Nucleic Acids Research, 2006, Vol. 34, Web Server issue ***W555–W559***
doi:10.1093/nar/gkl224

Stubb: a program for discovery and analysis of *cis*-regulatory modules

Saurabh Sinha*, Yupu Liang[1] and Eric Siggia[1]

Department of Computer Science, University of Illinois, Urbana-Champaign and [1]Center for Studies in Physics and Biology, The Rockefeller University, NY, USA

Received February 14, 2006; Revised March 6, 2006; Accepted March 27, 2006

ABSTRACT

Given the DNA-binding specificities (motifs) of one or more transcription factors, an important bioinformatics problem is to discover significant clusters of binding sites for the transcription factors(s). Such clusters often correspond to *cis*-regulatory modules mediating regulation of an adjacent gene. In earlier work, we developed the Stubb program that uses a probabilistic model and a maximum likelihood approach to efficiently detect *cis*-regulatory modules over genomic scales. It may optionally exploit a second related genome to improve module prediction accuracy. We describe here the use of a web-based interface for the Stubb program. The interface is equipped with a special post-processing step for in-depth analysis of specific modules, in order to reveal individual binding sites predicted in the module. The web server may be accessed at the URL http://stubb.rockefeller.edu/.

DESCRIPTION

Dissection of a gene regulatory pathway, such as that involved in segmentation of the fruitfly embryo (1), requires identification of *cis*-regulatory sequences that mediate the combinatorial action of multiple transcription factors on target genes. These *cis*-regulatory sequences, called 'modules', are typically 500–1000 bp long, and harbor one to many binding sites for different transcription factors. In a typical scenario, the scientist has a small set of transcription factors whose binding specificities (motifs) are known, and wants to find modules (and genes) targeted by these factors, over genomic scales.

Computational approaches to the module detection problem are based on discovering statistically significant clusters of predicted occurrences of input transcription factor motifs (2,3). The Cister (4), Ahab (5) and Stubb (6) programs use hidden Markov models (HMM) to obtain a statistically sound score for modules, namely, the likelihood that the module sequence was generated by a model, and eliminate the need for *ad hoc* thresholds on what constitutes a motif occurrence. Ahab and Stubb use maximum likelihood to infer the relative weights of the input motifs, leaving no parameters to be adjusted by hand. Module predictions made using this approach have been subjected to extensive experimental validation, and have led to several new modules being discovered for the segmentation gene network in fruitfly (1). Note that Stubb differs from existing tools like MAPPER (7) that predict binding sites of a single given motif in an input sequence.

The Stubb algorithm can additionally exploit a second, closely related genomic sequence, to improve the accuracy of module discovery, as demonstrated for two fruitfly genomes in (8). This is done by incorporating a probabilistic model of binding site evolution into the HMM framework used in the single-species case. An important by-product of the algorithm (for either the single- or the two-species case) is prediction of the number, locations and strengths (posterior probabilities) of sites of each transcription factor, for any candidate module. These predictions, called the 'module composition', provide invaluable clues on the function of the module, given prior knowledge about the functions of the individual transcription factors. This approach has led to a comprehensive study of evolution in the regulation of anterior–posterior patterning between *Drosophila melanogaster* and *Drosophila pseudoobscura* (9).

Here we report on a user-friendly web interface to Stubb, available at http://stubb.rockefeller.edu/. The reader is referred to earlier work (6) for details on the Stubb algorithm. The source code for the Stubb program is available at http://edsc.rockefeller.edu/cgi-bin/stubb/download.pl, and includes a few features not available through the web interface, such as the ability to exploit the relative order and positions of factor binding sites. Users of the web server may cite this document, or the original Stubb manuscript (6).

*To whom correspondence should be addressed. Tel: +1 217 333 3233; Fax: +1 217 244 6500; Email: sinhas@cs.uiuc.edu

WEB INTERFACE

Stubb web is organized into three major components

(i) *Sequence upload*: This allows the user to upload the DNA sequence(s) to be analyzed. If working with the *D.melanogaster* genome, the user may specify the genomic coordinates of the DNA sequence(s), and the server automatically extracts them, along with orthologous sequences from one of several closely related species.
(ii) *Stubb*: This is the interface for running Stubb on long genomic sequences (tens of Kb long), with one or more user-input motifs, to locate potential modules.
(iii) *Windowfit*: This is the interface for analyzing the composition (as determined by Stubb) of one or more modules (typically 1–2 Kb long, each).

