Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category

Systems Biology

MoDentify: phenotype-driven module identification in metabolomics networks at different resolutions

Kieu Trinh Do[1], David J.N.-P. Rasp[1], Gabi Kastenmüller[2,3], Karsten Suhre[4], and Jan Krumsiek[1,2,5*]

[1]Institute of Computational Biology, Helmholtz Zentrum München, Neuherberg, Germany, [2]German Center for Diabetes Research (DZD), Neuherberg, Germany, [3]Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum, Neuherberg, Germany, [4]Department of Physiology and Biophysics, Weill Cornell Medical College - Qatar, Education City, Doha, Qatar, [5]Institute for Computational Biomedicine, Englander Institute for Precision Medicine, Department of Physiology and Biophysics, Weill Cornell Medicine, New York, USA

*To whom correspondence should be addressed.

Abstract

Summary: Associations of metabolomics data with phenotypic outcomes are expected to span functional modules, which are defined as sets of correlating metabolites that are coordinately regulated. Moreover, these associations occur at different scales, from entire pathways to only a few metabolites; an aspect that has not been addressed by previous methods. Here we present *MoDentify*, a free R package to identify regulated modules in metabolomics networks at different layers of resolution. Importantly, *MoDentify* shows higher statistical power than classical association analysis. Moreover, the package offers direct interactive visualization of the results in Cytoscape. We present an application example using complex, multifluid metabolomics data. Due to its generic character, the method is widely applicable to other types of data.

Availability and Implementation: https://github.com/krumsieklab/MoDentify (vignette includes detailed workflow)

Contact: jak2043@med.cornell.edu

Supplementary Information: Supplementary materials are available at *Bioinformatics* online.

1 Introduction

Associations with clinical phenotypic outcomes in large-scale metabolomics datasets are complex. They typically span entire modules, which are defined as groups of correlating molecules that are functionally coordinated, coregulated, or generally driven by a common biological process (Mitra *et al*, 2013). The systematic identification of modules is often based on networks, where the aim is to identify highly connected parts containing nodes that are coordinately associated with a given phenotype. Systematic module identification algorithms are well established (Polanski *et al*, 2014; Chuang *et al*, 2007; May *et al*, 2016; Martignetti *et al*, 2016); however none of the previously published methods consider that phenotype associations can occur at different scales, ranging from global associations spanning entire pathways or even sets of pathways ("dense" associations, e.g., between metabolites and phenotypic traits such as gender or BMI), to localized associations with only a few metabolites ("sparse" associations, e.g., between metabolites and phenotypic traits such as insulin-like growth-factor I (IGF-I) levels or asthma) (Do *et al*, 2017). For sparse associations, the identification and interpretation of modules is usually straightforward. However, modules for dense phenotype associations at the metabolite level are challenging to interpret due to their overwhelming number. To facilitate interpretation, the plethora of information at the fine-grained metabolite level can be condensed to a hierarchically superordinate level, such as a pathway network (i.e., a network of pathways).

We have recently introduced a module identification algorithm for multifluid metabolomics data (Do *et al*, 2017), which has been successfully applied to IGF-I and gender as examples of sparse and dense phenotype associations, respectively. We here present *MoDentify*, a free R package implementing the approach for general use. *MoDentify* offers network inference, module identification, and interactive module visualization at different levels of resolution. In particular, it increases statistical power compared with classical association analysis and can easily be applied to any type of quantitative data due to its generic character.

2 Description

MoDentify identifies network-based modules that are highly affected by a given phenotype. The underlying network is either directly inferred from the data at the single metabolite or pathway level (see below) or can be provided from an external source. Any external network can be used for the module identification procedure. This includes (1) networks obtained from public databases such as KEGG (Kanehisa *et al*, 2012) or Recon3D (Brunk *et al*, 2018), (2) networks inferred from statistical approaches such as partial, Pearson, or Spearman correlations, or (3) networks produced by newly emerging hybrid prior-knowledge / data-based approaches (e.g., Zuo *et al*, 2017). Regardless of the source of the network, all nodes in the network must be measured in the given dataset. Details can be found in the Supplementary Information.

