BIOINFORMATICS APPLICATIONS NOTE
Vol. 28 no. 14 2012, pages 1931–1932
doi:10.1093/bioinformatics/bts293

Sequence analysis
Advance Access publication May 15, 2012

MotifSuite: workflow for probabilistic motif detection and assessment

Marleen Claeys[1], Valerie Storms[1], Hong Sun[1,2], Tom Michoel[3] and Kathleen Marchal[1,4,*]

[1]Department of Microbial and Molecular Systems, Katholieke Universiteit Leuven, Kasteelpark Arenberg 20, [2]Department of Electrical Engineering, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, 3001 Heverlee, Belgium, [3]School of Life Sciences—LifeNet, Freiburg Institute for Advanced Studies (FRIAS), University of Freiburg, Albertstr. 19, 79104 Freiburg, Germany and [4]Department of Plant Biotechnology and Bioinformatics (VIB), Ghent University, Technologiepark 927, B-9052 Ghent, Belgium

Associate Editor: Alex Bateman

ABSTRACT

Motivation: Probabilistic motif detection requires a multi-step approach going from the actual *de novo* regulatory motif finding up to a tedious assessment of the predicted motifs. MotifSuite, a user-friendly web interface streamlines this analysis flow. Its core consists of two post-processing procedures that allow prioritizing the motif detection output. The tools offered by MotifSuite are built around the well-established motif detection tool MotifSampler and can also be used in combination with any other probabilistic motif detection tool. Elaborate guidelines on each of its applications have been provided.

Availability: http://homes.esat.kuleuven.be/~bioi_marchal/MotifSuite/Index.htm

Contact: kamar@psb.ugent.be

Received on February 22, 2012; revised on April 23, 2012; accepted on May 10, 2012

1 INTRODUCTION

Probabilistic methods, which search *de novo* for statistically overrepresented motifs in co-regulated genes, have been proven successful for the prediction of regulatory motifs. Due to the presence of local optima in the search space of possible overrepresented motifs, different initializations of a deterministic algorithm, such as MEME (Bailey *et al.*, 2006) or different runs of a stochastic algorithm, such as MotifSampler (Thijs *et al.*, 2002a) will output non-identical motif predictions even when performed under identical parameter settings. A tedious post-processing is required to extract from this set of multiple candidate predictions, the most significant ones. MotifSuite streamlines this multi-step approach from *de novo* motif detection to the assessment of motif significance.

2 MOTIFSUITE

MotifSuite (Fig. 1) guides users through the procedure of probabilistic motif detection. It consists of six different applications, each with an own entry page where input files are uploaded and user parameters are defined. The applications can be used separately or the whole analysis flow can be run at once with consecutive applications being compatible.

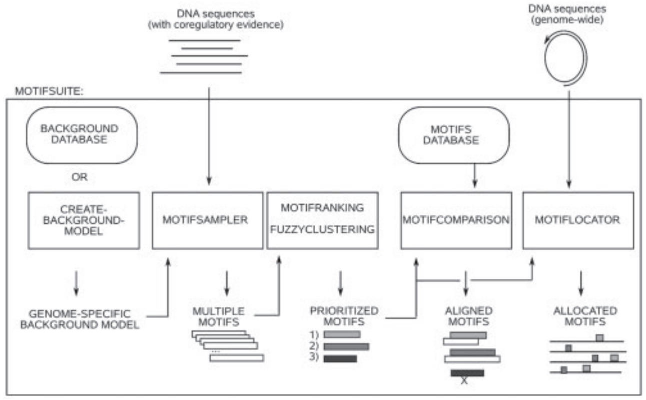

Fig. 1. Overview of the MotifSuite applications: CreateBackgroundModel, MotifSampler, MotifRanking, FuzzyClustering, MotifComparison and MotifLocator. The arrows point out the default integrated use of our applications

2.1 MotifSuite applications

A higher order background model (required for motif detection and scanning) can be selected from our database (Marchal *et al.*, 2003) or created with 'CreateBackgroundModel' (Thijs *et al.*, 2001). *De novo* motif detection in co-regulated sequences is executed by multiple Gibbs sampling runs in 'MotifSampler' (Thijs *et al.*, 2002a, b). Compared with the previous release of MotifSampler, the current release provides extensions that allow better approximating the true number of motif instances in each sequence of a given sequence set. Each of the multiple motifs reported by MotifSampler is represented by a set of instances and is summarized by a position weight matrix (PWM). Probabilistic motif detection offers the advantage of prioritizing the most reliable motif predictions by performing a significance analysis on the combined output of multiple motif detection runs. To this end, two applications are provided: (1) 'MotifRanking' sorts a list of predicted motifs (PWMs) by their motif model score and groups together PWMs that represent the same candidate motif. The representatives of each group (i.e. the motif with the highest score in the group) are prioritized by their motif model score and the number of times they re-occurred among the multiple motif detection runs (i.e. only high scoring motifs that are representatives of a large group of similar PWMs are retained);

*To whom correspondence should be addressed.