All three components of Stubb web are accessible from links on the main page (http://stubb.rockefeller.edu/). We describe each of these components next.

SEQUENCE UPLOAD

This is an optional but recommended first step that prepares the sequence data for input to Stubb, in the correct format. It is particularly useful if the user is working with the *D.melanogaster* genome.

Naming the dataset

The user must first input a 'project name' and a 'sequence file name' for the data. Data are organized by projects, which are like folders, and may contain multiple sequence files. For example, a project may be named '*Drosophila*', while a sequence file in this project may be named 'segmentation', in case the sequences to be analyzed are upstream regions of segmentation genes in *Drosophila*. The user should also choose if the analysis is going to be based on 'single-species' or 'multiple-species' which defines whether the individual sequences in the file are analyzed singly or in pairs.

Specifying the sequence

- The first option here (Option A) is to specify one or more loci in the *D.melanogaster* genome (Release 3.1 or 4.3 coordinates); if specifying multiple loci, the GFF format must be used. (To facilitate the recovery of coordinates for a feature of interest, a link is provided to Gbrowse, Fruitfly Release 3.1 or 4.3, that allows searches.) Additionally, if preparing for multiple-species analysis, the user must specify the 'secondary species', (one of *D.pseudoobscura*, *Drosophila yakuba*, *Drosophila ananassae*, *Drosophila mojavensis* or *Drosophila virilis*), whose genomes have been pre-aligned with the 'reference species' (*D.melanogaster*) to facilitate ortholog extraction. There are two additional parameters, 'select matching contig' that helps resolve cases of multiple matches with the secondary genome, and 'augment query interval' that helps define precise orthology boundaries. Both these parameters are described in detail on the web page, and may be left at their default values by the beginning user.
- The second way to specify sequences (Option B) is to directly upload the sequence data in Fasta format. This method may be used regardless of the species, and in case of multiple-species analysis, orthologous sequences from reference and secondary species must be interlaced.

Clicking on the 'GO' button causes Stubb to extract the sequences from one or more species, as specified by the user, and store them on the server. The resulting page provides a link to this file, prints details of where the orthologous sequences were extracted from, and prompts the user to proceed to the next step of analysis, i.e. Stubb ('RunStubb') or Windowfit ('RunWindowfit').

STUBB

(i) Sequence input: The Stubb page requires the input sequences to be specified first. If the analysis is going to be on pre-specified sequences, the project name and sequence file name are sufficient to identify the dataset. If this page was reached via the 'Sequence Upload' step, these fields are already entered. This page may also be reached directly, in which case the Fasta file containing the sequences must be uploaded here ('Upload sequences' or 'Paste sequences'), and the project and sequence file names must be created.
(ii) Motif input: Motifs capturing the DNA-binding specificity of known transcription factors must be provided, as position weight matrices, through 'Upload your own matrices'. Alternatively, the user may choose from a small set of internally stored motifs associated with the segmentation gene network in fruitfly, or from a larger compendium of 75 *Drosophila* motifs compiled by Daniel Pollard (http://rana.lbl.gov/~dan/matrices.html), based on the *Drosophila* DNase I Footprint database (10).
(iii) Algorithm parameters: The remaining parameters may be left at their default values, especially by a beginning user. These include the following

- Parameters to specify the background sequence model,
- The 'phylogeny (mu)' parameter to specify evolutionary distance between the reference and secondary species,
- Sliding window parameters ('Window shift' and 'Window length') to specify how the genomic sequences will be scanned,
- Advanced Stubb options to specify thresholds for reporting modules and binding sites, and
- Lagan parameters. These are relevant in multiple-species analysis, where the first step is to align the two sequences using the Lagan alignment program [(11), see http://lagan.stanford.edu/lagan_web/index.shtml]. The alignment output by Lagan is post-processed to extract ungapped orthologous blocks of high percent identity, and the last two of the 'Lagan parameters' can be used to control this post-processing step.

