

Minireview

## Of mice and men: phylogenetic footprinting aids the discovery of regulatory elements

Zhaolei Zhang and Mark Gerstein

Address: Department of Molecular Biophysics and Biochemistry, Yale University, 266 Whitney Avenue, New Haven, CT 06520-8114, USA.

Correspondence: Mark Gerstein. E-mail: Mark.Gerstein@yale.edu

Published: 6 June 2003

*Journal of Biology* 2003, **2**:11

The electronic version of this article is the complete one and can be found online at <http://jbiol.com/content/2/2/11>

© 2003 BioMed Central Ltd

### Abstract

Phylogenetic footprinting is an approach to finding functionally important sequences in the genome that relies on detecting their high degrees of conservation across different species. A new study shows how much it improves the prediction of gene-regulatory elements in the human genome.

It has been a great challenge for biologists to understand the complicated and often myriad mechanisms of gene regulation. The recent success of genome sequencing projects [1,2], combined with very effective gene-prediction algorithms, has generated abundant gene sequences, but our understanding of gene regulation has remained very limited. In human and other higher eukaryotes, gene expression is modulated by the binding of various transcription factors onto *cis*-regulatory regions of a gene. Binding of different combinations of transcription factors may result in a gene being expressed in different tissue types or at different developmental stages. To fully understand a gene's function, therefore, it is essential to identify the transcription factors that regulate the gene and the corresponding transcription-factor-binding sites (TFBSs) within the DNA sequence. Traditionally, these regulatory sites were determined by labor-intensive wet-lab techniques such as DNase footprinting or gel-shift assays [3]; several online databases, such as TRRD, COMPEL and TRANSFAC [4,5] have been constructed to store experimentally determined TFBSs. Now, Lenhard and colleagues [6] describe a new addition to the toolkit for TFBS prediction.

In recent years, various computational methods have been developed to model and predict gene-regulatory elements. But predicting TFBSs has proved to be much harder than predicting genes, the intrinsic difficulty being that TFBSs are in general very short and often degenerate in sequence. Most TFBSs are short sequences of 6-12 base-pairs located in the non-coding regions of a gene, most often in the 5' flanking region but sometimes in the 3' region or even introns. Only between four and six bases within each TFBS are fully conserved, however, with the other positions being highly variable from gene to gene. As a result, TFBSs are often modeled using position-specific weight matrices (PWMs) [7], which in essence summarize the relative frequencies of each of the four nucleotides at each position. Figure 1 shows an example of such a matrix, for the human transcription factor GATA-1, from the widely used TRANSFAC database [5].

Given a PWM and a reliable scoring function, one can scan genomic DNA sequences and identify potential TFBSs. But because TFBSs are highly degenerate, the majority of predicted sites are 'false positives' that have no biological

<b>TRANSFAC accession number:</b> M00127					
<b>TRANSFAC identifier:</b> VSGATA1_03					
<b>Name:</b> GATA-1					
<b>Description:</b> GATA-binding factor 1					
Position	A	C	G	T	Consensus sequence
1	4	1	2	0	R
2	1	1	3	2	N
3	1	2	4	0	S
4	2	2	2	1	N
5	3	0	2	2	D
6	0	0	12	0	G
7	12	0	0	0	A
8	0	0	0	12	T
9	12	0	0	0	A
10	8	1	3	0	A
11	1	4	4	3	N
12	3	4	3	2	N
13	3	1	7	1	G
14	2	4	4	2	N
<b>Statistical basis:</b>		12 selected binding sequences			

**Figure 1**

An example of a position-specific weight matrix (PWM) adapted from the TRANSFAC database [5]. The sequences that have been shown experimentally to bind to the human transcription factor GATA-1 have 14 positions, among which only positions 6-10 are fully conserved. Abbreviations: R, G or A (purine); N, any; S, G or C (strong); D, G or A or T. Twelve sequences were used to build this matrix.

significance [8]. Several strategies have therefore been developed to reduce the false-positive rate; these include combining predictions with gene-expression data [9] or using prior knowledge of gene co-regulation [10]. Another approach is to take advantage of the fact that genes are often regulated by multiple transcription factors, so potential TFBSs tend to be clustered or adjacent to each other [11]. Alternatively, some researchers have tried to create more precise and sensitive tools for local sequence alignment and pattern discovery [12,13].