Network inference: *MoDentify* estimates Gaussian graphical models, which have been shown to reconstruct metabolic pathways from metabolomics data (Krumsiek *et al*, 2011). At the fine-grained level, the network consists of nodes corresponding to metabolites, while at the pathway level, the nodes correspond to entire pathways (sets of metabolites). Such pathway definitions are available from public databases such as the Human Metabolome Database (HMDB) (Wishart *et al*, 2007), MetaCyc (Caspi *et al*, 2014), KEGG (Kanehisa *et al*, 2012), or Recon3D (Brunk *et al*, 2018). Edges represent significant (partial) correlations between two nodes after multiple testing correction.

Pathway representation: To build a network of interacting pathways, new variables are defined as representatives for each pathway, aggregating the total abundance of metabolites from the pathway into a single value. *MoDentify* implements two approaches: (1) *eigenmetabolite* approach, where the first principal component (*eigenmetabolite*) from a Principal Component Analysis is used as a representative value (Langfelder and Horvath 2007); (2) *average* approach, where the pathway representative is calculated as the average of all z-scored metabolite concentrations in the pathway.

Module identification and scoring: Given a network, a scoring function, and a starting node (seed node) as initial candidate module, the algorithm identifies an optimal module by score maximization. To this end, candidate modules are extended along the network edges until no further score improvement can be achieved. The score of a candidate module is calculated as the negative logarithmized p-value obtained from a multivariable linear regression model with the candidate module as dependent and the phenotype and optional covariates as independent variables. The procedure is repeated for each node in the network as seed node. Overlapping optimal modules are combined into single modules in an optional consolidation step. The combined module is then re-evaluated by the scoring function.

If multiple resolution levels are available, each resolution level is represented by its own network and module identification is performed at each resolution level separately.

Module visualization: In addition to returning R data structures and producing flat-file results, *MoDentify* offers visualization of the identified modules within an interactive network in the open source software Cytoscape (Shannon *et al*, 2003) for external visualization.

Complexity and runtime: The algorithm has a complexity of $O(n^2)$, which will lead to quadratic runtime in the worst-case scenario of a fully connected network. In practice, we assume biological networks to be sparse, i.e., with constant neighborhood sizes, leading to an approximate complexity of $O(n)$. On a 64-bit Windows 8 system with Intel(R) Core(TM) i7-4600U CPU @ 2.10GHz, network inference took ~21 seconds, module identification ~100s, and module visualization ~48s for a network with 1524 nodes.

3 Application example

We demonstrate the easy usage of *MoDentify* on plasma, urine, and saliva metabolomics data from the Qatar Metabolomics Study on Diabetes (QMDiab, see Supplementary Information) (Mook-Kanamori *et al*, 2014), aiming to identify functional modules associated with type 2 diabetes (T2D). Pathway annotations were provided by Metabolon, Inc., the metabolomics platform on which metabolomics measurements were performed. The dataset is also available via https://doi.org/10.6084/m9.figshare.5904022.

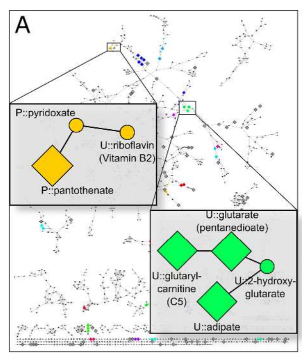
MoDentify was applied to the dataset at metabolite and pathway levels. To produce the list of T2D associated modules, as well as their interactive visualization in Cytoscape (Figure 1A), only three lines of code are required. Briefly, `generate.network` estimates partial correlations between metabolites, `identify.modules` searches network modules for the given phenotype, and `draw.modules` visualizes the results in Cytoscape.

```
###--- Load MoDentify
library(MoDentify)

###--- Network inference
met.graph <- generate.network(data = qmdiab.data, annotations = qmdiab.annos)

###--- Module identification
modules.summary <- identify.modules(graph = met.graph, data = qmdiab.data,
                                    annotations = qmdiab.annos,
                                    phenotype = qmdiab.phenos$T2D)

###--- Module visualization
draw.modules(graph = met.graph, summary = modules.summary)
```



