PHYLOGENETIC FOOTPRINT: A NEW METHOD FOR PROMOTER ALIGNMENT

[1]Cheremushkin E., [1,2]Kel A.

[1] Institute of Cytology and Genetics SB RAS, Novosibirsk

[2] BIOBASE GmbH, Wolfenbuettel

We have developed a new alignment method that takes into account similarity in distribution of potential binding sites. This method has been used effectively for promoter alignment and for revealing new potential binding sites for various transcription factors. We have developed a database of predicted potential TF binding sites in human genome by analyzing human/mouse conserved non-coding sequences (CNS).

Resume

Motivation: Phylogenetic Footprint is a new approach for revealing potential transcription factor binding sites in promoter sequences. The idea is based on an assumption that functional sites in promoters should evolve much slower then other regions that do not bear any conservative function. Therefore, potential transcription factor (TF) binding sites that are found in the evolutionally conservative regions of promoters have more chances to be considered as "real" sites. The most difficult step of the Phylogenetic Footprint is alignment of promoter sequences between different organisms (f.e. human and mouse). The conventional alignment methods often can not align promoters due to the high level of sequence variability.

Results: We have developed a new alignment method that takes into account similarity in distribution of potential binding sites. This method has been used effectively for promoter alignment and for revealing new potential binding sites for various transcription factors. We have developed a database of predicted potential TF binding sites in human genome by analyzing human/mouse conserved non-coding sequences (CNS).

Availability: http://compel.bionet.nsc.ru/FunSite/footprint/

Methods and algorithms

We developed a new method of alignment of regulatory sequences that include information about TF binding sites. To search the sites we apply position weight matrices (PWM) from TRANSFAC database (www.biobase.de) (Wingender et al., 2001). Every nucleotide in a sequence can potentially be belong to one or several TF binding sites. We estimate the probability $w_p(\bar{S},k)$ of k-th nucleotide of sequence $\bar{S}$ to be belong to a binding site of a factor T_p ($p \in [1,P]$) by using the following formulism:

$$w_p(\bar{S},k) = \sum_{j=k-L+1}^{k} F\left(\mathrm{X}_p(\bar{S},j)\right), \quad \vec{w}(\bar{S},k) = \langle w_1(\bar{S},k),\ldots,w_P(\bar{S},k)\rangle$$

where $\mathrm{X}_p(\bar{S},j)$ - score of p -th matrix at j -th position of sequence, L - length of $\bar{S}$.

$$F(x) = \frac{\exp(\lambda \cdot x)}{\exp(\lambda)}$$

The corresponding scores for different weight matrices can be seen in the Figure 1. We use different smoothing functions that weight differently the core positions of the sites (Fig. 1 b and c).

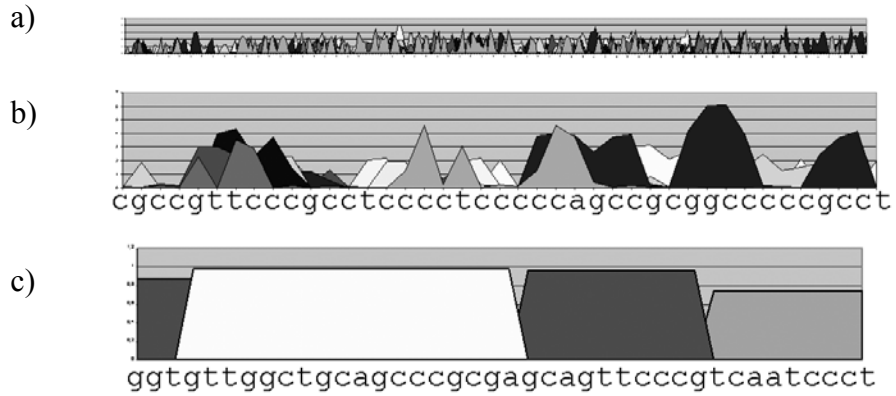

Figure 1. Distribution of nucleotide weights in a sequence. For each nucleotide in sequence we compute a vector of weights that reflects the probability of the nucleotide to be belong to a TF binding site. Different colors correspond to different TFs. (b,c) – usage of two different smoothing functions.

