A suite of web-based programs to search for transcriptional regulatory motifs

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Published Version</td>
<td>doi:10.1093/nar/gkh461</td>
</tr>
<tr>
<td>Citable link</td>
<td><a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:27002640">http://nrs.harvard.edu/urn-3:HUL.InstRepos:27002640</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA">http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA</a></td>
</tr>
</tbody>
</table>
A suite of web-based programs to search for transcriptional regulatory motifs

Yueyi Liu, Liping Wei³, Serafim Batzoglou¹, Douglas L. Brutlag², Jun S. Liu⁴ and X. Shirley Liu⁵,*

Stanford Medical Informatics, 251 Campus Drive X215, ¹Department of Computer Science and ²Department of Biochemistry, Stanford University, Stanford, CA 94305, USA, ³Center of Bioinformatics, College of Life Sciences, Peking University, Beijing 100871, People’s Republic of China, ⁴Department of Statistics, Harvard University, Cambridge, MA 02138, USA, and ⁵Department of Biostatistics, Harvard School of Public Health, Dana-Farber Cancer Institute, Boston, MA 02115, USA

Received February 15, 2004; Revised and Accepted April 27, 2004

ABSTRACT
The identification of regulatory motifs is important for the study of gene expression. Here we present a suite of programs that we have developed to search for regulatory sequence motifs: (i) BioProspector, a Gibbs-sampling-based program for predicting regulatory motifs from co-regulated genes in prokaryotes or lower eukaryotes; (ii) CompareProspector, an extension to BioProspector which incorporates comparative genomics features to be used for higher eukaryotes; (iii) MDscan, a program for finding protein–DNA interaction sites from ChIP-on-chip targets. All three programs examine a group of sequences that may share common regulatory motifs and output a list of putative motifs as position-specific probability matrices, the individual sites used to construct the motifs and the location of each site on the input sequences. The web servers and executables can be accessed at http://seqmotifs.stanford.edu.

INTRODUCTION
Regulatory elements are short sequences of DNA (5–20 bp in length) that determine the timing, location and level of gene expression (1). In recent years, sequences that contain regulatory elements have become easily available due to the effort for large-scale sequencing of many genomes. Meanwhile, technologies such as microarray and ChIP-on-chip (or GSLA for Genome-Scale Location Analysis) make it feasible to identify potential targets of transcription factors. Incidentally, many computational methods, such as MEME (2), AlignACE (3) and Consensus (4), have been developed in the past decade to predict regulatory elements and regulatory motifs. Compared with experimental procedures to determine regulatory elements, computational motif-finding programs are fast and inexpensive. Their predictions provide biologists with valuable hypotheses for experimental validation.

Our group has developed a suite of specialized regulatory motif-finding programs. The first program is BioProspector, which is based on the original Gibbs Motif Sampler (5), but has several important improvements. It has been applied successfully to prokaryotes and lower eukaryotes such as Bacillus subtilis and yeast, respectively. In higher eukaryotes, upstream regions are usually longer and noisier. CompareProspector, with a comparative genomics component on top of BioProspector, takes in cross-species sequence comparison information to help guide the search in higher eukaryotes. We also developed MDscan, a program for motif finding from ChIP-on-chip targets. The algorithms for these three programs have been tested and published. We have since developed them into interactive web-based applications available as a public resource at http://seqmotifs.stanford.edu. Also on the website is comprehensive information about the program, format for input/output and stand-alone executables for local use. These programs are complementary in nature, and together they provide a useful resource for the study of gene expression regulation.

DESCRIPTION OF THE PROGRAMS
BioProspector
BioProspector (http://seqmotifs.stanford.edu or http://bioprospector.stanford.edu) is a program that uses Gibbs sampling to search a list of sequences (e.g. promoters of co-regulated genes) for potential regulatory motifs (6). Gibbs sampling first initializes the motif matrix using $w$mers ($w$ being the width of the motif) randomly selected from the input.