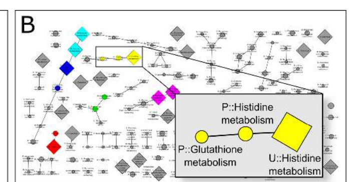


Figure 1. Visualization of identified modules for type 2 diabetes. The metabolomics networks with embedded modules at metabolite (**A**) and pathway (**B**) level are screenshots of the interactive versions in Cytoscape produced by *MoDentify*. Zoom-ins have been added to highlight examples for *MoDentify's* increased statistical power and its ability to extract biologically valuable insights. Round nodes correspond to metabolic entities not significantly associated with T2D when considered alone. Diamond nodes represent metabolic entities significantly related to T2D.

MoDentify identified 36 modules for T2D at the metabolite level (Figure 1A) and six modules at the pathway level (Figure 1B). Many of these modules consist of metabolites or pathways that are not significantly associated with T2D if considered alone. In combination, however, they form modules that are more associated with T2D than all of their single components. This increased statistical power in *MoDentify* can be attributed to the reduction of uncorrelated technical noise by aggregation of multiple metabolites and allows the detection of links with the phenotype that would have been missed with classical association analysis.

4 Conclusion

To the best of our knowledge, *MoDentify* implements the first approach for the systematic identification of phenotype-driven modules based on networks at different layers of resolution. The algorithm utilizes pathway definitions in combination with network topology to search for functional modules. Due to its increased statistical power, novel links between phenotypic outcomes and molecular levels can be detected that would be missed by classical analysis. We presented an application example using complex multifluid metabolomics data, but our approach can be applied for any quantitative dataset.

Acknowledgments

We thank the study participants and research team of the QMDiab study. The study was approved by the Institutional Review Boards of HMC and WCM-Q (protocol number 11131/11). All study participants provided written informed consent.

Funding

This work was supported by the German Federal Ministry of Education and Research (01ZX1313C), the European Union's Seventh Framework Program (305280), the National Institute of Aging (1RF1AG057452-01), the Qatar National Research Fund (NPRP8-061-3-011), and the Weill Cornell Medical College in Qatar.

Conflict of Interest: none declared.

References

Brunk,E. *et al*. (2018) Recon3D enables a three-dimensional view of gene variation in human metabolism. *Nat. Biotechnol.*, 36, 272–281

Caspi,R. *et al.* (2014) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res.*, 42, D459–D471

Chuang,H.Y. *et al.* (2007) Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.*, 3, 140

Do,K.T. *et al*. (2017) Phenotype-driven identification of modules in a hierarchical map of multifluid metabolic correlations. *NPJ Syst. Biol. Appl.*, **3**, 28

Kanehisa,M. *et al.* (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.*, 40, D109-114

Krumsiek,J. *et al*. (2011) Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC Syst. Biol*., **5**, 21

Martignett,L. *et al*. (2016) ROMA: Representation and quantification of module activity from target expression data. *Front. Genet*., **7**, 18

May,A. *et al*. (2016) Metamodules identifies key functional subnetworks in microbiome-related disease. *Bioinforma. Oxf. Engl*., **32**, 1678–1685

Mitra K, *et al*. (2013) Integrative approaches for finding modular structure in biological networks. *Nat. Rev. Genet.,* **14**, 719–732

Mook-Kanamori D.O., *et al*. (2014) 1,5-anhydroglucitol in saliva is a noninvasive marker of short-term glycemic control. *J. Clin. Endocrinol. Metab*., **99**, E479–E483

Polanski K, *et al*. (2014) Wigwams: identifying gene modules co-regulated across multiple biological conditions. *Bioinformatics* **30**, 962–970

Shannon, P. *et al*. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 13: 2498–2504

Wishart,D.S. *et al.* (2007) HMDB: the Human Metabolome Database. *Nucleic Acids Res.*, 35, D521–D526

Zuo,Y. *et al.* (2017) Incorporating prior biological knowledge for network-based differential gene expression analysis using differentially weighted graphical LASSO. *BMC Bioinformatics*, 18, 99