It is known that the library of weight matrices contains matrices that are similar to each other. These are different matrices for the same transcription factor or for the transcription factors that are very similar in their DNA binding signature. We consider a similarity matrix M that takes into account similarities between weight matrices.

$\vec{\varphi}(\overline{S},k)=\vec{w}(\overline{S},k)\cdot \mathrm{M}$, where M - $P\times Q$ similarity matrix. We will use $\vec{\varphi}(a)$ instead of $\vec{\varphi}(\overline{S},k)$, where $a\in\Sigma\times\Phi$ - sequence element, $\gamma(a)\in\Sigma$ - nucleotide for this element.

Alignment algorithm

We have developed an alignment algorithm for pair-wise and multiple alignment of nucleotide sequences (Cheremushkin & Kel, 2002). The algorithm is similar to the generally accepted Needleman-Wunsch dynamic programming algorithm. A major modification is made in the way of calculating the nucleotide substitution weights and gap penalty. The PWM scores were considered at every sequence positions in order to compute the corresponding substitution weights and gap penalty (see Fig. 2).

Figure 2. We consider alignment as a favorable one, if sites are aligned to each other

Gap penalty, while inserting gap in $\overline{S}^1$ between $k-1$ and k under position l in $\overline{S}^2$:

$$GAP(\overline{S}^1,\overline{S}^2,k,l)=\frac{G(\overline{S}^1,k)+R(\overline{S}^2,l)}{2},$$

Substitution weight:

$$SUB(\overline{S}^1,\overline{S}^2,k,l)=\mathrm{Z}(s_k^1,s_l^2),$$

where $G(\overline{S}^1,k)=\mathrm{Y}(s_{k-1}^1,s_k^1)$,

$$R(\bar{S}^2, l) = \frac{\mathrm{Y}(s_{l-1}^2, s_l^2) + \mathrm{Y}(s_l^2, s_{l+1}^2)}{2}$$

$$\mathrm{Y}(a,b) = \frac{C_{gap}}{N} + W_{gap} \cdot s_{gap}(a,b),$$

$$Z(a,b) = \frac{\Delta}{N} \cdot C_{sub} - W_{sub} \cdot \frac{\sum_{i=1}^{3} \lambda_i \cdot s_i(a,b)}{\sum_{i=1}^{3} \lambda_i}, \text{ for } a,b \in \Sigma \times \Phi, \text{ where,}$$

$$\Delta = \begin{cases} 1, \gamma(a) \neq \gamma(b) \\ 0, \gamma(a) = \gamma(b) \end{cases}, \; s_{gap}(a,b) = \begin{cases} (\vec{\varphi}(a) + \vec{\varphi}(b))^2, \gamma(a) = \gamma(b) \\ \vec{\varphi}(a)^2 + \vec{\varphi}(b)^2, \gamma(a) \neq \gamma(b) \end{cases},$$

$$s_1(a,b) = s_{gap}(a,b), \; s_2(a,b) = \begin{cases} 0, m > C_{\min} \\ (C_{\min} - m)/C_{\min}, m \leq C_{\min} \end{cases}, \text{ where } m = \min_i |\varphi_i(a) - \varphi_i(b)|,$$

$$s_3(a,b) = \max_i (\varphi_i(a) \cdot \varphi_i(b)),$$

$\gamma(a) \in \Sigma$ - nucleotide, $\vec{\varphi}(a) \in \Phi$ - matrices weight vector, C_{corr}, C_{gap}, W_{corr}, W_{gap}, λ_i - constants.

N - number of sequences.

In the Figure 3 we present an example of alignment of two sequences that is done by the algorithm. The score values of the aligned sequences are shown above and under the sequences correspondingly. One can see that the score picks are aligned to each other.

