W442–W446 *Nucleic Acids Research, 2005, Vol. 33, Web Server issue*
doi:10.1093/nar/gki420

FOOTER: a web tool for finding mammalian DNA regulatory regions using phylogenetic footprinting

David L. Corcoran[1,2], Eleanor Feingold[1,2] and Panayiotis V. Benos[1,3,4,*]

[1]Department of Human Genetics, Graduate School of Public Health, [2]Department of Biostatistics, Graduate School of Public Health and [3]Department of Computational Biology and [4]University of Pittsburgh Cancer Institute, School of Medicine, University of Pittsburgh, Pittsburgh, PA, USA

Received February 14, 2005; Revised and Accepted March 21, 2005

ABSTRACT

FOOTER is a newly developed algorithm that analyzes homologous mammalian promoter sequences in order to identify transcriptional DNA regulatory 'signals'. FOOTER uses prior knowledge about the binding site preferences of the transcription factors (TFs) in the form of position-specific scoring matrices (PSSMs). The PSSM models are generated from known mammalian binding sites from the TRANSFAC database. In a test set of 72 confirmed binding sites (most of them not present in TRANSFAC) of 19 TFs, it exhibited 83% sensitivity and 72% specificity. FOOTER is accessible over the web at http://biodev.hgen.pitt.edu/Footer/.

INTRODUCTION

Identifying DNA regulatory 'signals' in the promoter regions of genes is still one of the challenging problems in computational biology. Part of the problem is that the transcription factor binding sites (TFBSs) are usually short DNA sequences (6–20 bp) with high degree of degeneracy. Algorithms, such as AlignACE (1), ANN-Spec (2), Consensus (3), Co-bind (4) and MEME (5) to name a few, try to address the problem of low signal-to-noise ratio by looking at sets of genes considered to be enriched in one or more TFBS motifs [for a recent comparative study of these methods, we refer the reader to the excellent review of Tompa *et al*. (6)]. The various oligonucleotides in the test set that could be targets of a transcription factor (TF) are scored against some 'random' background. The methods primarily differ on the way they calculate the background and the objective function they try to optimize. These widely used methods fall under the category generally known as *de novo* 'pattern discovery' methods. A consequence of this methodology is that usually there is not much information about the TFs that bind to the identified patterns.

Although these methods are very useful, when knowledge about the binding preferences of a TF exists, there is no reason for one to ignore it. Thus, a second category of methods has been developed, the 'pattern identification' methods, that use existing information about the binding preferences of certain TFs and tries to identify the exact location of the motifs in a given promoter sequence. The same problem of low signal-to-noise ratio exists here, especially when one analyzes promoter sequences from complex eukaryotes, such as human, mouse or fly. The gene regulation in these organisms is usually more complex and the promoter length can extend to many kilobases from the transcription start site (TSS). In this case, evolutionary information comes to the rescue. Homologous promoter sequences can be compared in order to identify the evolutionary conserved DNA regulatory signals. This is commonly known as phylogenetic footprinting, a term first coined by Tagle *et al*. (7). Two of the most widely used algorithms for analyzing mammalian sequences are rVista (8) and ConSite (9). These algorithms are using the evolutionary conservation information in a fundamentally different way: rVista scans one of the promoters for high-scoring TFBSs and then uses position conservation to eliminate false positives. ConSite scans both promoters for motifs that score higher than a position-specific scoring matrix (PSSM) score threshold and then it uses a 'sliding window' approach to decide about the position conservation of the putative TFBSs.

We recently developed a novel phylogenetic footprinting algorithm, named FOOTER, which combines two statistics in order to score a pair of putative regulatory sites. Our method scans both promoter sequences and for each TF, it retains the top K scoring sites ('seed' TFBSs) in each promoter and then it compares all against all in order to find the best matching pairs according to the two criteria. This method has been shown to perform very well in a set of 72 confirmed TFBS of 19 TFs (S_N = 83%, S_P = 72%) (25).

*To whom correspondence should be addressed. Tel: +1 412 648 3315; Fax: +1 412 624 3020; Email: benos@pitt.edu

METHODS

Given two homologous promoter sequences and a number of putative motifs identified in each of them (by default FOOTER retains one top scoring motif per TF per 300 bp of promoter sequence), our method performs all pairwise comparisons of the motifs. A scoring scheme based on two statistics has been employed. The first statistic scores a pair of motifs according to their position conservation in the sequence. The second statistic scores the pair of motifs according to their agreement with the corresponding PSSM model(s). A PSSM model is the most commonly used way to represent the binding preferences of a TF (10). Typically, a set of aligned sequences is used to calculate a $4 \times L$ weight matrix (L is the length of the pattern). In each column, the weights correspond to the log-likelihood of the preferences of the TF to each of the four bases (sometimes normalized for the background), and in some cases it has been shown that they correspond to the actual binding energies of the protein–DNA interactions (11–13).