*To whom correspondence should be addressed. Tel: +1 617 632 2472; Fax: +1 617 632 2444; Email: xsliu@jimmy.harvard.edu

The online version of this article has been published under an open access model. Users are entitled to use, reproduce, disseminate, or display the open access version of this article provided that: the original authorship is properly and fully attributed; the Journal and Oxford University Press are attributed as the original place of publication with the correct citation details given; if an article is subsequently reproduced or disseminated not in its entirety but only in part or as a derivative work this must be clearly indicated.

© 2004, the authors

Nucleic Acids Research, Vol. 32, Web Server issue © Oxford University Press 2004; all rights reserved
Then it samples from all \( w \) mers in the input sequences to update the motif matrix. The probability of selecting a \( w \) mer is proportional to the likelihood of generating the \( w \) mer from the current motif matrix over the likelihood of generating it from the non-motif background. The motif matrix is updated until convergence, or until a certain number of sampling iterations has been reached.

BioProspector has several significant improvements compared with the original Gibbs Motif Sampler (5). It uses a Markov model estimated from all promoter sequences in the genome to model adjacent nucleotide dependency and improve motif specificity. It also adopts two thresholds to allow each input sequence to contain zero to multiple copies of the motif. As prokaryotic motifs often occur in two blocks with a gap of variable length, BioProspector is capable of modeling motifs with two blocks or with palindromic patterns. In MDscan (7), a better motif scoring function has been identified via simulations and theoretical analyses, and adopted in the current version of BioProspector.

**CompareProspector**

Non-coding sequences in higher eukaryotes are typically much longer than those of prokaryotes. As a result, eukaryotic regulatory motif prediction faces the challenge of sifting for motif signals with much more noise. Based on our finding that known regulatory elements in non-coding sequences are more likely to be conserved than background non-coding sequences (8), we extended BioProspector to bias the motif search towards regions that are conserved across species. Even when only a subset of the input sequences has identifiable orthologs, CompareProspector (http://seqmotifs.stanford.edu or http://compareprospector.stanford.edu) shows improved performance over BioProspector for human/mouse or *Caenorhabditis elegans/C. briggsae* sequences.

To utilize cross-species sequence information, CompareProspector takes as input an array of window percentage identity values (WPIDs) for each input sequence with available ortholog. Although we calculated the WPIDs based on the global alignment of each input sequence with its ortholog generated using LAGAN (9), the WPIDs can be calculated using any alignment method the user chooses. In Gibbs sampling iterations, CompareProspector biases the motif search towards sequences conserved across species. First of all, users can specify two WPID thresholds. During initial iterations, only positions whose WPID values are above the high threshold are sampled. Subsequently, the WPID cutoff can be gradually decreased to the low conservation threshold to allow sampling of less conserved positions. In addition, the probability of selecting a \( w \) mer to update the motif matrix is weighted by sequence conservation to favor sampling of more conserved sequences. Sequences without orthologs are assigned the low conservation threshold, so as to participate in sampling only in later iterations.

**MDscan**

MDscan (http://seqmotifs.stanford.edu or http://mdscan.stanford.edu/) is designed for motif finding from sequences obtained from a ChIP-on-chip or GSLA experiment, a procedure now routinely used to characterize genome-wide protein–DNA interaction and transcription regulation (7). Since many of the sequences selected by ChIP-on-chip, especially highly enriched ones, often have multiple copies of the motif, MDscan first enumerates all the non-redundant \( w \) mers as seeds in the top \( t \) most enriched sequences. For each seed, MDscan constructs a candidate motif matrix using all the \( w \) mers in the top \( t \) sequences that are ‘neighbors’ of that seed. A pair of \( w \) mers is considered ‘neighbors’ if it shares at least \( m \) matched positions, where the probability for two random \( w \) mers sharing \( \geq m \) matched positions is 0.0015. All the motif matrices are evaluated using a semi-Bayesian scoring function and the best ones are saved. In the subsequent updating process, each retained motif matrix is refined by adding or removing \( w \) mers in all the sequences to increase the motif score, and the best refined motifs are reported.

