrixes, but for others more complex methods are necessary. Also recognition quality directly depends on parameters used in recognition algorithms.

Method

The recognition quality can be analyzed from distribution of two parameters: positive prediction value $\alpha = 1 - FP$ and sensitivity $\beta = 1 - FN$. They correspond to false positive (*FP)* and false negative *(FN)* errors. Let $S = \{s_1,...,s_n\}$ are known experimental sites. $Q = \{q_1,...,q_m\}$ are the sites found by a certain method. Denote $s_i \approx q_j$ that site s_i matches q_j (it is recognized by q_j). Let $Q' = \{q_j \in Q \mid \exists s_i \in S, s_i \approx q_j\}$ is a set of sites that recognizes sites from S correctly. $S' = \{s_i \in S \mid \exists q_j \in Q, s_i \approx q_j\}$ is a set of sites that are correctly recognized. Then $\alpha = \frac{|Q'|}{|Q|}, \beta = \frac{|S'|}{|S|}$.

The basic problem is that is not all of the sites are discovered and annotated. Denote $T = \{t_1,...,t_k\}$ are unknown sites. Joint of known and unknown sites is all set of sites $S^* = S \cup T$. Let's transfer PPV and sensitivity with consideration of unknown sites $\alpha^* = \frac{|Q'^*|}{|Q|}, \beta^* = \frac{|S'^*|}{|S^*|}$

where $Q'^* = \{q_j \in Q \mid \exists s_i \in S^*, s_i \approx q_j\}$, $S'^* = \{s_i \in S^* \mid \exists q_j \in Q, s_i \approx q_j\}$. We can also rewrite $S'^* = S' \cup T'$ and $Q'^* = Q' \cup Q'_T$. Let unknown sites are k_T times more than known: $|T| = k_T |S|$. Let methods recognizes smaller percentage of unknown sites, than of known: $\frac{|Q'_T|}{|Q'|} = k_\alpha \frac{|T|}{|S|} = k_\alpha \cdot k_T$, $k_\alpha \in (0,1]$. Let also the amount of recognized unknown sites depends on amount of recognized

known sites similarly: $\frac{|T'|}{|S'|} = k_\beta \frac{|T|}{|S|} = k_\beta \cdot k_T$, $k_\beta \in (0,1]$. Then we have $\alpha^* = \alpha \cdot (1 + k_\alpha \cdot k_T)$, $\beta^* = \beta \cdot \frac{1 + k_\beta \cdot k_T}{1 + k_T}$.

Let's notice, that generally for various transcription factors there are various constants k_T, k_α, k_β. The constant k_T does not depend on observed search method. As sites from set T are not known, assume that other constants k_α and k_β also do not depend on a method. Then for the comparative analysis of methods it is enough to use distribution $\langle \alpha, \beta \rangle$, keeping in mind that it is not an absolute, but relative estimation of methods. Quality of a method of recognition varies for different factors, for different groups of sequences, as well as for parameters of a method. Parameters are not comparable for various groups of factors and groups of sequences, but they are comparable inside one group of factors and sequences.

Implementation

The system of comparison is implemented in GRESA DT system [4] as the module. For addition of a new method for testing it is enough to implement a function with use of the common mechanisms of calculation of statistics. If the method demands use of the additional data these data should be added in that way that for calculation of statistics in methods the same set of genes and transcription factors was used. During work of a method the statistics is stored in visualization friendly way.

Results

We tested 2 popular sites search methods: MATCH [1] and FOOTPRINT [5, 6.] MATCH is a method based on weight matrices. It is the most widely used method. FOOTPRINT - a method which is taking into account the evolutionary similarity of sequences. Firstly alignment of sequences is built, and then aligned sequences are scanned for binding sites which are recognized on both sequences in the same alignment block. Quality FOOTPRINT method highly depends on a homology of sequences taken. In the given method the data with a low threshold of homology is used. Genomic alignments from rVISTA Group [7] of human and mouse were used. These are alignments of the big DNA fragments (around 100000 bp) of mouse genome on complete human genome. Then only fragments containing known promoters were selected. For the analysis the data on genomic human sites were taken from TRANSFAC database [8]. Consider two distributions $\langle \alpha_m, \beta_m \rangle$ - for method MATCH and $\langle \alpha_f, \beta_f \rangle$ - for method FOOTPRINT.