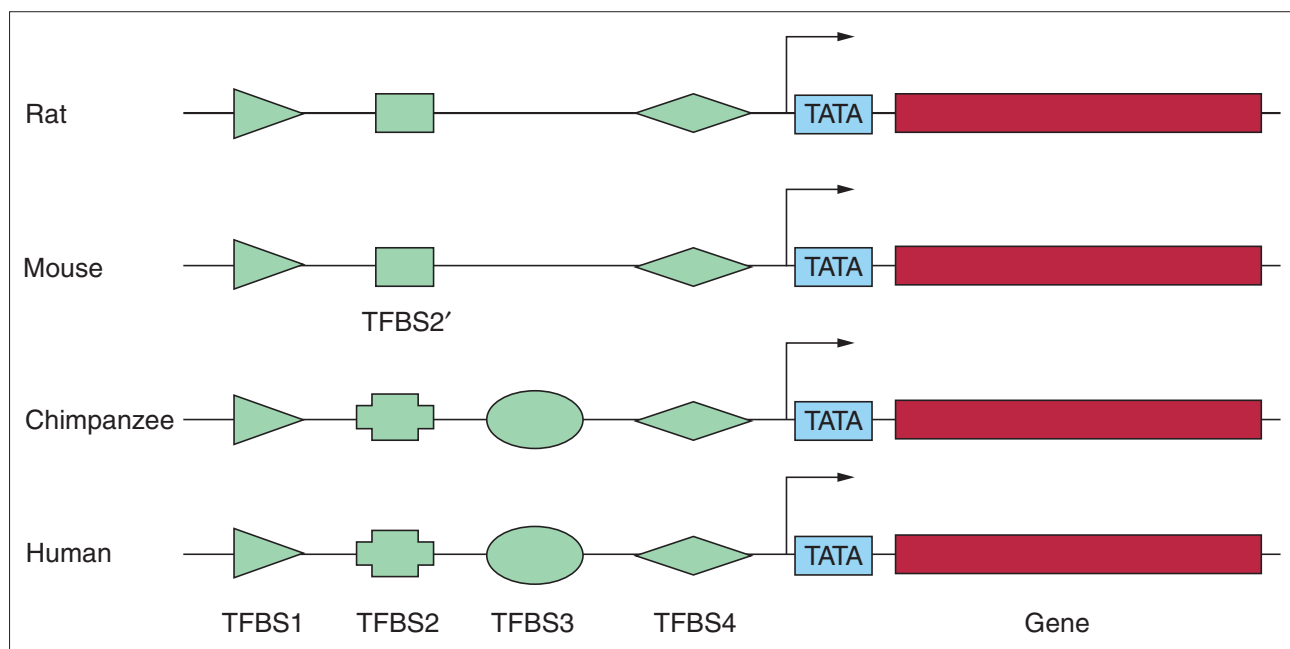
With the advance of genome sequencing projects, it has become obvious that comparing genomic sequences across species - 'comparative genomics' - is a very effective way to identify functionally important DNA sequences. At first comparative techniques were primarily applied to the coding regions of genomes, to identify genes or exon-intron boundaries [14]. More recently, such evolutionary approaches have become central to the efforts to predict gene-regulatory sites,

and the technique itself in this context has become known as 'phylogenetic footprinting' [15,16], a term inspired by the wet-lab technique of DNase footprinting. The reasoning behind the approach is that, just like coding sequences, regulatory elements are functionally important and are under evolutionary selection, so they should have evolved much more slowly than other non-coding sequences. Genome-wide sequence comparison and studies on individual genes have confirmed that regulatory elements are indeed conserved between related species [17-19]. Thus, if we align the non-coding regions of orthologous genes from two species that are sufficiently evolutionarily distant (but not too distant), we should be able to detect the conserved regulatory elements interspersed between the truly non-functional background sequences. This approach is illustrated schematically in Figure 2, in which a hypothetical human gene and its orthologs from mouse, rat and chimpanzee are shown together; alignment of the orthologous sequences reveals conserved TFBSs that are present in more than one species.

Phylogenetic footprinting was first performed by visually examining the alignment of orthologous sequences; then, automated computer programs were developed to assist the process. In this issue of *Journal of Biology*, Lenhard, Sandelin and colleagues describe their most recent success in predicting TFBSs by comparative genome analysis [6]. They also introduce an interactive, web-based computational platform, ConSite [20], which allows users to do their own phylogenetic footprinting.

