Whole Genome Human/Mouse Phylogenetic Footprinting of Potential Transcription Regulatory Signals

E. Cheremushkin, A. Kel

Pacific Symposium on Biocomputing 8:291-302(2003)

WHOLE GENOME HUMAN/MOUSE PHYLOGENETIC FOOTPRINTING OF POTENTIAL TRANSCRIPTION REGULATORY SIGNALS

E. CHEREMUSHKIN[1]; A. KEL[1,2]

[2]*Institute of Cytology & Genetics SB RAN, 10 Lavrentyev pr., 630090, Novosibirsk, Russia;*

[2]*BIOBASE GmbH, Halchtersche Strasse 33, 38304 Wolfenbuettel, Germany.*

Phylogenetic footprinting is an efficient 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 footprinting 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. We have developed a new alignment method that takes into account similarity in distribution of potential binding sites (motif-based alignment). This method has been used effectively for promoter alignment and for revealing new potential binding sites for various transcription factors. We made a systematic phylogenetic footprinting of human/mouse conserved non-coding sequences (CNS). 60 thousand potential binding sites were revealed in human and mouse genomes. We have developed a database of the predicted potential TF binding sites.
Availability: http://compel.bionet.nsc.ru/FunSite/footprint/; www.gene-regulation.com/

1. Introduction

Genes in genomes of higher eukaryotic organisms are regulated mainly by the means of multiple regulatory proteins - transcription factors (TF), acting through specific regulatory sequences (TF binding sites) that are located usually in the proximity of the genes when constituting a promoter, or at more remote locations when being a part of an enhancer. The unique pattern of regulation of every gene in the multiplicity of different cellular, tissue-specific and developmental conditions can be fully understood through identifying all functional TF binding sites in the regulatory regions of genes. New techniques of large scale analysis such as ChiP on chip (Ren et al, 2002) provide means for extensive identification of TF binding sites. But the large number, combinatory nature of regulation as well as practically unlimited verity of cellular conditions renders unlikely an experimental genome-scale identification of TF binding sites. Computational identification of potential TF binding sites in gene regulatory regions become very important for gene prediction, functional characterization of newly discovered genes (Bucher, 1999; Werner, 1999; Wasserman & Fickett, 1998) as well as for understanding of regulatory circuits in the cell. There are a number of pattern-based as well as matrix based methods for prediction of potential TF binding sites (Quant et al., 1995; Goessling et al, 2001) but the false positive prediction rates are pretty high (Tronche et al, 1997).

Phylogenetic footprinting (Gumucio et al., 1996; Duret and Bucher, 1997) is a very effective approach for identification of regulatory elements, which relies on the cross-species sequence comparison. Phylogenetic footprinting is based on the assumption that the parts of gene regulatory regions that are liable to no or little variations in the course of evolution have important functions (Duret & Bucher, 1997). The essence of the method is the alignment of the regulatory regions of the orthologous genes and finding of the most highly conserved regions.

Though, it is clear that this method works as far as the pattern of regulation of the particular gene and their control mechanisms are conserved across species, which can be true for rather similar organisms only. Sites will stay conserved when they provide basic regulatory function of a gene which is conserved between considered species. Species-specific regulatory nuances will not be recognized by this method. On the other side, sequence comparison of a closely related species does not help to distinguish functionally conserved features against a background similarity of recently evolved sequences. Therefore, human and mouse are generally accepted as a reasonably similar as well as pretty distant species to perform efficient phylogenetic footprinting of their genomes (Mueller et al, 2002).

In the last years, the approach of phylogenetic footprinting was used several times to reveal potential TF binding sites in different genes: in upstream region of the beta-like and ε-globine genes (Gumucio et al, 1993; Gumucio et al, 1996), in COX5B gene that encodes subunit Vb of cytochrome c oxidase (Bachman et al, 1996). Recently, phylogenetic footprinting helps to reveal new E2F sites and proposes a regulation in cell cycle for a number of new genes (Kel et al, 2001). Conservative combinations of two binding sites (composite elements) were revealed in many T-cell specific genes (Kel et al., 1999). Using principles of phylogenetic footprinting a number of novel binding sites were revealed in regulatory regions of 502 genes associated with a variety of different human diseases (Levy et al., 2001).