The two statistics FOOTER employs consist of the P-values of the observed data, under the null hypothesis that the two sites are unrelated. The position-related score is calculated using the following formula:

$$\mathrm{PF}_D = P(D_{XY} \leqslant d) = \frac{1}{N} + \sum_{k=1}^{d} \frac{2 \cdot (N-k)}{N^2}, \qquad \mathbf{1}$$

where D_{XY} is the random variable denoting the distance between two putative sites, d is the observed distance of the particular putative sites (measured from the 3′ closest conserved region boundary), N is the effective promoter length (i.e. the promoter length minus $L-1$, where L is the length of the pattern). Equation 1 calculates the tail probability that two high-scoring 'signals' will be found by chance at a distance d or less in the promoter with effective length N.

The PSSM-related score is calculated using the following formula:

$$\mathrm{PF}_S = P[(S+T) \leqslant (s+t) \,|\, \mathrm{M}_1, \mathrm{M}_2], \qquad \mathbf{2}$$

where M_1 and M_2 are the PSSM models for the two species; S and T are random variables following the models' score distributions; and s and t are the observed PSSM scores. The PF_S score is calculated using Gaussian approximation of mean and standard deviation estimated through random samplings from the PSSM model distributions. The results of the samplings are stored in each model. Similarly to PF_D, Equation 2 calculates the corresponding tail probability under the null hypothesis that the two high-scoring 'hits' are due to chance alone.

We have developed a novel scoring scheme that combines the above two statistics in a single similarity measure. The combined score, PF, consists of a weighted log-likelihood transformation:

$$\mathrm{PF} = -w_D \cdot \ln(\mathrm{PF}_D) - w_S \cdot \ln(\mathrm{PF}_S), \qquad \mathbf{3}$$

The weights w_D and w_S are positive numbers that sum to one (current default values: $w_D = 0.85$; $w_S = 0.15$). Summation of the logarithms is valid, since the individual PF scores are tail probabilities based on the null hypothesis that the human and the mouse patterns are not true binding sites, and hence the individual tail probabilities should be independent. Note that higher PF values correspond to higher probability that the human and mouse patterns are true sites. Since PF is the negative weighted average of the logarithms of P-values, exponentiation of $-$PF will return the weighted average P-value (WAP), which we use as a threshold cutoff in the server.

