MILES : a Java tool to extract node-specific enriched subgraphs in biomolecular networks

Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Applications note

Data and text mining

MILES: a Java tool to extract node-specific enriched subgraphs in biomolecular networks

Pieter Moris [1,2], **Danh Bui-Thi** [1,2], **Kris Laukens** [1,2*] **and Pieter Meysman** [1,2,]

[1] Adrem Data Lab, Department of Mathematics and Computer Science, University of Antwerp, Antwerp, 2020, Belgium
[2] Biomedical Informatics Research Network Antwerp (biomina), University of Antwerp, Antwerp, 2020, Belgium

* To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: The growing availability of biomolecular networks has led to a need for analysis methods that are able to extract biologically meaningful information from these complex data structures. Here we present MILES (MIning Labeled Enriched Subgraphs), a Java-based subgraph mining tool for discovering motifs that are associated to a given set of nodes of interest, such as a list of genes or proteins, in biomolecular networks. It provides a unique extension to the widely used enrichment analysis methodologies by integrating network structure and functional annotations in order to discern novel biological subgraphs which are enriched in the targets of interest. The tool can handle various types of input data, including (un)directed, (un)connected and multi-label networks, and is thus compatible with most types of biomolecular networks.
Availability and implementation: MILES is available as a platform-independent Java application at https://github.com/pmoris/miles-subgraph-miner alongside a user manual, example datasets and the source code.
Contact: kris.laukens@uantwerpen.be
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Networks (or graphs) have become ubiquitous in biological research thanks to rapid advances in high-throughput technologies. Typical examples of biomolecular networks are protein-protein interaction and gene regulatory networks, but network representations are also often used to represent the structural formula of biomolecules or the amino acid sequence of proteins. Biomolecular networks offer a wealth of knowledge, but their interpretation within different contexts requires specialized methods.

Various algorithms and tools have been described in literature for subgraph analysis (see Mrzic *et al.* (2018) for an overview of bioinformatics-specific applications). They differ, among other things, in whether they deal with a single or multiple networks, and in their subgraph ranking methods. However, these methods mostly focus on identifying subgraphs as properties of the local structure of entire networks (Hočevar and Demšar, 2014; Przulj, 2007) or for inferring higher-order organization (Benson *et al.*, 2016), while it is also relevant to extract them as properties of sets of nodes within a network.

Enumerating subgraphs associated with a predefined set of nodes, is conceptually similar to a pathway or gene ontology enrichment analysis. Despite the prevalence of both molecular networks and enrichment analyses, no tools exist that can identify subgraphs enriched in gene or protein lists or in other biomolecular node sets.

To address this need, we have developed the Java application MILES (MIning Labeled Enriched Subgraphs), a subgraph enrichment tool that is based on the significant subgraph mining algorithm originally introduced in (Meysman *et al.*, 2016). It is a versatile tool that can handle (un)directed, (un)connected and multi-label networks, and is thus compatible with most types of biomolecular networks, ranging from regulatory gene networks to protein-protein interaction networks and many other situation where network data arises.

2 Application description

MILES aims to retrieve those subgraphs (or motifs) in a network that are significantly enriched in, or associated with, a set of nodes of interest. These nodes of interest might arise from experimental results (e.g. differentially

Table 1. Contingency table with occurrences for the enrichment test of the subgraph in Figure 1.

	Nodes of interest	Background nodes	
Source nodes	4	1	5
Non-source nodes	0	12	12
	4	13	17

expressed genes) or essentially any other method of delineating a subset of entities within the network based on certain properties (e.g. the protein receptors of a specific ligand, genes associated with a disease, residues with specific coordinates in a protein structure graph, etc.).

The first step in discovering enriched subgraphs is the generation of candidate subgraph patterns through a depth-first search.

These patterns can consist either of a purely topological structure (e.g. a feed-forward loop in a regulatory network) or they can include relevant biological annotation labels (e.g. a self-regulating transcription factor acted upon by a kinase).

Each subgraph pattern is then subjected to a hypergeometric (or equivalently a one-tailed Fisher's exact) test that compares the pattern's frequency among the nodes of interest to its frequency in the entire network. When significantly more than expected nodes of interest are at the basis of instances of a specific subgraph pattern, that subgraph is considered to be enriched in the nodes of interest. Note that for a subgraph to be associated with the subset of interest, it is not required to be completely embedded within it. Instead, each subgraph pattern contains a *source* node from which the pattern is gradually extended throughout the search, and it is the number of instances of a specific subgraph whose source nodes belong to the subset of interest that counts towards the enrichment statistic.

Finally, since a new hypothesis test is carried out for each distinct subgraph pattern, a multiple testing correction is applied to control the type I error rate.

