

Sequence analysis

## BEST: Binding-site Estimation Suite of Tools

Dongsheng Che<sup>1,\*</sup>, Shane Jensen<sup>2</sup>, Liming Cai<sup>1</sup> and Jun S. Liu<sup>3,\*</sup>

<sup>1</sup>Department of Computer Science, University of Georgia, Athens, GA 30602, USA, <sup>2</sup>Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19102, USA and <sup>3</sup>Department of Statistics, Harvard University, Cambridge, MA 02138-2931, USA

Received on December 14, 2004; revised on March 24, 2005; accepted on March 31, 2005

Advance Access publication April 6, 2005

### ABSTRACT

**Summary:** The purpose of our Binding-site Estimation Suite of Tools (BEST) is two-fold: to provide a platform for using and comparing different motif-finding programs for transcription factor binding site prediction, and to improve the accuracy of these predictions by further optimization. Our software package BEST includes four commonly used motif-finding programs: AlignACE, BioProspector, CONSENSUS and MEME, as well as the optimization program BioOptimizer. BEST allows the user to run programs either separately or sequentially and manages all programs by automating the common inputs and the optimization procedure. The BEST system was implemented in Qt, a C++ application development framework, and was compiled and executed on Linux operating systems.

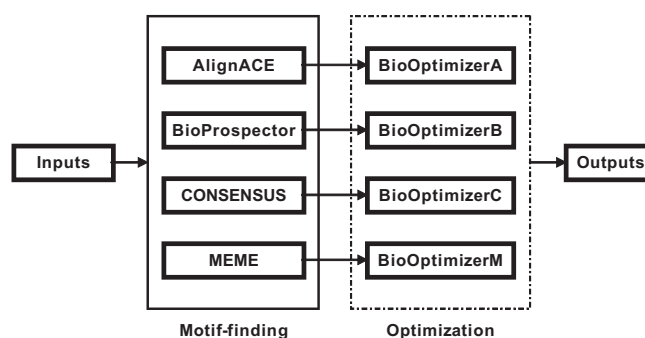
**Availability:** BEST is available for download at <http://www.cs.uga.edu/~che/BEST> and <http://www.fas.harvard.edu/~junliu/BEST>

**Contact:** dsche@uga.edu, jliu@stat.harvard.edu

### INTRODUCTION

A transcription factor binding site (TFBS) is a short DNA segment typically located upstream of a gene and is recognized by a specific transcription factor (TF). Accurate identification of these binding sites contributes to our understanding of gene regulatory networks and aids in the design of biological experiments. Experimental procedures for TFBS determination such as DNase footprinting (Galas and Schmitz, 1978) and gel-shift assays (Garner and Revzin, 1981) are reliable, but are time-consuming and expensive, making them impractical for genome-wide transcription regulation studies. In the past decade, many computational tools have been developed for TFBS prediction. The premise of these computational methods is that the upstream sequences of co-regulated genes tend to be 'enriched' with binding sites recognizable by one or a few TFs. Thus, the computational problem is to find over-represented sequence patterns, called *motifs*, in these upstream sequences.

Existing motif-finding algorithms can be categorized by their underlying computational procedure. AlignACE (Roth *et al.*, 1998), BioProspector (Liu *et al.*, 2001) and Gibbs motif sampler (Liu *et al.*, 1995) are based on a Bayesian statistical model and Gibbs sampling methods (Lawrence *et al.*, 1993; Jensen *et al.*, 2004). MEME (Bailey and Elkan, 1994) uses the EM-algorithm to find the maximum likelihood estimates for a similar statistical model. CONSENSUS (Hertz and Stormo, 1999) uses a sequential search strategy to find motifs



**Fig. 1.** Architectural outline of the BEST system. AlignACE, BioProspector, CONSENSUS and MEME are the motif-finding programs, while BioOptimizerA, BioOptimizerB, BioOptimizerC and BioOptimizerM are their corresponding optimization programs. Each program accepts inputs via their GUI, and the output of each motif-finding program is used as input to Biooptimizer.

