A comprehensive evaluation of module detection methods for gene expression data

Wouter Saelens WOUTER.SAELENS@UGENT.BE
VIB Inflammation Research Center, Ghent, Belgium

Department of Internal Medicine, Ghent University, Ghent, Belgium

Robrecht Cannoodt ROBRECHT.CANNOODT@UGENT.BE
VIB Inflammation Research Center, Ghent, Belgium

Center for Medical Genetics, Ghent University Hospital, Ghent, Belgium

Bart N. Lambrecht BART.LAMBRECHT@UGENT.BE
VIB Inflammation Research Center, Ghent, Belgium

Department of Internal Medicine, Ghent University, Ghent, Belgium

Yvan Saeys YVAN.SAEYS@UGENT.BE
VIB Inflammation Research Center, Ghent, Belgium

Department of Internal Medicine, Ghent University, Ghent, Belgium

Keywords: module detection, gene expression, clustering, biclustering

Abstract

Module detection is central in every analysis of large scale gene expression data. While numerous methods have been developed, the relative merits and drawbacks of these different approaches is still unclear. In this work we use known gene regulatory networks to do an unbiased comparison of 41 module detection methods, spanning clustering, biclustering, decomposition, direct network inference and iterative network inference. This analysis showed that decomposition methods outperform current clustering methods. Our work provides a first comprehensive evaluation to guide the biologist in their choice but also serves as a guideline for the evaluation of novel module detection methods.

Appearing in *Proceedings of Benelearn 2016*. Copyright 2016 by the author(s)/owner(s).

1. Introduction

Module detection methods form a cornerstone in the analysis of genome wide gene expression compendia. Modules in this context are defined as groups of genes with a similar expression profile, and therefore frequently share certain functions, are co-regulated and cooperate to produce a certain phenotype.

Over the last years, dozens of module detection methods have been developed, which can be classified in five different categories. The most popular method is undoubtedly clustering, which will group genes into modules based on global similarity in expression profiles. Within the transcriptomics community these methods have received a considerable amount of criticism. This is mainly due to three drawbacks: (i) clustering cannot detect so called local co-expression effects, (ii) most clustering methods are unable to detect overlapping modules and (iii) clustering methods do not model the underlying gene regulatory network. Alternative approaches have therefore been developed which either handle both overlap and local co-expression (biclustering and decomposition) or model the gene regulatory network (direct network inference and iterative

network inference).

Given this methodological diversity, it is important that existing and new approaches are evaluated on robust and objective benchmarks. However, evaluation studies in the past were limited in the number of methods, use unrealistic synthetic data or do not correctly assess the balance between false positives and false negatives. In this study we therefore provide a novel unbiased and comprehensive evaluation strategy, and used it to evaluate 42 state-of-the-art module detection methods.

2. Methods

The key of our approach is that we use golden standard regulatory networks to define sets of known modules. These can be used to directly assess the sensitivity and specificity of the different module detection methods. We used four different large scale gene expression compendia, two from E. coli, two from S. cerevisae and three from human samples. For each of these organisms a substantial part of the regulatory network is already known, either based on the integration of small-scale experiments or based on large, genome wide datasets. We use these networks to define groups of known modules using by looking at genes which either share one regulator, all regulators or are strongly interconnected. We used four different metrics to compare a set of observed modules with known modules: recovery and recall control the type II errors, while the relevance and specificity control the type I errors. Because of the large number of false positive and false negative edges within the human regulatory networks, we used an alternative scoring system on the human data based on the coverage of known transcription factors enrichment in the observed modules.

Parameter tuning is a necessary but often overlooked challenge of module detection methods. As default parameters of a tool are usually optimized for some specific test cases by the authors, they do not necessarily reflect general good performance on other datasets. On the other hand, one should be careful of overfitting parameters on specific characteristics of the data, as such parameters will lead to suboptimal results when using the same parameter settings on other datasets. In this study we first optimized parameters using a grid-based approach. Next, to avoid overfitting we used the optimal parameters on one dataset to score the performance on another dataset, in an approach akin to cross-validation.

3. Results

We evaluated 42 different module detection methods covering all five approaches. Overall, our analysis showed that certain decomposition methods, those based on the independent component analysis, outperform current state-of-the-art clustering methods. However, despite their theoretical advantages, neither biclustering nor network inference methods are able to outperform clustering methods. Importantly, our results are stable across datasets, module definitions and scoring metrics, demonstrating the robustness of our evaluation methodology.

We next used the evaluation workflow to further investigate several practical issues of module detection methods: parameter estimation, data requirements and alternative similarity measures. We tested several measures that can be used to find the most optimal parameter setting, and found that cluster validity indices only perform well when used together with non-overlapping clustering methods, while measures based on functional enrichment can effectively estimate parameters across a wide range of methods. Next, we tested the performance of the methods at different number of samples and found that although decomposition methods outperform clustering methods in large datasets, they have a smaller relative performance when the number of samples is limited. We finally tested several alternative similarity measures, which are able to better handle outliers, non-linearity and/or inverse relationships between two genes. Our evaluation workflow however does not indicate a clear advantage of any of the measures tested over the ubiquitously used Pearson correlation.

The applications of our work are twofold. First, if local co-expression and overlap are of interest, we discourage the use of biclustering methods and suggest the use of decomposition instead. Secondly, we provide a new comprehensive evaluation methodology which can be used to compare novel methods with the current state-of-the-art.

Acknowledgments

W.S. and R.C. are supported by an FWO aspirant grant.