A few computational implementation of the idea of the phylogenetic footprinting have been developed. These are: FunSiteFootprint (http://compel.bionet.nsc.ru/FunSite/footprint/; Cheremushkin & Kel, 2002), Consite (http://forkhead.cgr.ki.se/cgi-bin/consite, Wasserman et al., 2000), rVISTA (http://pga.lbl.gov/rvista.html, Loots et al, 2002), TRES (http://bioportal.bic.nus.edu.sg/tres/info.htm), FootPrint2.0 (http://bio.cs.washington.edu/software.html, Blanchette et al, 2002). The most critical step of the phylogenetic footprinting is alignment of regulatory sequences such as promoters and enhancers. The conventional alignment methods often can not align promoters due to the high level of sequence variability. We have developed a new alignment method that takes into account similarity in distribution of potential binding sites. We call this method: motif-based alignment. This method has been used effectively for promoter alignment and for revealing new potential binding sites for various transcription factors. We made a systematic phylogenetic footprinting of human/mouse conserved non-coding sequences (CNS) that were downloaded from the Berkeley Genome Pipeline site (http://pipeline.lbl.gov/) of the global comparison of human and mouse genomes. 60 thousand potential binding sites were revealed in human and mouse genomes. We have developed a database of the predicted potential TF binding sites.

This database as well as the program for motif-based alignment and phylogenetic footprinting are available at: (compel.bionet.nsc.ru/FunSite/footprint).

2 Method

2.1 Search for TF binding sites.

We developed a new method for alignment of regulatory sequences that includes information about TF binding sites. To search for 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 $w_p(\overline{S},k)$ of k-th nucleotide of a sequence $\overline{S}$ to be belong to a binding site of a factor T_p ($p \in [1,P]$):

$$w_p(\overline{S},k) = \alpha \times \sum_{j=k-L+1}^{k} \exp(\beta \times s_p(\overline{S},j)), \ \vec{w}(\overline{S},k) = \langle w_1(\overline{S},k) \ldots, w_P(\overline{S},k) \rangle$$

where $s_p(\overline{S},j)$ - score of p -th matrix at j -th position of sequence, L - length of $\overline{S}$, α and β are two normalization constants.

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 a and b). First smoothing function gives more weight to the core positions of the site, the second function gives similar weights to all positions of the site.

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. We use M to convert the probability to a new function:

$\vec{\varphi}(\overline{S},k) = \vec{w}(\overline{S},k) \cdot \mathrm{M}$, 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. The components of the vector $\vec{\varphi}(a)$ we will call TF belonging coefficients.

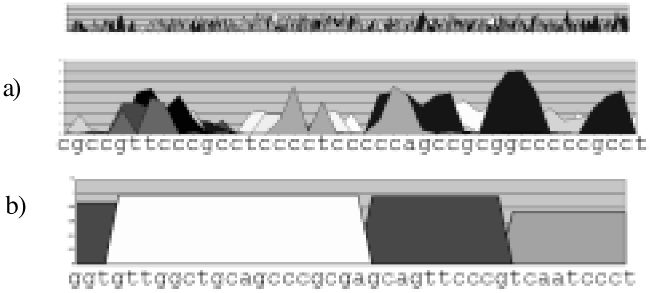

Figure 1. Distribution of TF belonging coefficients in a sequence. For each nucleotide in the sequence we compute a vector of weights that reflects the probability of the nucleotide to be belong to a TF binding site. Different gray colors correspond to different TFs. (a,b) – usage of two different smoothing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

2.2 Motif-based alignment algorithm.

We have developed an alignment algorithm for pair-wise and multiple alignment of nucleotide sequences. 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).