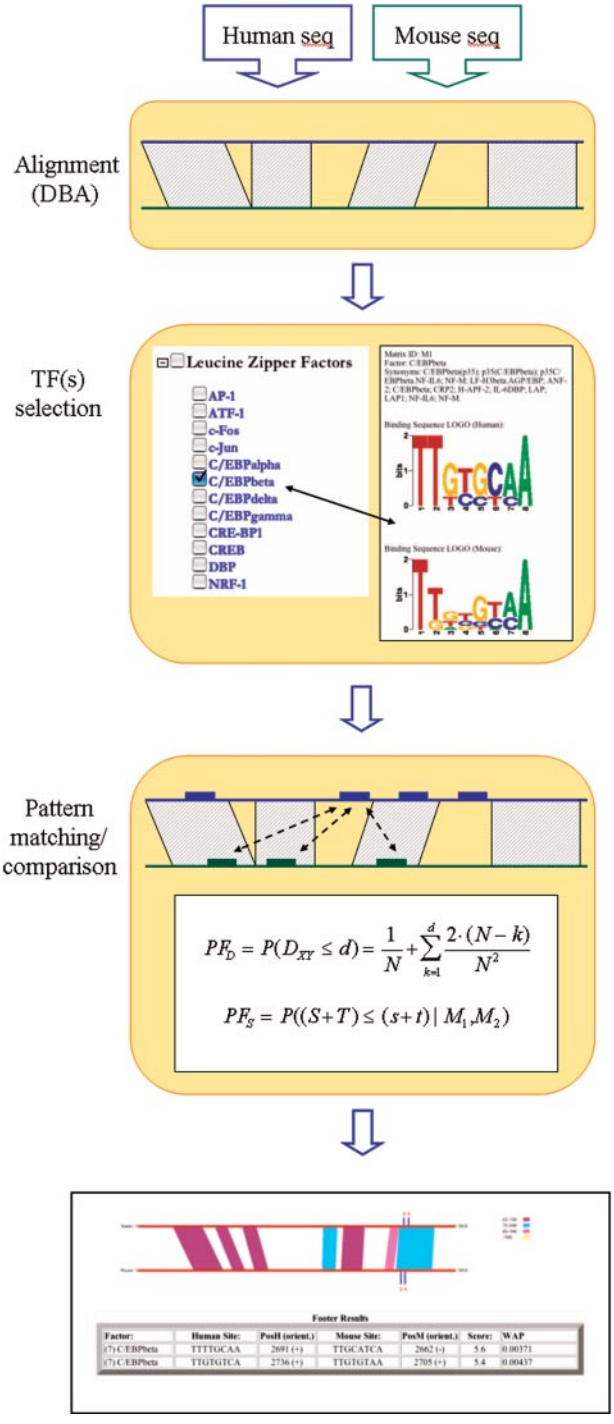
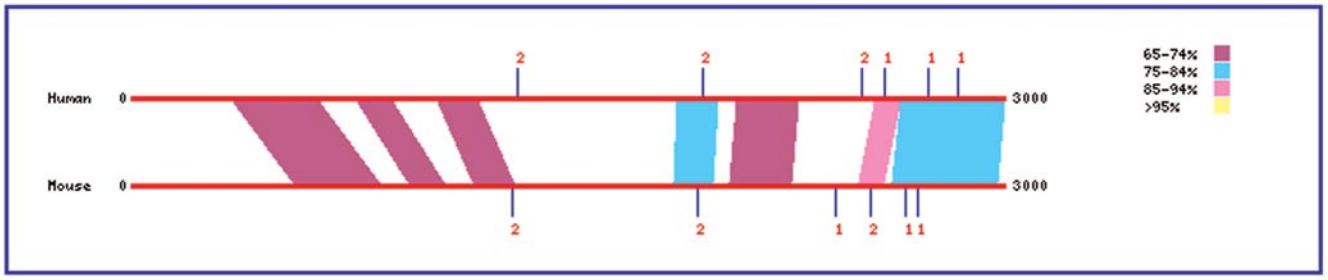


Figure 1. Flowchart of the execution of web-tool FOOTER. Protein or DNA sequences can be entered in the input. In the first case, a series of BLAST (15) searches will be performed to identify the homologous promoter sequences. The DNA sequences will be aligned and putative motifs will be compared pairwise as we describe in the text.

Footer Results



Footer Results

Factor	Human Site	PosH (orient.)	Mouse Site	PosM (orient.)	Score	WAP
(1) C/EBPbeta	TTGTGTCA	2736 (+)	TTGTGTAA	2705 (+)	8.8	0.00015
(1) C/EBPbeta	TTTTGCCA	2589 (+)	TTGCATCA	2662 (-)	8.3	0.00025
(1) C/EBPbeta	TGGCCCAA	2840 (-)	TGAGGTAA	2419 (+)	6.4	0.00162
(2) C/EBPalpha	TTGCCCAAG	2509 (-)	TTTGACAAC	2541 (+)	7.8	0.00040
(2) C/EBPalpha	ATGGCTAAT	1326 (-)	TTTTCCCAA	1310 (-)	7.5	0.00056
(2) C/EBPalpha	TTGCTTAAG	1967 (-)	GTTGCTCAA	1948 (-)	7.1	0.00079

Figure 2. Example of FOOTER output. The predicted sites are presented in table format and in the PNG formatted figure. The figure displays the alignment of the two promoter sequences, colour-coded by conservation percentage.

Algorithm

Once the promoter sequences have been specified and the PSSM models have been selected, FOOTER runs program DBA (14) to calculate the alignment between them (Figure 1). Then, a series of conserved and non-conserved regions are defined. Subsequently, the promoter sequences are scanned with each of the selected species- or mammalian-specific PSSM models and the top K 'seed' sites (user-defined parameter) are retained in each. These sites are analyzed pairwise, scored according to Equation 3 and matched using a greedy algorithm. The pairs that score above a user-specified WAP threshold are reported in the output. We should note here that since FOOTER compares pairwise all 'seed' sites in the two promoters (irrespectively of the DNA conservation in their surrounding region), it eliminates the need for a sliding window to identify 'conserved' sites.

