MotifSuite: workflow for probabilistic motif detection and assessment

Marleen Claeys1, Valerie Storms1, Hong Sun1,2, Tom Michael3 and Kathleen Marchal1,4,*

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

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.


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

© The Author 2012. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com
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.

REFERENCES


Downloaded from https://academic.oup.com/bioinformatics/article-abstract/28/14/1931/219479 by guest on 05 August 2018