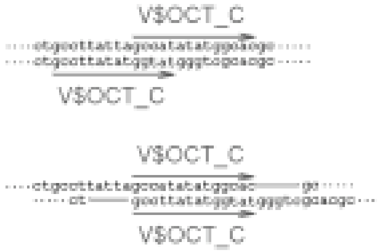

Figure 2. We consider alignment as a favorable one, if sites are aligned to each other.

For calculation of the gap penalty we construct a disjoining score function that can be applied to any two neighbor positions *a* and *b* in one nucleotide chain:

$$X_{gap}(a,b) = (\vec{\varphi}(a) + \vec{\varphi}(b))^2 .$$

This score estimates how similar are the two TF belonging vectors for these two positions. If both these positions have high belonging coefficients for the similar sets of transcription factors then the score X_{gap} is high. It means that most of the predicted TF binding sites in this region span over these two neighbor positions and disjoining of these positions by a gap will cause braking of these TF binding sites and is considered as an unfavorable event.

For N sequences in the alignment we use the following variant of the disjoining score function, where C_{gap}, W_{gap} are optimized constants:

$$\Upsilon(a,b) = C_{gap}/N + W_{gap} \cdot X_{gap}(a,b) \cdot$$

We use this function for calculation of the gap penalty, while inserting gap in $\bar{S}^1$ between $k-1$ and k under position l in $\bar{S}^2$:

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

where

$$G\left(\bar{S}^1, k\right) = \Upsilon\left(s_{k-1}^1, s_k^1\right),$$

which describes the score of disjoining of positions *k-1* and k due to a possible insertion in the sequence $\bar{S}^2$ during evolution;

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

which describes the score of simultaneous disjoining of positions *l-1* and l as well as l and *l+1* due to a possible deletion in the sequence $\bar{S}^1$ during evolution. Both deletions and insertions are considered to be equally probable that is why the gap penalty is a mean value between G and R.

Substitution weight for two aligned positions k and l in the sequences $\bar{S}^1$ and $\bar{S}^2$ correspondingly:

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

$$\mathrm{Z}(a,b) = \frac{\Delta}{N} \cdot C_{sub} - W_{sub} \cdot \sum_{i=1}^{3} \lambda_i \cdot E_i(a,b) \Big/ \sum_{i=1}^{3} \lambda_i, \text{ where, } \Delta = \begin{cases} 1, \gamma(a) \neq \gamma(b) \\ 0, \gamma(a) = \gamma(b) \end{cases},$$

$$E_1(a,b) = \begin{cases} \left(\vec{\varphi}(a) + \vec{\varphi}(b)\right)^2, \gamma(a) = \gamma(b), \\ \vec{\varphi}(a)^2 + \vec{\varphi}(b)^2, \gamma(a) \neq \gamma(b) \end{cases} \quad E_2(a,b) = \max_i \left(\varphi_i(a) \cdot \varphi_i(b)\right),$$

$$E_2(a,b) = \begin{cases} 0, m > C_{\min} \\ \left(C_{\min} - m\right)/C_{\min}, m \leq C_{\min} \end{cases}, \text{ where } m = \min_i \left|\varphi_i(a) - \varphi_i(b)\right| \cdot$$

$\gamma(a) \in \Sigma$ - nucleotide, C_{sub}, C_{gap}, W_{sub}, W_{gap}, λ_i - constants.

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

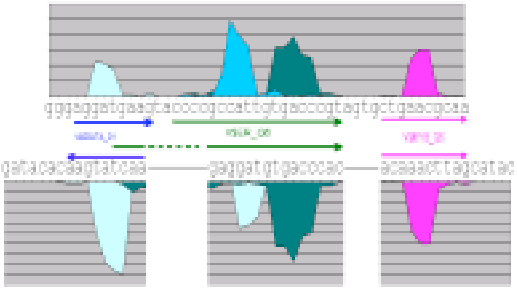

Figure 3. Example of alignment of a sequences. Graphical representation of the TF belonging coefficients.