The power of any TFBS prediction algorithm that uses PWMs depends on the quality of the matrix models that it uses, since the matrices represent an abstraction of experimentally verified TFBSs. Lenhard and colleagues [6] collected TFBSs from both *in vivo* and *in vitro* assays and used an improved motif discovery algorithm, ANN-Spec [21], to construct over 100 distinct and high-quality TFBS profile matrices. These comprehensive profiles were collected into an online database JASPAR [22], which is freely available to the scientific community. Users of ConSite can either provide an existing alignment of two orthologous sequences or input just the sequences alone and the program will generate the alignment. The program then scans the individual sequences for potential TFBSs and compares the potential sites between the aligned sequences. Only those conserved sites that are present in both sequences and also, more importantly, are located in equivalent positions in the two aligned sequences, are selected and reported in the output. The remainder of the sites, which are not conserved between the two species, are considered to be false positives and are eliminated.

This phylogenetic filtering procedure significantly improves the power of TFBS prediction, as is demonstrated by an



**Figure 2**

Using phylogenetic footprinting to detect conserved TFBSs. This schematic diagram shows a hypothetical human gene aligned with its orthologs from three other mammals. Cross-species sequence comparison reveals conserved TFBSs in each sequence. Sequence motifs of the same shape (colored in green) represent binding-sites of the same class of transcription factors. TFBS1 and TFBS4 are conserved in all four mammals; TFBS3 represents a newly acquired, primate-specific binding site. TFBS2 and TFBS2' represent orthologous regulatory sites that have diverged significantly between the primate and rodent lineages. Blue rectangles represent TATA boxes.

example described in detail in the article by Lenhard *et al.* [6]. The authors compared the human  $\beta$ -globin promoter sequence with the orthologous sequences from mouse and cow; this dramatically reduced the false-positive prediction of TFBSs and they were able to identify a previously documented regulatory site. The authors also studied a larger set of human-mouse gene pairs and compared the results predicted by ConSite with the previously verified regulatory sites. On average, phylogenetic footprinting improved the selectivity of TFBS prediction by 85% compared to using matrix models alone, and could detect the majority of verified sites. When compared with other available systems, ConSite has a flexible and easy-to-use web interface. Users of the website can choose to search for binding sites for any numbers of transcription factors or can even provide their own defined PWMs. The entire procedure and the output graphs can be modulated by many user-specified parameters such as the extent of required conservation (cut-off), and the length of sequence to search (window size).

It is becoming evident that comparative genome analysis is very powerful and will be of use not only for genome annotation but also as an adjunct to more traditional disciplines, such as molecular biology and genetics. Just like the

sequence-alignment programs that emerged in the early 1990s, ConSite and other similar programs [23,24] will prove very valuable and timely research tools for the scientific community. Many new research directions are currently being pursued in this area; for example, pair-wise sequence comparisons can be expanded to include multiple species and to make use of additional information, such as evolutionary distance and phylogenetic relationships [25]. More precise and effective sequence alignment programs have been created to handle genome-scale sequences [26,27]. In addition to the human-mouse comparisons, some researchers are also proposing cross-species comparison between human and other primates, which has been described as 'phylogenetic shadowing' [28]. This approach complements human-rodent comparisons and will detect primate-specific regulatory elements (see Figure 2). On the 'wet' experimental front, recent developments include microarray-based technologies such as 'ChIP-chip', which combines chromatin immunoprecipitation (ChIP) with analysis of the precipitated DNA on a microarray (chip), to detect TFBSs within a whole genome [29]. It can be imagined that, with the emergence of more mammalian genome sequences in the near future, we can finally identify all the gene regulatory elements in the human genome and use

them as a blueprint for understanding the mysteries of gene regulation.