Input data types

FOOTER accepts two types of input data (Figure 1): either a single protein sequence (human or mouse/rat) or two DNA sequences (presumed human and mouse promoters). The input sequence files should be in FASTA format. If the specified input sequence is a human (mouse) protein, FOOTER will employ a BLASTP search (15) in the human/mouse proteome to identify its homologous mouse (human) protein, then it will use these protein sequences to perform TBLASTN searches against Unigene database (16) to identify the longest mRNA sequences. Finally, these mRNA sequences will be used in BLASTN searches against the corresponding genomes in order to identify the locations of the TSSs and thus automatically retrieve the corresponding promoter sequences.

FOOTER parameters

The input parameters for FOOTER are the w_D and w_S weights (see Equation 3) that should sum to one (if not, the program will automatically adjust them proportionally); the WAP threshold (default value is 0.005, which corresponds to a FOOTER combined score of 7.6); and the number of seed sites that will be initially retained and analyzed (default: an average of one site per 300 bp). In the case that a protein sequence is specified in input (see above), the user should also specify the promoter length (upstream and downstream sequence from the TSS) to be analyzed.

Testing

FOOTER has been tested in a set of 72 confirmed TFBS of 19 TFs. FOOTER was able to predict correctly 60 of these binding sites (S_N = 83%) while it made an additional (unverified) 24 predictions. A table with the results mentioned above is provided at http://biodev.hgen.pitt.edu/Footer/webNAR05/Table.pdf. A more detailed description and extensive evaluation of the algorithm has been submitted for publication (25).

IMPLEMENTATION

FOOTER is written in Perl (CGI). For the graphical representation of the aligned promoters (PNG file), the PG package of Perl is used. The program uses program DBA (14) for the promoter alignment, which is currently its main performance bottleneck. DBA alignment time depends on the size of the promoter region, though it usually takes 45–75 s for a 3 kb promoter region. Once the two promoters have been aligned, FOOTER requires only a couple of seconds to identify the optimal matching patterns for a 3 kb region (on a Dell PowerEdge 2650 machine with 2.8 GHz dual-processor Xeon machine with HT technology and 2 GB of RAM). The time increases linearly with the number of TFs that it considers and exponentially with the number of seed patterns. With the default parameters, the complete search, including automatic identification of the promoter regions and alignment using DBA, does not usually require more than 3 min for a promoter length of 3 kb. FOOTER is available at http://biodev.hgen.pitt.edu/Footer/.

RESULTS

The final result of the program is a list of predicted sites in a table format (Figure 2). The results include the name of the TF, the sequence and position of the predicted TFBS in both the human and mouse promoters, the FOOTER calculated score (Equation 3) and the WAP value. This table can be copied into a spreadsheet program and analyzed further. In addition, FOOTER produces a PNG image with the alignment of the two sequences, color-coded by percent identity. The PNG image also displays all predicted sites. Finally, the results page provides links to the individual promoter sequences, the DBA alignment output and a summary of the program run, including all predicted sites (not just those above the cutoff).

Current limitations/future improvements

At the present stage, FOOTER has two limitations. One is the availability of PSSM models. We currently use models we constructed using TRANFAC (17) binding sites. In this way, we have calculated mammalian PSSM models for 127 TFs. With the development of high-throughput techniques for binding site identification, such as ChIP (18) and SELEX (19), construction of mammalian-specific matrix for many TFs will not be a problem. Recently, well-curated sets of binding sites have started to become publicly available (20). Nevertheless, we plan to add a feature to FOOTER that will allow for a user-defined PSSM model to be uploaded and used to scan the promoter sequences. We also plan to hyperlink the PNG image so that by moving the cursor over it, the user will receive information on various features of the predictions.

We have noticed that DBA (14) sometimes becomes slow in aligning long DNA sequences. For this purpose, we are currently exploring other algorithms (21,22) and strategies in order to further speed up FOOTER performance. For example, another way to speed the performance is to use pre-calculated alignments or databases of aligned promoter regions [e.g. (23,24)].

ACKNOWLEDGEMENTS

P.V.B. was partly supported by NSF grant MCB0316255. Funding to pay the Open Access publication charges for this article was provided by intramural funds of the Department of Computational Biology and the University of Pittsburg Cancer Institute, School of Medicine, University of Pittsburgh.

Conflict of interest statement. None declared.