3 Implementation and results

3.1 Implementation and availability.

The motif-based alignment algorithm was implemented as a Java standalone program. It takes two sequences as input and align them. First it runs the Match program (Goessling et al., 2001) 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).

3.2 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 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 motif-based 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. It is worth mentioning here that both alignment algorithms were always reporting a variant of alignment. Although, in the case of the failure the ClustalW alignments does not match the correct implanted sites but often matches wrong sites that were not implanted.

3.3 Phylogenetic footprinitng 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 footprinting of all of them. Two types of alignment were used. First, we made an alignment using ClustalW, and second, using the developed motif-based alignment algorithm. An example of VISTA large scale analysis of the PTH gene located in the human chromosome 11 is presented in the Fig. 4. 3 conservative non/coding sequences

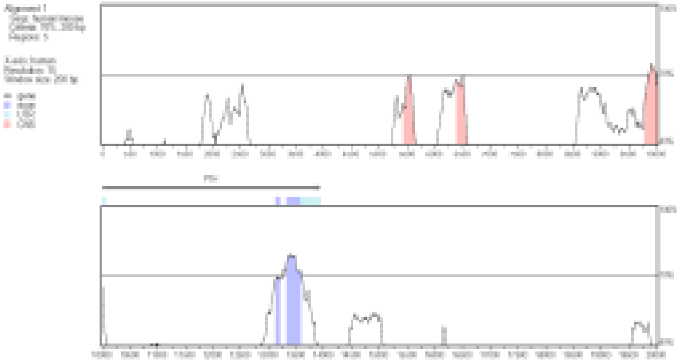

Figure 4. An example of large scale sequence comparison between orthologous PTH genes from human and mouse. The sequence comparison was done by the VISTA tool (http://www-gsd.lbl.gov/VISTA/). The program makes a global alignment and presents the result in a form of plot showing the percent of homology in the 200nt window sliding along the alignment. The position 10000 corresponds to the start of transcription of the human PTH gene. Regions of the highest homology are located in the coding regions of the second and third exons. 3 conservative non-coding sequences (homology > 70%) were found in the 5' sequence of the gene: CNS1 (5431-5639), CNS2 (6395-6599), CNS3 (9772-9999).

(CNS) are marked at the homology profile.

Two examples of detailed phylogenetic fotprinting are shown at the Figures 5 and 6. We can confirm most of the experimentally known sites. In addition a number of new sites are found.

```
                        <============V$SP1_Q6(0.89)
          ===>V$NFAT_Q6(0.82)                          VDRE
pth_human_ACTTAAGAAGAGTGTGCACCGCCCAATGGGTGTGTGTATGTGCTGCTTTGAACCTATAGT   -118
                        <============V$SP1_Q6(0.86)
          ===>V$NFAT_Q6(0.82)
pth_mouse_ACTTTAGAGGAGTGGGCACCACCCCATGAGGGTATGT---GGCTGTTCTGATCCTGTGAT    -74
           **** *** ***** ***** *** *** * ** ***    **** * *** *** *  *
======================================================================
                          <===============V$T3R_01(0.83)
=========.V$NF1_Q6(0.81)
                             ==========>V$AP1_Q2(0.91)
                             ===========>V$CREB_Q4(0.86)
                                         ===========>V$SRY_02(0.96)
                                 CREB        =========>V$TATA_C(0.93)
pth_human_TGAGATCCAGAGAATTGGGAGTGACATCATCTGTAACAATAAAAGAGCCTCTCTTGGTAA    -58
                          <===============V$T3R_01(0.84)
=========.V$NF1_Q6(0.81)
                             ==========>V$AP1_Q2(0.91)
                             ===========>V$CREB_Q4(0.86)
                                         ===========>V$SRY_02(0.97)
                                             =========>V$TATA_C(0.92)
pth_mouse_TGAGAGCCAGAGAACCAGGAGTGACATCATCCTTAACAATAAAATA-CTCCTCTTGGTGA    -15
           ***** ********   **************  *********** * *  ******** *
=======================================================================
          ========>V$NF1_Q6(0.81)
                      <=============V$OCT1_06(0.86)

<=======.V$MYOD_Q6(0.93)
pth_human_GCAGAAGACCTATATATAAAAGTCACCATTTAAGGGGTCTGCAGTCCAATTCATCAGTTG     +2
          ========>V$NF1_Q6(0.81)
                      <=============V$OCT1_06(0.92)

<=======.V$MYOD_Q6(0.93)
pth_mouse_GCAAAAAGCCTGCATATGAAACTCAGACTTGAAGAA--CTGCAGTCCAGTTCATCAGCTG    +43
           *** **  ***  **** *** ***   ** ***    ********** ******** **
======================================================================
          ==V$MYOD_Q6(0.93)
                             ==  =======>V$AP4_Q5(0.84)
pth_human_TCT---TTAGT-----TTACT-CAGCATCAGCTACTAACATACCTGAACGAAGATCTTGT    +53
                              =========>V$AP4_Q5(0.80)
          ==V$MYOD_Q6(0.93)
pth_mouse_TCTGGTTTACTCCAGCTTACTACAGCATCAGTTTGTG-CATCCCCGAAGGATCCCCT--T   +106
           ***   *** *     ***** ********* *  *  *** ** *** **    **  *
======================================================================
               <===========V$SRY_02(0.87)
pth_human_TCTAAGACATTGTATGGTAAG
               <===========V$SRY_02(0.88)
pth_mouse_TGAGAGTCATTGTATGGTAAG
           *   ** **************
```