and (2) 'FuzzyClustering' evaluates predicted motifs at their instance instead of motif model level. Ensemble motifs obtained by merging instances of multiple motif detection runs have been shown to more accurately describe true motifs than any of the individual motif solutions (Newberg *et al.*, 2007; Reddy *et al.*, 2007). In FuzzyClustering, subsets of instances that were frequently detected together in multiple motif detection runs are grouped together into cluster(s) (Joshi *et al.*, 2008). A cluster represents an ensemble motif from which spurious instance predictions have been removed and in which instances are prioritized according to their membership score. 'MotifComparison' computes the PWM similarity to curated motif models from precompiled or user-supplied databases to analyze whether detected motifs correspond to any known motifs. Besides the original similarity metric (KL), the current release of MotifComparison provides an alternative similarity metric [p-BLiC, based on Habib *et al.* (2008)] that assigns more importance to similarity in motif positions that differ significantly from the genomic background assuming that such positions contribute most to the sequence-specific binding of a motif. 'MotifLocator' uses the PWM of a detected motif to screen (genome-wide) DNA sequences for potential novel motif instances.

2.2 Optimal use

To encourage the user to fully exploit the potential of our applications, we provide elaborated guidelines explaining the basic design of each application, the impact of its parameters and how to optimally evaluate its output. For most datasets, running our applications in the default workflow and at default parameter settings provides a reasonable answer. For particular datasets (e.g. with a different number of instances in different sequences), tuning parameter settings improves the detection of true motifs or at least provides a more accurate description of the detected motif. We provide elaborated case studies demonstrating the use of MotifSuite on 43 *Escherichia coli* sets of co-regulated sequences containing known motifs (Gama-Castro *et al.*, 2008). The MotifSampler case study shows, for example, that using a non-default setting for the expected number of instances per sequence (Statements 6 and 7) allows the detection of some motifs (four cases) that were missed under default setting and provided a more accurate prediction of the number of true instances for another set of five motifs. Another example is the use of the p-BLiC metric in MotifComparison instead of the default metric based on mutual information (used in MotifRanking) to assess the similarity to detected motifs (MotifRanking case study, Statement 2; Table 3a).

The case studies also show the complementarity between MotifRanking and Fuzzyclustering in post-processing the results of multiple MotifSampler runs. MotifRanking is best suitable to quickly assess whether a dataset contains any significantly overrepresented motifs (MotifRanking case study, Statement 1), whereas FuzzyClustering is better in retrieving the more reliable instances of a detected motif (FuzzyClustering case study, Statement 2). Alternatively, FuzzyClustering can be used to summarize the results of running MotifSampler at different parameter settings. Employing this scenario is, for example most suitable to find motifs having different unknown motif lengths in different sequences (FuzzyClustering will report instances of different lengths and build a single consensus PWM of most optimal motif length).

3 DISCUSSION

Conclusively, MotifSuite replaces INCLUSive (Coessens *et al.*, 2003; Thijs *et al.*, 2002b) which offered online access to the first release of MotifSampler. MotifSuite not only offers an improved release of MotifSampler but also provides a set of complementary methods for prioritizing and comparing motifs obtained by multiple runs of the Gibbs sampling tool. Because of their modular structure, the applications provided by MotifSuite can be used in combination with any probabilistic motif detection tool other than MotifSampler.

ACKNOWLEDGEMENT

We thank G. Thijs for his valuable contributions in several MotifSuite applications.

Funding: KU Leuven funding: (GOA/08/011, CoE EF/05/007); Agentschap voor Innovatie door Wetenschap en Technologie (IWT): (SBO-BioFrame); Interuniversity Attraction Poles (IAP) (P6/25); Fonds Wetenschappelijk Onderzoek-Vlaanderen (FWO): (IOK-B9725-G.0329.09, 5) and Ghent University: (Multidisciplinary Research Partnership 'M2N').

Conflict of Interest: none declared.

REFERENCES

Bailey,T.L. *et al.* (2006) MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.*, **34**, W369–W373.

Coessens,B. *et al.* (2003) INCLUSive: a web portal and service registry for microarray and regulatory sequence analysis. *Nucleic Acids Res.*, **31**, 3468–3470.

Gama-Castro,S. *et al.* (2008) RegulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Res.*, **36**, D120–D124.

Habib,N. *et al.* (2008) BLiC : A Novel Bayesian DNA Motif Comparison Method for Clustering and Retrieval. *PLoS Comput. Biol.*, **4**, e1000010.

Joshi,A. *et al.* (2003) Analysis of a Gibbs sampler method for model based clustering of gene expression data. *Trends Microbiol.*, **11**, 61–66.

Newberg,L.A. *et al.* (2007) A phylogenetic Gibbs sampler that yields centroid solutions for cis regulatory site prediction. *Bioinformatics*, **23**, 1718–1727.

Reddy,T.E. *et al.* (2007) Binding Site Graphs : a new graph theoretical framework for prediction of transcription factor binding sites. *Nucleic Acids Res.*, **35**, e20.

Thijs,G. *et al.* (2001) A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics*, **17**, 1113–1122.

Thijs,G. *et al.* (2002a) A Gibbs Sampling method to detect over-represented motifs in upstream regions of co-expressed genes. *J. Comput. Biol.*, **9**, 447–464.

Thijs,G. *et al.* (2002b) INCLUSive: Integrated Clustering, Upstream sequence retrieval and motif Sampling. *Bioinformatics*, **18**, 331–332.