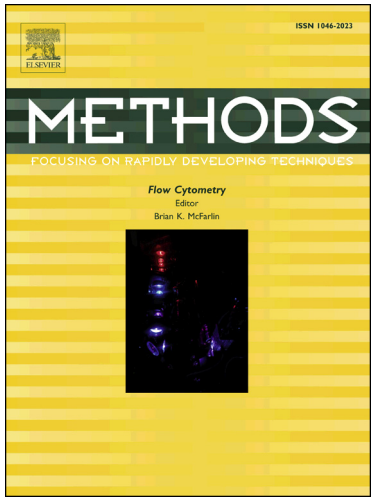

Network module identification—a widespread theoretical bias and best practices

Iryna Nikolayeva, Oriol Guitart-Pla, Benno Schwikowski

PII: S1046-2023(17)30037-3
DOI: https://doi.org/10.1016/j.ymeth.2017.08.008
Reference: YMETH 4301

To appear in: *Methods*

Received Date: 23 April 2017
Revised Date: 14 August 2017
Accepted Date: 18 August 2017

Please cite this article as: I. Nikolayeva, O. Guitart-Pla, B. Schwikowski, Network module identification—a widespread theoretical bias and best practices, *Methods* (2017), doi: https://doi.org/10.1016/j.ymeth.2017.08.008

Network module identification—a widespread theoretical bias and best practices

Iryna Nikolayeva, Oriol Guitart-Pla, Benno Schwikowski

Abstract

Biological processes often manifest themselves as coordinated changes across modules, i.e., sets of interacting genes. Commonly, the high dimensionality of genome-scale data prevents the visual identification of such modules, and straightforward computational search through a set of known pathways is a limited approach. Therefore, tools for the data-driven, computational, identification of modules in gene interaction networks have become popular components of visualization and visual analytics workflows. However, many such tools are known to result in modules that are large, and therefore hard to interpret biologically.

Here, we show that the empirically known tendency towards large modules can be attributed to a statistical bias present in many module identification tools, and discuss possible remedies from a mathematical perspective. In the current absence of a straightforward practical solution, we outline our view of best practices for the use of the existing tools.

Keywords:
Subnetwork identification, pathway, modules, algorithms, jActiveModules, size bias, extreme value distribution

1. Introduction

The organisation of cells is thought to be inherently modular [1, 2]. Modules can be identified from high-dimensional, genome-wide datasets. Typically, in a first step, gene-wise scores—often obtained from a statistical test—are calculated. These scores reflect the degree of involvement of each gene in a biological process. In a second step, one tries to identify gene modules from plausible sets of candidates, based on their scores.