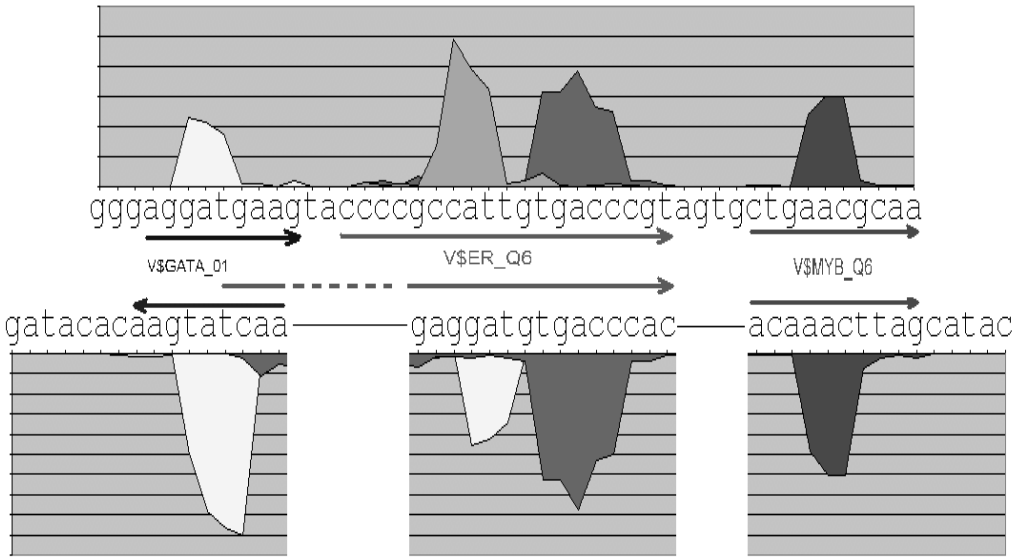

Figure 3. Example of alignment of a sequences. Graphical representation with the nucleotide weights in the alignment.

Implementation and results

The algorithm was implemented as a Java standalone program. It takes two sequences an input and align them. First it runs an Match algorithm that finds potential TF binding sites in the sequences. Specific collection of weight matrices with predefined cut-off values for every matrix can be specified by the user: taxon-dependent collection, tissue or function specific, minimizing false positive or false negative error. User can build his own profile with the help of TRANSPLORER program (http://www.biobase.de/pages/products/transplorer.html).

Testing of the alignment using a model of orthologous promoter sequences

In order to validate the developed alignment algorithm we have constructed a computer model of evolution of promoter sequences. An ancestor sequence of a length L is randomly created. In this sequence we implant N_{sites} binding sites with $N_{sites}+1$ spacers between them and on flanks. From this sequence we generate two descendant sequences by introducing a R_{spacer} random mutations (insertions, deletions and substitutions) in the spacer regions and R_{site} substitutions in the sites. We require that after each iteration all sites should remain "functional". For that, we check the PWM score for each of them and discard cases when the score drops below a certain cut-off (CO_{site}). Then, these two sequences are aligned and positions of the alignment blocks are compared with the sites that were originally implanted. In the case of misalignment of one of the sites we report a failure.

We have compared the developed alignment algorithms with the ClustalW by counting the percentage of failures. Our algorithm shows much better performance in finding correct alignment. With the homology of sequences equals to homology between human and mouse, the failure rate of our algorithms was about 0.1% whereas ClustalW gives approximately 2.5% of failures.