REFERENCES

1. Hughes,J.D., Estep,P.W., Tavazoie,S. and Church,G.M. (2000) Computational identification of *cis*-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J. Mol. Biol.*, **296**, 1205–1214.
2. Workman,C.T. and Stormo,G.D. (2000) ANN-Spec: a method for discovering transcription factor binding sites with improved specificity. *Pac. Symp. Biocomput.*, 467–478.
3. Hertz,G.Z., Hartzell,G.W.,III and Stormo,G.D. (1990) Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Comput. Appl. Biosci.*, **6**, 81–92.
4. GuhaThakurta,D. and Stormo,G.D. (2001) Identifying target sites for cooperatively binding factors. *Bioinformatics*, **17**, 608–621.
5. Bailey,T.L. and Elkan,C. (1995) The value of prior knowledge in discovering motifs with MEME. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **3**, 21–29.
6. Tompa,M., Li,N., Bailey,T.L., Church,G.M., De Moor,B., Eskin,E., Favorov,A.V., Frith,M.C., Fu,Y., Kent,W.J. *et al.* (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.*, **23**, 137–144.
7. Tagle,D.A., Koop,B.F., Goodman,M., Slightom,J.L., Hess,D.L. and Jones,R.T. (1988) Embryonic epsilon and gamma globin genes of a prosimian primate (Galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *J. Mol. Biol.*, **203**, 439–455.
8. Loots,G.G., Ovcharenko,I., Pachter,L., Dubchak,I. and Rubin,E.M. (2002) rVista for comparative sequence-based discovery of functional transcription factor binding sites. *Genome Res.*, **12**, 832–839.
9. Lenhard,B., Sandelin,A., Mendoza,L., Engstrom,P., Jareborg,N. and Wasserman,W.W. (2003) Identification of conserved regulatory elements by comparative genome analysis. *J. Biol.*, **2**, 13.
10. Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.
11. Benos,P.V., Lapedes,A.S. and Stormo,G.D. (2002) Probabilistic code for DNA recognition by proteins of the EGR family. *J. Mol. Biol.*, **323**, 701–727.
12. Benos,P.V., Bulyk,M.L. and Stormo,G.D. (2002) Additivity in protein–DNA interactions: how good an approximation is it? *Nucleic Acids Res.*, **30**, 4442–4451.
13. Benos,P.V., Lapedes,A.S., Fields,D.S. and Stormo,G.D. (2001) SAMIE: statistical algorithm for modeling interaction energies. *Pac. Symp. Biocomput.*, **6**, 115–126.
14. Jareborg,N., Birney,E. and Durbin,R. (1999) Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. *Genome Res.*, **9**, 815–824.
15. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI–BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
16. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S., Helmberg,W. *et al.* (2005) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **33**, D39–D45.
17. Wingender,E. (2004) TRANSFAC, TRANSPATH and CYTOMER as starting points for an ontology of regulatory networks. *In Silico Biol.*, **4**, 55–61.
18. Orlando,V. (2000) Mapping chromosomal proteins *in vivo* by formaldehyde-crosslinked-chromatin immunoprecipitation. *Trends Biochem. Sci.*, **25**, 99–104.
19. Choo,Y. and Klug,A. (1994) Selection of DNA binding sites for zinc fingers using rationally randomized DNA reveals coded interactions. *Proc. Natl Acad. Sci. USA*, **91**, 11168–11172.

20. Sandelin,A., Alkema,W., Engstrom,P., Wasserman,W.W. and Lenhard,B. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res*., **32**, D91–D94.
21. Nix,D.A. and Eisen,M.B. (2005) GATA: a graphic alignment tool for comparative sequence analysis. *BMC Bioinformatics*, **6**, 9.
22. Brudno,M., Do,C.B., Cooper,G.M., Kim,M.F., Davydov,E., Green,E.D., Sidow,A. and Batzoglou,S. (2003) LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res*., **13**, 721–731.
23. Palaniswamy,S.K., Jin,V.X., Sun,H. and Davuluri,R.V. (2005) OMGProm: a database of orthologous mammalian gene promoters. *Bioinformatics*, **21**, 835–836.
24. Zhao,F., Xuan,Z., Liu,L. and Zhang,M.Q. (2005) TRED: a Transcriptional Regulatory Element Database and a platform for *in silico* gene regulation studies. *Nucleic Acids Res*., **33**, D103–D107.
25. Corcoran,D.L., Feingold,E., Dominick,J., Wright,M., Harnaha,J., Trucco,M., Giannoukakis,N. and Benos,P.V. (2005) FOOTER: a quantitative comparative genomics method for efficient recognition of cis-regulatory elements. *Genome Res*., in press.