W366–W368 *Nucleic Acids Research, 2006, Vol. 34, Web Server issue*
doi:10.1093/nar/gkl069

MicroFootPrinter: a tool for phylogenetic footprinting in prokaryotic genomes

Shane Neph and Martin Tompa*

Department of Computer Science and Engineering and Department of Genome Sciences,
University of Washington, Box 352350, Seattle, WA 98195-2350, USA

Received February 11, 2006; Revised February 19, 2006; Accepted March 1, 2006

ABSTRACT

Phylogenetic footprinting is a method for the discovery of regulatory elements in a set of homologous regulatory regions, usually collected from multiple species. It does so by identifying the most conserved motifs in those homologous regions. This note describes web software that has been designed specifically for this purpose in prokaryotic genomes, making use of the phylogenetic relationships among the homologous sequences in order to make more accurate predictions. The software is called MicroFootPrinter and is available at http://bio.cs.washington.edu/software.html.

INTRODUCTION

One of the current challenges facing biologists is the discovery of novel functional elements in noncoding genomic sequence. With the rapidly increasing number of genomes being sequenced, a comparative genomics approach called 'phylogenetic footprinting' has become a favored method for such discovery. The idea underlying phylogenetic footprinting is that selective pressure causes functional elements to evolve at a slower rate than the nonfunctional surrounding sequence. Therefore the most conserved motifs in a collection of homologous regions are excellent candidates as functional elements.

This note focuses on phylogenetic footprinting for the discovery of novel *cis*-regulatory elements in prokaryotic genomes. A web tool for this purpose has been implemented in a program called MicroFootPrinter, available at http://bio.cs.washington.edu/software.html. One reason to focus on prokaryotes is that over 300 prokaryotic genomes are completely sequenced at the time of this writing, making this by far the richest current medium for phylogenetic footprinting. MicroFootPrinter gives the user automatic, full access to all these genomes.

USER INPUTS

MicroFootPrinter is actually a front end for the FootPrinter phylogenetic footprinting program (1), but specifically tailored to prokaryotic genomes. The user simply supplies a prokaryotic species and gene of interest. MicroFootPrinter automatically takes care of the laborious tasks of (i) finding homologous genes in related prokaryotes, (ii) inferring their phylogenetic gene tree, (iii) extracting the noncoding *cis*-regulatory regions of each of these homologous genes, (iv) setting the most difficult of FootPrinter's parameters and (v) running FootPrinter on these regulatory regions. The result is the identification of motifs that are well conserved across the *cis*-regulatory regions of these homologous genes. [The reader is referred to earlier work (1,2) for details on FootPrinter and examples of its applications to biological data].

MicroFootPrinter's 'Search' feature is very useful for quickly finding species and genes of interest. The user enters any search terms, separated by spaces. All search fields are considered, and any partial or complete match found is included in the results. For instance, if the user enters 'coli' for the species search; MicroFootPrinter offers the list of all *Escherichia coli* strains available. After choosing a species, if the user enters 'pyrim' for the gene search, MicroFootPrinter offers a list of all genes with this text in their gene product descriptions, notably genes involved in processing of pyrimidines.

After choosing a species and gene, the user is asked to supply a few simple parameters (or leave them at their default values). These are the length of the desired motif (in base pairs), the target number of motifs for MicroFootPrinter to display, the target number of species in which to locate homologous genes, and the maximum parsimony score (number of mutations) to allow among the instances of each displayed motif. If desired, the search for other species can also be restricted to any taxonomic clade containing the user's chosen species, for instance, restricted to just γ–proteobacteria.

*To whom correspondence should be addressed. Tel: +1 206 543 9263; Fax: +1 206 543 8331; Email: tompa@cs.washington.edu

For each of these user inputs there are links marked '?' that lead to further description. These include explanations of the input parameters and advice on adjusting them.

After the user has set the parameters, it typically takes 1 to 2 min of elapsed time for MicroFootPrinter to perform all its computations and display FootPrinter's output. For a description and interpretation of FootPrinter's output, the reader is referred to earlier work (1).

METHODS USED BY MicroFootPrinter

MicroFootPrinter uses protein-level BLAST to find the closest homologs to the user's chosen gene. Specifically, it uses NCBI's BLink facility, which provides the results of BLAST searches that have been done for every protein sequence in the Entrez Proteins data domain. If there are close homologs in multiple sequenced strains of the same species, MicroFootPrinter will select only the single strain whose homolog's protein sequence is most similar to the query sequence.

FootPrinter requires as input a phylogeny relating the homologous sequences. MicroFootPrinter infers this phylogeny by using ClustalW (3) to align the homologous protein sequences. The guide tree returned by ClustalW is used as a reasonable approximation of the true gene tree.

For each of these homologous genes, MicroFootPrinter next extracts the *cis*-regulatory regions in which FootPrinter will report conserved motifs. Each of these regions consists of up to 500 bp of noncoding sequence upstream of the start codon. (It may be shorter, if there is another coding region fewer than 500 bp upstream.) Note that these regulatory regions typically contain both 5′ untranslated region (5′ UTR) and promoter sequences. The fact that 5′ UTR is included makes MicroFootPrinter useful for discovery of *cis*-regulatory mRNA elements such as riboswitches. Indeed, it has already proven useful in this role (4).