Clicking on the 'GO' button has one of two possible results. If in multiple-species mode, the results of running Lagan (and post-processor) are output in the new page, with an option of re-doing this step with changed parameter values ('Rerun Lagan'). Clicking 'Continue to Stubb' causes Stubb to be run. In case of single-species analysis, this intermediate step is skipped and Stubb is run directly.

The resulting page has separate links to the results page for each of the input sequences (or genomic loci), as well as a link to a combined results page. A Stubb results page (Figure 1) has the following options:

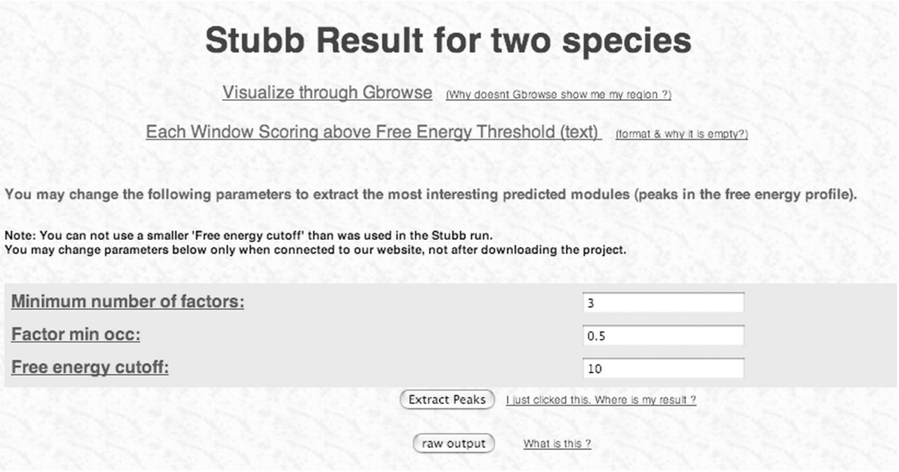
- 'Visualize through Gbrowse', a link to display the Stubb score (free energy) profile on a genomic atlas, using the Gbrowse software (Lincoln Stein). Figure 2 shows an example of such a display. Options are provided to manage and simultaneously display multiple free energy profiles to facilitate the comparison of related Stubb runs. Other annotation files may be overlayed on the display using features of Gbrowse itself. If using sequences other than *Drosophila* loci, Gbrowse is used only to provide a visualization interface, rather than the richly annotated genomic atlas that it is typically used as.
- The user may view, in plain text, a list of all candidate modules that satisfy certain criteria, which the user may control ('Extract Peaks').
- The link 'raw output' leads to a page where the files output by Stubb are accessible in their original format.

WINDOWFIT

Windowfit is a program that displays binding sites (predicted by Stubb) in a graphical format. A by-product of running Stubb on any sequence window is the probability of each substring of the sequence being a site for each of the input PWMs. Windowfit collects this information from Stubb's output and graphically displays the high probability (above a user-specified threshold) predicted sites for all motifs.

User input required for Windowfit is almost identical to that of Stubb (see above). The compulsory inputs are the sequences and motifs, and the remaining inputs are parameters that may be left at their default values. The parameters exclusive to Windowfit are the 'advanced Windowfit options' that specify thresholds for displaying motif occurrences under two different prediction schemes. Clicking on 'GO' runs Windowfit, and displays links to result pages for each of the input sequences.

A Windowfit results page (Figure 3) has a graphical display (plot) of predicted binding sites in the sequence. In case of multiple-species analysis, there are two additional plots, one for single-species prediction on the secondary species, and one



Stubb Result for two species

Visualize through Gbrowse (Why doesnt Gbrowse show me my region ?)

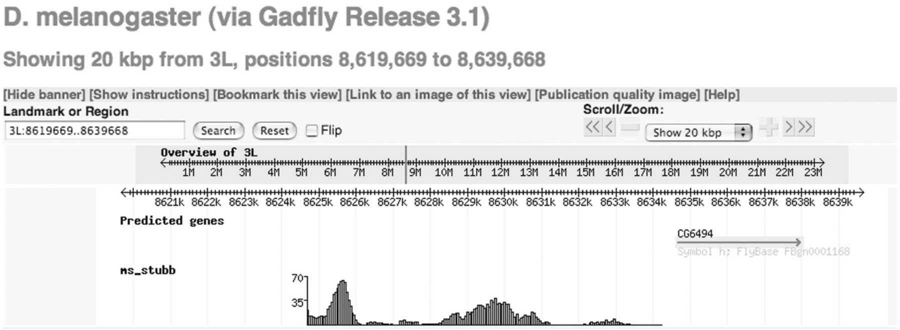
Each Window Scoring above Free Energy Threshold (text) (format & why it is empty?)