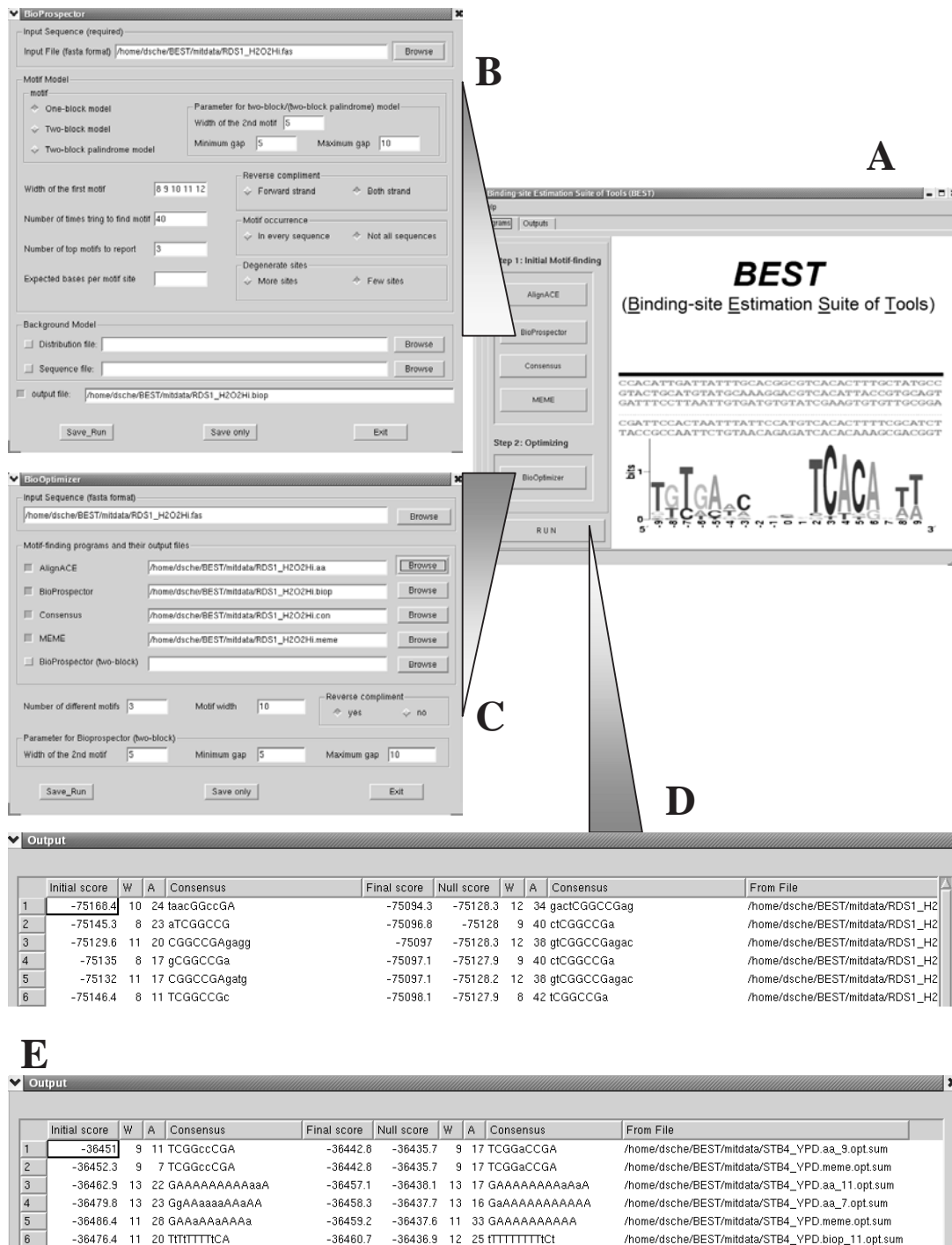
that optimize an information-based scoring function. CUBIC (Olman *et al.*, 2003) formulates the motif prediction problem as a cluster identification problem. MDscan (Liu *et al.*, 2002) combines word enumeration with the Bayesian modeling approach.

However, for a particular dataset, these different algorithms often provide different motif predictions, and no algorithm has been shown to dominate others in all cases. Jensen and Liu (2004) developed BioOptimizer to compare and further refine the output of any motif-finding algorithm. BioOptimizer can handle unknown site abundance and unknown motif width, and can also optimize two-block motifs with variable-length gaps. Using BioOptimizer in conjunction with AlignACE, BioProspector, CONSENSUS or MEME was shown in simulation and real data studies to be superior to using any one of these programs alone.

### THE BEST SYSTEM

Our software package BEST includes a graphic user interface to integrate BioOptimizer with several motif-finding programs, thereby allowing users to compare their results on a single platform. The BEST system contains two core components: the motif-finding programs supported by BEST (currently AlignACE, BioProspector, CONSENSUS and MEME) and the corresponding BioOptimizer algorithm for each program (Fig. 1). If the user is only interested in a particular program, they can set up the fields and click

\*To whom correspondence should be addressed.



**Fig. 2.** Sample inputs and outputs for BEST. (A) Main window of the BEST program. (B) BioProspector GUI with RDS1 sequence file. (C) BioOptimizer GUI with parameters automatically filled based on the BioProspector GUI. (D) Snapshot of the summary output for RDS1 motifs. (E) Snapshot of the summary output for STB4 motifs.

the ‘run’ button for that program. Alternatively, the user can use several programs and compare their results by setting up the GUIs for each one. The ‘run’ button in the main window is used to trigger the workflow engine, which runs the selected programs sequentially.

BEST treats input variables such as the sequence file and the motif width as global parameters, so that inputs into a single GUI are automatically fed to other programs. Outputs of each motif-finding program are automatically filled into the corresponding BioOptimizer GUI. BEST accepts sequence files in the FASTA format, which is

the default format for AlignACE, BioOptimizer, BioProspector and MEME. CONSENSUS uses a specialized sequence format that is generated by BEST when needed. The 'Help' menu in the main window contains detailed information about the plugged-in programs. The color of the button for each program indicates the status of that program—with green indicating running and pink indicating finish.