The comparative analysis shows, that at parameters of methods at which $\alpha \in (0.0043, 0.0081)$ method MATCH gives the best prediction, and at other accessible values method FOOTPRINT is more appropriate.

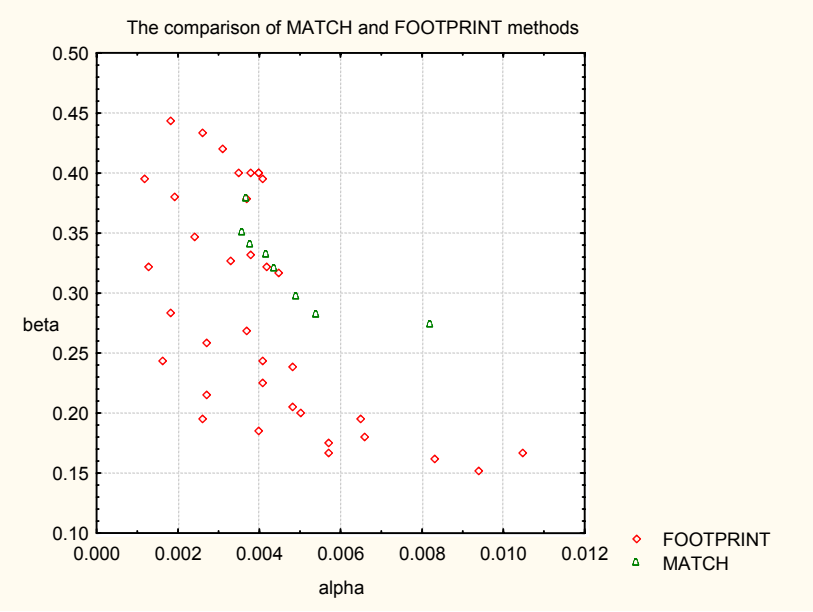

Fig 1. Distribution of positive prediction value $\alpha = 1 - FP$ *and sensitivity* $\beta = 1 - FN$ *for MATCH and FOOTPRINT methods of binding sites search. The more* α *and* β*, the quality of recognition is better. Various combinations of parameters are applied for various practical tasks, and give different combinations of* $\langle \alpha, \beta \rangle$*. Therefore all cumulative distribution is shown.*

References

1. MATCH: A tool for searching transcription factor binding sites in DNA sequences. Kel AE, Gossling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E.
2. A biophysical approach to transcription factor binding site discovery. Djordjevic M, Sengupta AM, Shraiman BI.
3. Bucher P. «Regulatory elements and expression profiles. » *Curr Opin Struct Biol.* 9,400-407 (1999)
4. GRESA Development Tools: The environment of development and testing of applications in the field of the DNA regulation sequences analysis. Evgeny Cheremushkin, Tatiana Konovalova (Ivanova), Maria Lobanova, Dmitry Tchekmenev, Eugene Beschastnov, Alexander Kel.
5. Blanchette M., Schwikowski B., Tompa M. «Algorithms for phylogenetic footprinting. » *J Comput Biol.* 9, 211-223 (2002)
6. Whole genome human/mouse phylogenetic footprinting of potential transcription regulatory signals. Cheremushkin E, Kel A. Pac Symp Biocomput. 2003; 291-302.
7. Loots G.G., Ovcharenko I., Pachter L., Dubchak I., Rubin E.M. rVista for comparative sequence-based discovery of functional transcription factor binding sites. *Genome Res.* 12, 832-839 (2002)
8. Wingender, E., Chen, X., Fricke, E., Geffers, R., Hehl, R., Liebich, I., Krull, M., Matys, V., Michael, H., Ohnhäuser, R., Prüß, M., Schacherer, F., Thiele, S. and Urbach, S. «The TRANSFAC system on gene expression regulation. » *Nucleic Acids Res.* 29, 281-283 (2001)