You may change the following parameters to extract the most interesting predicted modules (peaks in the free energy profile).

Note: You can not use a smaller 'Free energy cutoff' than was used in the Stubb run.
You may change parameters below only when connected to our website, not after downloading the project.

Minimum number of factors:	3
Factor min occ:	0.5
Free energy cutoff:	10

Figure 1. A typical stubb results page.



D. melanogaster (via Gadfly Release 3.1)

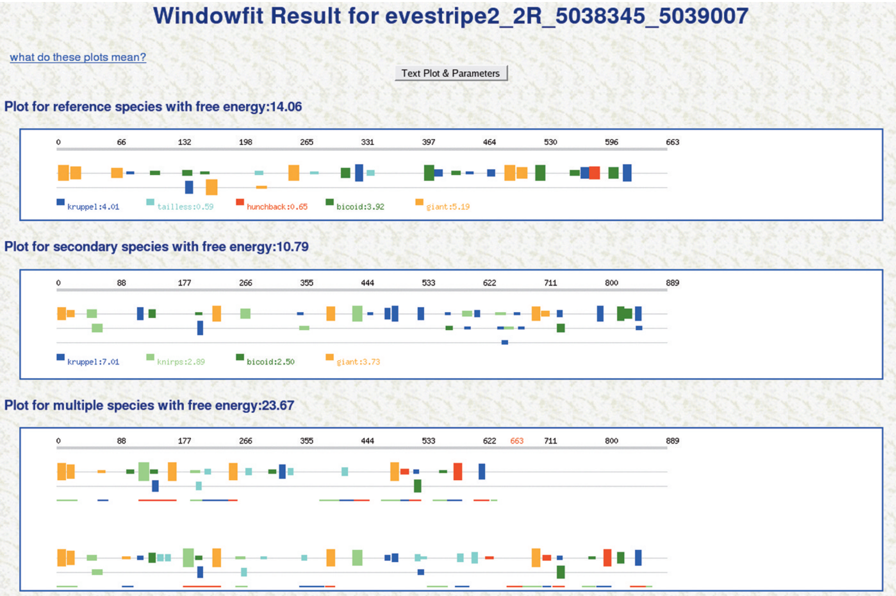
Showing 20 kbp from 3L, positions 8,619,669 to 8,639,668

Landmark or Region: 3L:8619669..8639668

Scroll/Zoom: Show 20 kbp

[Hide banner] [Show instructions] [Bookmark this view] [Link to an image of this view] [Publication quality image] [Help]
Search
Reset
□ Flip
Overview of 3L
1M 2M 3M 4M 5M 6M 7M 8M 9M 10M 11M 12M 13M 14M 15M 16M 17M 18M 19M 20M 21M 22M 23M
8621k 8622k 8623k 8624k 8625k 8626k 8627k 8628k 8629k 8630k 8631k 8632k 8633k 8634k 8635k 8636k 8637k 8638k 8639k
Predicted genes
CG6494
Symbol h; FlyBase FBgn0001168
ms_stubb
70
35

Figure 2. Gbrowse display of stubb free energy profile for the hairy gene locus.



Windowfit Result for evestripe2_2R_5038345_5039007

Plot for reference species with free energy:14.06

Plot for secondary species with free energy:10.79

Plot for multiple species with free energy:23.67

Figure 3. A typical windowfit results page.

showing predictions made by Stubb in the two-species mode. Each plot has colored bars indicating the position and type of each binding site. The bar height encodes the site strength (posterior probability). There are additional lines with bars if binding sites overlap. The cumulative strength of each factor's sites is listed next to the factor's name. In case of multiple-species analysis, the plot for 'multiple species' has colored line segments (in alternating colors green, blue and red) that depict corresponding aligned blocks between the species. Binding sites overlapping these blocks are fit with an evolutionary model and are necessarily conserved (with the same posterior probability) in the two species.