## References

- Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, Fitzhugh W, et al.: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409**:860-921.
- Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, et al.: **Initial sequencing and comparative analysis of the mouse genome.** *Nature* 2002, **420**:520-562.
- Galas DJ, Schmitz A: **DNAse footprinting: a simple method for the detection of protein-DNA binding specificity.** *Nucleic Acids Res* 1978, **5**:3157-3170.
- Heinemeyer T, Wingender E, Reuter I, Hermjakob H, Kel AE, Kel OV, Ignatieva EV, Ananko EA, Podkolodnaya OA, Kolpakov FA, et al.: **Databases on transcriptional regulation: TRANSFAC, TRRD and COMPEL.** *Nucleic Acids Res* 1998, **26**:362-367.
- Wingender E, Chen X, Fricke E, Geffers R, Hehl R, Liebich I, Krull M, Matys V, Michael H, Ohnhauser R, et al.: **The TRANSFAC system on gene expression regulation.** *Nucleic Acids Res* 2001, **29**:281-283.
- Lenhard B, Sandelin A, Mendoza L, Engström P, Jareborg N, Wasserman WW: **Identification of conserved regulatory elements by comparative genome analysis.** *J Biol* 2003, **2**:13.
- Frech K, Herrmann G, Werner T: **Computer-assisted prediction, classification, and delimitation of protein binding sites in nucleic acids.** *Nucleic Acids Res* 1993, **21**:1655-1664.
- Fickett JW, Wasserman WW: **Discovery and modeling of transcriptional regulatory regions.** *Curr Opin Biotechnol* 2000, **11**:19-24.
- Zhu Z, Pilpel Y, Church GM: **Computational identification of transcription factor binding sites via a transcription-factor-centric clustering (TFCC) algorithm.** *J Mol Biol* 2002, **318**:71-81.
- Kielbasa SM, Korb J, Beule D, Schuchhardt J, Herzel H: **Combining frequency and positional information to predict transcription factor binding sites.** *Bioinformatics* 2001, **17**:1019-1026.
- Berman BP, Nibu Y, Pfeiffer BD, Tomancak P, Celniker SE, Levine M, Rubin GM, Eisen MB: **Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the *Drosophila* genome.** *Proc Natl Acad Sci USA* 2002, **99**:757-762.
- Quandt K, Frech K, Karas H, Wingender E, Werner T: **MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data.** *Nucleic Acids Res* 1995, **23**:4878-4884.
- Hertz GZ, Stormo GD: **Identifying DNA and protein patterns with statistically significant alignments of multiple sequences.** *Bioinformatics* 1999, **15**:563-577.
- Batzoglou S, Pachter L, Mesirov JP, Berger B, Lander ES: **Human and mouse gene structure: comparative analysis and application to exon prediction.** *Genome Res* 2000, **10**:950-958.
- Gumucio DL, Heilstedt-Williamson H, Gray TA, Tarle SA, Shelton DA, Tagle DA, Slightom JL, Goodman M, Collins FS: **Phylogenetic footprinting reveals a nuclear protein which binds to silencer sequences in the human gamma and epsilon globin genes.** *Mol Cell Biol* 1992, **12**:4919-4929.
- Duret L, Bucher P: **Searching for regulatory elements in human noncoding sequences.** *Curr Opin Struct Biol* 1997, **7**:399-406.
- Oeltjen JC, Malley TM, Muzny DM, Miller W, Gibbs RA, Belmont JW: **Large-scale comparative sequence analysis of the human and murine Bruton's tyrosine kinase loci reveals conserved regulatory domains.** *Genome Res* 1997, **7**:315-329.
- Hardison RC: **Conserved noncoding sequences are reliable guides to regulatory elements.** *Trends Genet* 2000, **16**:369-372.
- Wasserman WW, Palumbo M, Thompson W, Fickett JW, Lawrence CE: **Human-mouse genome comparisons to locate regulatory sites.** *Nat Genet* 2000, **26**:225-228.
- ConSite [<http://phylofoot.org/>]
- Workman CT, Stormo GD: **ANN-Spec: a method for discovering transcription factor binding sites with improved specificity.** *Pac Symp Biocomput* 2000:467-478.
- JASPAR [<http://forkhead.cgb.ki.se/JASPAR/>]
- Loots GG, Ovcharenko I, Pachter L, Dubchak I, Rubin EM: **rVista for comparative sequence-based discovery of functional transcription factor binding sites.** *Genome Res* 2002, **12**:832-839.
- Aerts S, Thijs G, Coessens B, Staes M, Moreau Y, De Moor B: **Toucan: deciphering the cis-regulatory logic of coregulated genes.** *Nucleic Acids Res* 2003, **31**:1753-1764.
- Blanchette M, Schwikowski B, Tompa M: **Algorithms for phylogenetic footprinting.** *J Comput Biol* 2002, **9**:211-223.
- Schwartz S, Zhang Z, Frazer KA, Smit A, Riemer C, Bouck J, Gibbs R, Hardison R, Miller W: **PipMaker - a web server for aligning two genomic DNA sequences.** *Genome Res* 2000, **10**:577-586.
- Bray N, Dubchak I, Pachter L: **AVID: A global alignment program.** *Genome Res* 2003, **13**:97-102.
- Boffelli D, McAuliffe J, Ovcharenko D, Lewis KD, Ovcharenko I, Pachter L, Rubin EM: **Phylogenetic shadowing of primate sequences to find functional regions of the human genome.** *Science* 2003, **299**:1391-1394.
- Horak CE, Mahajan MC, Luscombe NM, Gerstein M, Weissman SM, Snyder M: **GATA-I binding sites mapped in the beta-globin locus by using mammalian ChIP-chip analysis.** *Proc Natl Acad Sci USA* 2002, **99**:2924-2929.