The prevalence of operons in prokaryotic genomes complicates the extraction of the regulatory regions. Operons are contiguous collections of genes on the same DNA strand that are transcribed together. Typically the intergenic distance between consecutive genes in an operon is extremely small. The complication in this case is that the desired regulatory region may be upstream of the entire operon rather than immediately upstream of the selected gene. For most prokaryotes, it is not known which genes comprise operons.

To handle this complication in a conservative manner, MicroFootPrinter extracts and concatenates the noncoding sequences upstream of the gene and upstream of its plausible operon. Specifically, if the next coding region upstream is in the same orientation and fewer than 100 bp upstream, this short intergenic sequence is concatenated with the result of applying this same procedure to the upstream gene. This process continues until interrupted either by a coding region in the opposite orientation or an intergenic region longer than 100 bp. Up to 500 bp of this final intergenic region are also concatentated to the result. These concatenated noncoding sequences are actually separated from each other by the sequence NNNNNNNNNN so that, when inspecting the ultimate FootPrinter output, the user can identify when such concatenation has taken place.

In addition to providing the user with FootPrinter's output, MicroFootPrinter also provides the protein sequences, *cis*-regulatory sequences and gene tree. With these, the user can rerun FootPrinter directly, adjusting FootPrinter's parameters if desired, or use another motif discovery tool.

DISCUSSION

There are many programs available for motif discovery. Most of these are not intended for phylogenetic footprinting, as they implicitly assume that the input sequences are independent rather than homologous. The traditional approach to phylogenetic footprinting has been via multiple sequence alignment. We believe that, for sequences as diverged as the prokaryotes that are currently sequenced, this approach is less effective than the use of FootPrinter, which searches for conserved motifs directly in unaligned sequences.

MicroFootPrinter provides the microbiologist with a convenient front end for FootPrinter, whereby specification of only the species and gene of interest is sufficient for the extraction of all the data necessary for phylogenetic footprinting on that gene. Ultimately, we would like to extend this service to the eukaryotes, but this is still premature. For the few eukaryotes that are currently completely sequenced, a static catalog of all regulatory elements discovered by phylogenetic footprinting (5–8) is probably more appropriate at this time.

Another extension that could be very helpful is the ability to analyze multiple genes from a single species for common regulatory elements, using the homologs of each gene as well. This is a more difficult problem than simple phylogenetic footprinting, one for which FootPrinter was not intended. For discussion of what makes this problem more difficult and some approaches to its solution, the reader is referred to recent work (9–12).

ACKNOWLEDGEMENTS

The authors thank Jieyang Hu, Martha Mercaldi, Scott Rose, Larry Ruzzo, Travis Wright and the NCBI User Service for advice and assistance in this project. This material is based upon work supported in part by the National Science Foundation under grant DBI-0218798 and by the National Institutes of Health under grant R01 HG02602. The Open Access publication charges for this article were waived by Oxford University Press.

Conflict of interest statement. None declared.

REFERENCES

1. Blanchette,M. and Tompa,M. (2003) FootPrinter: a program designed for phylogenetic footprinting. *Nucleic Acids Res.*, **31**, 3840–3842.
2. Blanchette,M. and Tompa,M. (2002) Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res.*, **12**, 739–748.
3. Chenna,R., Sugawara,H., Koike,T., Lopez,R., Gibson,T.J., Higgins,D.G. and Thompson,J.D. (2003) Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.*, **31**, 3497–3500.
4. Yao,Z., Weinberg,Z. and Ruzzo,W.L. (2006) CMfinder–a covariance model based RNA motif finding algorithm. *Bioinform.*, **22**, 445–452.

5. Cliften,P., Sudarsanam,P., Desikan,A., Fulton,L., Fulton,B., Majors,J., Waterston,R., Cohen,B.A. and Johnston,M. (2003) Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science*, **301**, 71–76.
6. Kellis,M., Patterson,N., Endrizzi,M., Birren,B. and Lander,E.S. (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, **423**, 241–254.
7. Xie,X., Lu,J., Kulbokas,E.J., Golub,T.R., Mootha,V., Lindblad-Toh,K., Lander,E.S. and Kellis,M. (2005) Systematic discovery of regulatory motifs in human promoters and $3'$ UTRs by comparison of several mammals. *Nature*, **434**, 338–345.
8. Prakash,A. and Tompa,M. (2005) Discovery of regulatory elements in vertebrates through comparative genomics. *Nat. Biotechnol.*, **23**, 1249–1256.
9. Wang,T. and Stormo,G.D. (2003) Combining phylogenetic data with coregulated genes to identify regulatory motifs. *Bioinform.*, **19**, 2369–2380.
10. Moses,A.M., Chiang,D.Y. and Eisen,M.B. (2004) Phylogenetic motif detection by expectation-maximization on evolutionary mixtures. In Altman,R.B., Dunker,A.K., Hunter,L., Jung,T.A. and Klein,T.E. (eds), *Pacific Symposium on Biocomputing*. World Scientific Publishing Co., pp. 324–335.
11. Sinha,S., Blanchette,M. and Tompa,M. (2004) PhyME: a probabilistic algorithm for finding motifs in sets of orthologous sequences. *BMC Bioinform.*, **5**, 170.
12. Siddharthan,R., Siggia,E.D. and van Nimwegen,E. (2005) PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput. Biol.*, **1**, e67.