Above the plots is a link ('Text plot and parameters') that leads to more detailed plots, where the actual sequences are spelled out. Below the Windowfit plots, there is one Information Content Plot for each species analyzed. (Data not shown in Figure 3.) This displays binding sites predicted based on the probability distribution induced by each PWM. In other words, for every substring, the probability of sampling it from a motif is calculated, normalized against background, and the substring is predicted as a site if the score is above some threshold.

GENERAL FEATURES

(i) Stubb web uses cookies on the user's browser to implement a transparent authentication scheme to protect the privacy of the user's data. No explicit usernames and passwords are required, and the user is unaware of the underlying security mechanism unless attempting to access an earlier project built from another machine.

(ii) All data and result files in a project may be downloaded to the user's local machine and be viewed without connection to the server. Clicking the 'download the project' button presents a 'tar' archive of the project for download.

(iii) We limit the total sequence length for Stubb runs, so for genome-wide scans the user will have to run the program locally.

(iv) Project folders that have not been accessed recently are removed from the system.

ACKNOWLEDGEMENTS

Support was provided by the NIH under grant GM66434. Funding to pay the Open Access publication charges for this article was provided in part by the NIH under grant GM66434 and in part by SS's start-up funds from the University of Illinois, Urbana-Champaign.

Conflict of interest statement. None declared.

REFERENCES

1. Schroeder,M.D., Pearce,M., Fak,J., Fan,H., Unnerstall,U., Emberly,E., Rajewsky,N., Siggia,E.D. and Gaul,U. (2004) Transcriptional control in the segmentation gene network of *Drosophila*. *PLoS Biol*., **2**, E271.
2. Berman,B.P., Nibu,Y., Pfeiffer,B.D., Tomancak,P., Celniker,S.E., Levine,M., Rubin,G.M. and Eisen,M.B. (2002) Exploiting transcription factor binding site clustering to identify *cis*-Regulatory modules involved in pattern formation in the *Drosophila* genome. *Proc. Natl Acad. Sci. USA*, **99**, 757–762.
3. Halfon,M.S., Grad,Y., Church,G.M. and Michelson,A.M. (2002) Computation-based discovery of related transcriptional regulatory modules and motifs using an experimentally validated combinatorial model. *Genome Res*., **12**, 1019–1028.
4. Frith,M.C., Hansen,U. and Weng,Z. (2001) Detection of *cis*-element clusters in higher eukaryotic DNA. *Bioinformatics*, **17**, 878–889.
5. Rajewsky,N., Vergassola,M., Gaul,U. and Siggia,E.D. (2002) Computational detection of genomic *cis*-regulatory modules applied to body patterning in the early *Drosophila* embryo. *BMC Bioinformatics*, **3**, 30.
6. Sinha,S., van Nimwegen,E. and Siggia,E.D. (2003) A probabilistic method to detect regulatory modules. *Bioinformatics*, **19**, i292–i301.
7. Marinescu,V.D., Kohane,I.S. and Riva,A. (2005) MAPPER: a search engine for the computational identification of putative transcription factor binding sites in multiple genomes. *BMC Bioinformatics*, **6**, 79.

8. Sinha,S., Schroeder,M.D., Unnerstall,U., Gaul,U. and Siggia,E.D. (2004) Cross-species comparison significantly improves genome-wide prediction of *cis*-regulatory modules in *Drosophila*. *BMC Bioinformatics*, **5**, 129.
9. Sinha *et al*. (2005) Evolution of *cis*-regulatory modules in the segmentation gene network. In *Proceedings of the 46th Annual Drosophila Research Conference*. San Diego, p. 101.
10. Bergman,C.M., Carlson,J.W. and Celniker,S.E. (2005) *Drosophila* DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, *D.melanogaster*. *Bioinformatics*, **21**, 1747–1749.
11. Brudno,M., Do,C.B., Cooper,G.M., Kim,M.F., Davydov,E., Green,E.D., Sidow,A., Batzoglou,S. and NISC Comparative Sequencing Program. (2003) LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res*., **13**, 721–731.