We recently developed a new program, Motif Regressor (10), to better utilize mRNA expression level or ChIP-on-chip enrichment information to improve the performance of MDscan. Motif Regressor first identifies a set of non-redundant candidate motifs using MDscan, and scans the promoter region of every gene in the genome with each candidate motif to measure how well a promoter matches a motif (in terms of both the number of sites and the strength of matching). It then uses linear regression analysis to select motifs whose promoter matching scores are significantly correlated with ChIP-on-chip enrichment or downstream gene expression values. When ranking motifs by linear regression \( p \)-value, Motif Regressor automatically picks the best motif and optimal motif width.

Due to its computational intensity, Motif Regressor is not currently available as a web server. However, for interested users to explore the program locally, the program is available for download at http://www.techtransfer.harvard.edu/Software/MotifRegressor/.

**INPUTS/OUTPUTS OF THE PROGRAMS**

**Inputs and parameters**

The following inputs are required for the web servers of all three programs above.

(i) User’s email. Results of the search will be emailed to the user.
(ii) A user-defined job name, which will be sent in the result email.
(iii) A list of sequences that may share regulatory motif(s). These sequences can be obtained from microarray experiments, ChIP-on-chip experiments, manually selected sequences hypothesized to share sequence motifs or other methods.
(iv) The width(s) of the motif to search for.
(v) Background model. Users can specify their own background model or choose from a list of pre-computed models for many genomes.
(vi) Whether the motif occurs in every sequence or just some of the sequences.
(vii) Whether to search in both strands of each sequence or just the forward strand.
(viii) Number of top motifs to report.
The highest scoring 15 motifs are:

<table>
<thead>
<tr>
<th>Motif #1:</th>
<th>CAGCCTGTC/GACAGCTGC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Width (8, 0); Gap [0, 0]; MotifScore 3.159; Sites 21</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Block</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
<th>Consensus</th>
<th>rConsensus</th>
<th>rDegConsensus</th>
<th>SeqLen</th>
<th>Site</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.20</td>
<td>99.38</td>
<td>0.23</td>
<td>0.20</td>
<td>C</td>
<td>C</td>
<td>G</td>
<td>Site #1 f 3120</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>99.35</td>
<td>0.23</td>
<td>0.23</td>
<td>0.20</td>
<td>A</td>
<td>T</td>
<td>A</td>
<td>T</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.20</td>
<td>42.72</td>
<td>65.49</td>
<td>0.20</td>
<td>C</td>
<td>G</td>
<td>C</td>
<td>Site #1 r 10857</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.20</td>
<td>99.38</td>
<td>0.23</td>
<td>0.20</td>
<td>C</td>
<td>G</td>
<td>C</td>
<td>G</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.20</td>
<td>0.23</td>
<td>0.23</td>
<td>99.35</td>
<td>T</td>
<td>A</td>
<td>T</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.20</td>
<td>0.23</td>
<td>99.38</td>
<td>0.20</td>
<td>G</td>
<td>C</td>
<td>G</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.20</td>
<td>33.28</td>
<td>66.30</td>
<td>0.23</td>
<td>T</td>
<td>A</td>
<td>Y</td>
<td>R</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.20</td>
<td>56.89</td>
<td>0.23</td>
<td>42.69</td>
<td>C</td>
<td>G</td>
<td>Y</td>
<td>R</td>
<td></td>
</tr>
</tbody>
</table>

Figure 1. Typical output for BioProspector, CompareProspector and MDscan.

DISCUSSION

We present three specialized regulatory motif-finding programs and their web servers. BioProspector, a program that can search for two-block motifs with variable gap and palindrome motifs, is ideal for identifying motifs in prokaryotes and lower eukaryotes. CompareProspector is more powerful when orthologous sequences from other species are available. It gives better performance than several other motif-finding programs using datasets from both human and Caenorhabditis elegans (8). MDscan works best with ChIP-on-chip data, where it can take advantage of enrichment level information, or any dataset where the user can rank the sequences according to how likely motif sites are to occur.

ACKNOWLEDGEMENTS

The authors would like to thank Marina Siroti for her help with webpage design, Russ Altman and Stuart Kim for their insight and help during the development of CompareProspector, and Erin Conlon for her contribution in developing Motif Regressor.

REFERENCES