Figure 5. The result of applying of the phylogenetic footprinting tool to the proximal part of the human/mouse PTH CNS3. The beginning of the fist exon in both sequences is marked by a shadow. Two sites are marked (CREB and VDRE) that have been described previously for this region in human PTH gene. Conservative sites found by this analysis are shown by the arrows above each sequence. A number of new potential sites is revealed: CREB site, OCT site near start of transcription, a couple of NF-AT sites, SP-1 sites, GATA sites and some others.

Phylogenetic footprinting 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 exceeding 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 of the phylogenetic footprinting.

```
                 <============V$CETS1P54_02(0.95)
                 <===========V$NFAT_Q6(0.97)
                             <============V$E2F_1Q6(0.93)
                              E2F-4  (R08679)     c-Ets-2  (R04343)
human    CTCTAGCCGCCCTTTCCTCTTTCTTTCGCGCTCTAGCCACCCGGGAAGGCCTGCCCA-GC    -93
                 <============V$CETS1P54_02(0.96)
                 <===========V$NFAT_Q6(0.95)
                             <============V$E2F_1Q6(0.94)
mouse    CTCTGGCCGCGGTTTCCGCTCCCTTTCGCGCTCTGCGCTCCCAGGCAGCCCCGCCCAAGC    -91
         **** *****  ***** **  ************   * *** ** ** ** ***** **
=====================================================================
                <================V$NFY_01(1.00)    =========>V$GATA1_01(0.99)
                  =========>V$GATA1_01(0.98)       <==========V$NFY_Q6(0.96)
                     <===========V$CAAT_01(1.00)
                      CAAT-box  (R04344)               CAAT-box  (R12441)
human    GTAGCTGGGCTCTGATTGGCTGCTTTGAAAGTCTACGGGCTACCCGATTGGTGAATCCGG    -33
                <================V$NFY_01(1.00)    =========>V$GATA1_01(0.98)
                  =========>V$GATA1_01(0.98)        <==========V$NFY_Q6(0.92)
                     <===========V$CAAT_01(0.99)
mouse    GTAGCTCAGCTCTGATTGGCTCCTTTGAACGTCTACGTGCAATCGGATTGGCGGATCCCG    -31
         ******  ************* ******* ******* ** * * ****** * **** *
=====================================================================
                       <=========V$MYB_Q6(0.95)
                              =========>V$MYB_Q6(0.97)
               <==========V$E2F_2Q6 _(0.94)
                      R  box  (p55)R04345
human    GGCCCTTTAGCGCGGTGAGTTTGAAACTGCTCG   +1
                       <=========V$MYB_Q6(0.96)
                              =========>V$MYB_Q6(0.97)
               <==========V$E2F_2Q6 _(0.83)
mouse    GGAGCTTTACCGCGGCGAGTTTGAAACTGCT--   +1
         **  ***** ***** ***************
```