Phylogenetic footprint of human/mouse conserved non-coding sequences (CNS)

Evolutionary conserved non-coding regulatory sequences (CNS) could serve as good landmarks on genome to find functionally important promoters, enhancers or silencers (Duret & Bucher, 1997). Phylogenetic footprinting of CNS will help us to reveal TF binding sites and assign a regulatory function to the regulatory regions and to the adjacent genes. We use results of the Berkeley Genome Pipeline (http://pipeline.lbl.gov/) of the global comparison of human and mouse genomes. We have download the complete list of CNS and made the phylogenetic footprint of all of them. Two types of alignment were used. First, we used the original alignment that was done by the VISTA program (http://www-gsd.lbl.gov/vista/), and second, we made our own alignment using the developed algorithm. Phylogenetic footprint was done by the previously developed tool (http://compel.bionet.nsc.ru/FunSite/footprint/) that takes two or several aligned sequences, finds conservative binding sites and display them. Binding sites with the score acceding a predefined cut-off, for transcription factors that belong to the same family and that have overlapping location on the alignment are considered as the positive match at the phylogenetic footprint. The list of 17117 CNS of the total length of alignment 2418267 bp was analysed. We applied a set of 240 weight matrices from TRANSFAC rel. 5.3 with the cut-offs optimized to minimize the sum of false positive and false negative errors. Using VISTA alignments we found 54075 conservative TF binding sites. Using alignment by our own algorithm we found 58106 conservative TF binding sites. So, our algorithm that includes information about potential TF sites at the very early stage of analysis allows us to reveal 4031 more binding sites then using standard alignment algorithm. In the figure 4 one can see the comparison of the number of revealed sites using VISTA alignment versus our alignment.

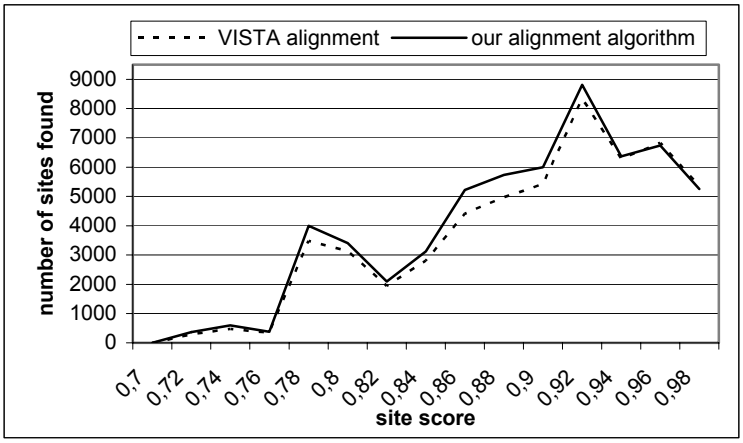

Fig.4. Comparison of the number of revealed sites using VISTA alignment versus our alignment algorithm. More sites with the score values from 0.78 to 0.92 can be revealed by our alignment algorithm.

It is interesting to observe that our alignment algorithm helps to reveal more sites with the score values from 0.78 to 0.92, which are the most functionally relevant sites. Low scoring sites (lower then 0.72) and pick scoring (higher then 0.92) are revealed in the same amount as using the VISTA alignment.

We have developed a database of predicted potential TF binding sites in human genome by analyzing the human/mouse CNS. Using this database user can retrieve all conservative sites for a selected chromosome or for a region at the chromosome and can visualize gene information for the nearest upstream and downstream genes, that can be targets for regulation through found TF binding sites. Using the developed database molecular biologists can plan their experiments for validation of found target genes and can make regulatory functional annotation of human and mouse genome.

Acknowledgments

The authors are indebted to Edgar Wingender for fruitful discussion of the results. Parts of this work was supported by Siberian Branch of Russian Academy of Sciences and by the grant of Volkswagen-Stiftung (I/75941).

Literature

1. *Cheremushkin E., Kel A.* (2002) PromoterFootprint: A new method for alignment of regulatory genomic sequences. Phylogenetic footprinting of TF binding sites. In Liliana Florea, Brian Walenz, Sridhar Hannenhalli (eds) Currents in Computational Molecular Biology 2002. RECOMB 2002, Washington D.C., pp. 40-41.
2. *Duret, L. & Bucher, P.* (1997). Searching for regulatory elements in human noncoding sequences. *Curr. Opin. Struct. Biol.* **7**, 399-406.
3. *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.* (2001) The TRANSFAC system on gene expression regulation. *Nucleic Acids Res***. 29**, 281-283.