A limitation of most motif-finding programs is that the user must provide a fixed motif width, which is not usually known a priori. BEST addresses this issue by accepting a range of possible widths and lets BioOptimizer to rank the predictions. In the final output, BEST produces a table for all reported motifs sorted according to their BioOptimizer scores. For each of the top 10 motifs, the table provides the motif score, width, number of predicted sites and consensus sequence from both the original motif-finding program and BioOptimizer.

## IMPLEMENTATION

BEST was implemented in Qt, a complete C++ application development framework, for several reasons. Qt is easy to use and provides a class library and many tools. It supports multiple platforms, including Unix/Linux and Windows platforms. The current version of BEST was compiled and executed on Linux operating systems since many of the plugged-in executable motif-finding programs operate in such systems. Programs included in BEST are AlignACE3.0, BioProspector, CONSENSUS v6c and MEME.3.0.8, as well as BioOptimizer versions for each.

## APPLICATIONS OF BEST

### RDS1 binding sites prediction

This dataset contains upstream sequences of 49 genes selected by a ChIP-chip experiment (Harbison *et al.*, 2004) as targets of transcription factor RDS1. The main window of BEST (Fig. 2A) contains a button for each of the four motif-finding programs, as well as a button for BioOptimizer and a global 'run' button. Figure 2B shows the GUI of the BioProspector program. For this application, we used a one-block model with a width range of 8–12 bps, while all other parameters were left as defaults. The same parameters were used in the GUIs of the other three programs. All entered information can be saved by clicking the 'Save' button. The buttons for all plugged-in programs in the main window are green, indicating that the program is ready to run. After the 'run' button in the main window is clicked, the workflow engine triggers all programs to run sequentially, and is completed when the green buttons change color to pink. A portion of the program output is displayed in Figure 2D. This table shows that the best scored motif is 'gactCGGCCGag' [pattern 'kCGGCCGa' was reported in Harbison *et al.* (2004)].

### STB4 binding sites

Our second dataset consists of 28 regulatory sequences bound by STB4. The motif pattern 'TCGgnnCGA' has been predicted and

validated in Harbison *et al.* (2004). We ran all four motif-finding programs within BEST using a one-block model with a motif width ranging from 7 to 11 bps. The combined output (Fig. 2E) from the four motif-finding programs alone shows three distinct motifs. However, once optimized and ranked by BioOptimizer, the top two motifs are identical and matched the confirmed motif pattern in Harbison *et al.* (2004).

The foregoing two examples and other unreported ones demonstrate that BEST not only is easy and intuitive to use, but also improves TFBS prediction accuracy. BEST achieves these goals by enabling the biologist to simultaneously use, compare and combine the few most powerful motif-search programs available. An immediate future extension of our work is to include more motif-finding tools into the BEST framework.

## ACKNOWLEDGEMENTS

We thank the developers of each motif-finding program for allowing us to include their current versions in BEST. We also thank X Shirley Liu for helpful discussions. The work was partially supported by the NIH Grant R01-HG02518-01 and the NSF grant DMS-0244638.

## REFERENCES

- Bailey, T.L. and Elkan, C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, Stanford, CA. AAAI Press, Bethesda, MD, pp. 28–36.
- Galas, D.J. and Schmitz, A. (1978) A DNA footprinting: a simple method for the detection of protein–DNA binding specificity. *Nucleic Acids Res.*, **5**, 3157–3170.
- Garner, M.M. and Revzin, A. (1981) A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the *Escherichia coli* lactose operon regulatory system. *Nucleic Acids Res.*, **9**, 3047–3060.
- Harbison, C.T. *et al.* (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.
- Hertz, G.Z. and Stormo, G.D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, **15**, 563–577.
- Jensen, S.T. and Liu, J.S. (2004) BioOptimizer: a Bayesian scoring function approach to motif discovery. *Bioinformatics*, **20**, 1557–1564.
- Jensen, S.T. *et al.* (2004) Computational discovery of regulatory binding motifs: a Bayesian perspective. *Statist. Sci.*, **19**, 188–204.
- Lawrence, C.E. *et al.* (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.
- Liu, J.S. *et al.* (1995) Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *J. Am. Statist. Assoc.*, **90**, 1156–1170.
- Liu, X. *et al.* (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac. Symp. Biocomput.*, **6**, 127–138.
- Liu, X. *et al.* (2002) An algorithm for finding protein–DNA interaction sites with applications to chromatin immunoprecipitation microarray experiments. *Nat. Biotechnol.*, **20**, 835–839.
- Olman, V. *et al.* (2003) CUBIC: identification of regulatory binding sites through data clustering. *J. Bioinform. Comput. Biol.*, **1**, 21–40.
- Roth, F.R. *et al.* (1998) Finding DNA regulatory motifs within unaligned non-coding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.*, **16**, 939–945.