Figure 6. The result of phylogenetic footprinting of the promoter regions of human /mouse cdc2 gene. 5 sites annotated in TRANSFAC are marked. 3 of them are confirmed by the phylogenetic footprinting. On the bases of this analysis we can propose the structure of R box: it contains E2F and Myb sites. c-Ets site in human promoter is not conserved in mouse, although we revealed a new conserved c-Ets site 20bp upstream.

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 ClustalW alignments we found 54075 conservative TF binding sites. Using the motif-based alignment 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 other alignment algorithm. In the figure 7 one can see the comparison of the number of revealed sites using ClustalW alignment versus our motif-based alignment.

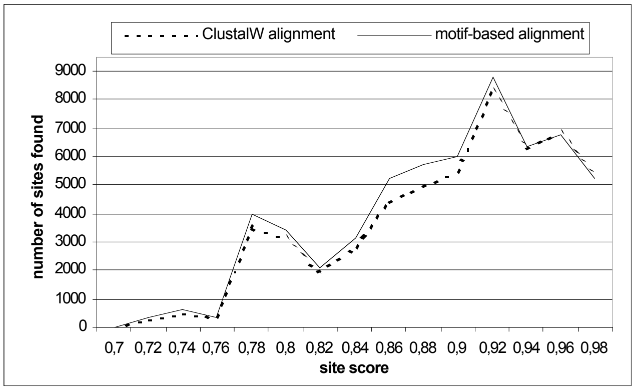

Figure 7. Comparison of the number of revealed sites using ClustalW alignment versus motif-based alignment algorithm. More sites with the score values from 0.78 to 0.92 can be revealed by the motif-based alignment algorithm.

It is interesting to observe that motif-based 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 peak scoring (higher then 0.92) are revealed in the same amount as using the ClustalW alignment.

We made a comparison between two alignment algorithms whether they identify the same sites. The comparison shows that only about 80% of sites are found the same (data not shown). Due to the higher failure rate for ClustalW to reveal the correct sites (see the simulation results) we can assert that motif-based alignment not only finds more genomic sites but all sites reported by this alignment are expected to be more often correct then using the ClustalW algorithm.

An example of comparison of results of two alignments done by ClustalW and motif-based alignment algorithm is presented in the Fig. 8. One can see that our algorithm enables revealing an additional conservative potential TF binding site that was missed by ClustalW.

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.

```
a)

GGTAT--------------AATTCTGTATT--------TGTTAAAA--------------    1274
GATTTTTTTTTAAAATATAAACTCAGTATTATCGATTATGTCAAAAATTTCTAGGTGGAC    1557
*** *              ** ** *****        *** ****


b)
                                  ============>V$CEBP_01(0.97)
GAT------------GGTATA------ATT-CTGTATTTGTTAAAA--------------    1274
                                  ============>V$CEBP_01(0.88)
GATTTTTTTTTAAAATATAAACTCAGTATTATCGATTATGTCAAAAATTTCTAGGTGGAC    1557
***              ** *      ***   *  * *** ****
```

Figure 8 An example of two alignments done by a) ClustalW and b) motif-based alignment algorithm. First alignment is more optimal (it have 20 matching positions vs 18 in the second alignment), whereas second alignment allows to reveal a conservative C/EBP site.

Acknowledgments

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

References

1. Bachman N.J., Yang T.L., Dasen J.S., Ernst R.E., Lomax M.I. "Phylogenetic footprinting of the human cytochrome c oxidase subunit VB promoter." *Arch Biochem Biophys* **333**, 152-162 (1996)
2. Blanchette M., Schwikowski B., Tompa M. "Algorithms for phylogenetic footprinting." *J Comput Biol.* **9**, 211-223 (2002)
3. Bucher P. "Regulatory elements and expression profiles." *Curr Opin Struct Biol.* **9**,400-407 (1999)
4. Cheremushkin E., Kel A. "PromoterFootprint: A new method for alignment of regulatory genomic sequences. Phylogenetic footprinting of TF binding sites." In Currents in Computational Molecular Biology (edited by L.Florea, B.Walenz, S. Hannenhalli), 40-41 (2002)
5. Duret, L. & Bucher, P. "Searching for regulatory elements in human noncoding sequences." *Curr. Opin. Struct. Biol.* **7**, 399-406 (1997)

6. Goessling E., Kel-Margoulis O.V., Kel A.E. and Wingender E. "MATCH^{TM} - a tool for searching transcription factor binding sites in DNA sequences. Application for the analysis of human chromosomes." In: *Proceedings of the German Conference on Bioinformatics (GCB2001)*, October 7-10, 2001, Braunschweig, pp.158-160 (2001)

7. Gumucio, D.L., Shelton, D.A., Zhu, W., Millinoff, D., Gray, T., Bock, J.H., Slightom, J.L., Goodman, M. "Evolutionary strategies for the elucidation of cis and trans factors that regulate the developmental switching programs of the beta-like globin genes." *Mol. Phylogenet. Evol.* 5, 18-32 (1996)

8. Gumucio D.L., Shelton D.A., Bailey W.J., Slightom J.L., Goodman M. "Phylogenetic footprinting reveals unexpected complexity in trans factor binding upstream from the epsilon-globin gene." *Proc Natl Acad Sci U S A.* **90**, 6018-6022 (1993)

9. Kel, A., Kel-Margoulis, O., Babenko, V., Wingender, E. " Recognition of NFATp/AP-1 Composite Elements within Genes Induced upon the Activation of Immune Cells" *J. Mol. Biol.* **288** , 353-376 (1999)

10. Kel A.E, Kel-Margoulis O.V., Farnham P.J., Bartley S.M., Wingender E., and Zhang M.Q. "Computer-assisted identification of cell cycle-related genes - new targets for E2F transcription factors." *J. Mol. Biol.* **309** , 99 – 120 (2001)

11. Loots G.G., Ovcharenko I., Pachter L., Dubchak I., Rubin E.M. *Genome Res*. **12**, 832-839 (2002)

12. Levy S., Hannenhalli S., Workman C. *Bioinformatics* **17**, 871-877 (2001)

13. Quandt, K., Frech, K., Karas, H., Wingender, E., and Werner, T. *Nucleic Acids Res*., **23**, 4878-4884 (1995)

14. Ren B., Cam H., Takahashi Y., Volkert T., Terragni J., Young R.A., Dynlacht B.D. "E2F integrates cell cycle progression with DNA repair, replication, and G(2)/M checkpoints." *Genes Dev.* **15;16**, 245-256 (2002)

15. Tronche, F., Ringeisen, F., Blumenfeld, M., Yaniv, M. & Pontoglio, M. "Analysis of the distribution of binding sites for a tissue-specific transcription factor in the vertebrate genome." *J. Mol. Biol.* **266**, 231-245 (2002)

16. Muller F., Blader P., Strahle U. *Bioessays* **24**, 564-572 (2002)

17. Wasserman W.W., Palumbo M., Thompson W., Fickett J.W., Lawrence C.E. "Human-mouse genome comparisons to locate regulatory sites." *Nat Genet.* **26**, 225-228 (2000)

18. Wasserman, W. W., Fickett, J. W. "Identification of regulatory regions which confer muscle-specific gene expression." *J. Mol. Biol.* **278** , 167-181 (1998)

19. Werner T. "Models for prediction and recognition of eukaryotic promoters." *Mamm. Genome* **10**, 168-175 (1999)

20. 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)