A more in-depth overview of the underlying algorithm and its nuances is provided in the Supplementary Note. In the next section we present two example analyses, but several additional experimental applications are described in the online documentation and the original publication of the mining algorithm (Meysman *et al.*, 2016), including the search for ion-binding motifs in amino acid networks and the analysis of duplicated genes in *Saccharomyces cerevisiae*.

2.1 Toy example

The aspects described above are illustrated by the toy example in Figure 1. Here, a directed network is depicted, which could for example represent a gene regulatory network. The nodes would then represent transcription factors and the edges correspond to the regulatory influence they exert on each other. The shape of the nodes represents their labels (circles and squares for two different types of transcription factors). The light coloured nodes were selected as the subset of interest, e.g. because they were found to be differentially expressed in an experimental assay. In this situation, MILES will enumerate subgraph patterns, originating from the nodes of interest, and count their occurrences in the nodes of interest and in the network as a whole. One such subgraph is shown in the box: a feed-forward motif. This pattern appears four times in the network, and in all of these instances the pattern originates from a node of interest (Table 1). The resulting enrichment p-value is 0.013 (after Holm multiple testing correction), allowing us to reject the null hypothesis that there is no association between the nodes of interest and the subgraph pattern (for a significance level $\alpha = 0.05$, and conclude that this subgraph is significantly enriched in the subset of interest.

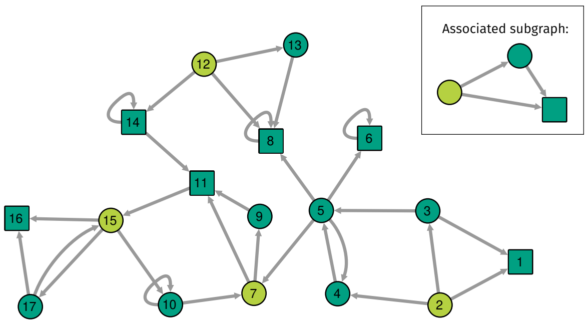

Fig. 1. A simple network to illustrate the discovery of enriched subgraphs. The node shapes represent labels and nodes of interest are shown in light green. An example of an enriched subgraph pattern, more specifically a feed-forward motif, is given in the inset.

2.2 A subgraph approach to annotation enrichment analysis

Searching for subgraphs enriched in certain genes or proteins offers an elegant extension of the widely utilised gene ontology and pathway enrichment analyses. Similar to the search for functional annotations that are enriched in a subset of biological entities (e.g. differentially expressed genes versus all genes measured in an expression assay), MILES recovers enriched labeled subgraphs from a network. By overlaying the network information (e.g. interacting protein partners or regulatory pathways) on top of the functionally labeled genes or proteins, there is more information available, in the form of the network structure and the combinations of annotations, to discern a specific subset of interest from its background.

This type of analysis was used in Bartholomeus *et al.* (2018), where gene expression measurements were taken in the context of a vaccination study. Proteins underlying differentially upregulated genes were marked as interesting in a human protein-protein interaction network, which was annotated with gene ontology terms pertaining to the immune system. The resulting enriched subgraphs revealed protein modules controlling immune effector processes and leukocyte activation (see Supplementary Note).

3 Implementation

The only inputs MILES requires are 1) a network, 2) a list of nodes of interest and optionally 3) node labels, all in the form of plain text files. The network file can consist of a single network or it can be made up of multiple distinct networks (or equivalently an unconnected network). Each node can either carry no, one (e.g. amino acids in a protein structure graph) or multiple labels (e.g. gene ontology labels in a protein-protein interaction network). If the label file is omitted entirely, MILES will only take the topology of the subgraphs into account.

Various additional analysis options can be set to adapt the algorithm to specific use cases. These include the possibility to specify a background reference set for the enrichment test, the choice of subgraph search algorithm (the original algorithm introduced by Meysman *et al.* (2016) or gSpan (Xifeng Yan and Jiawei Han, 2002)), a directionality toggle (e.g. edges in gene regulatory networks are directed), more liberal pruning of the search space (e.g. custom subgraph frequency thresholds) and a more stringent enrichment test that considers the subgraphs' parent patterns during the enrichment calculation (nested p-value approach). We refer to the Supplementary Note and the online documentation for a complete overview of all available options alongside examples.

The output of the mining analysis consists of a list of enriched subgraphs (i.e. those meeting the multiple testing adjusted p-value threshold of the hypergeometric tests), alongside their frequencies among the nodes of interest and in the network as a whole. In addition, an interactive visualisation in Cytoscape.js is provided (Franz *et al.*, 2015) from which the results can easily be exported for further post-processing or visualisation in external tools.

4 Availability

The MILES application, user documentation, example datasets and source code are available at https://github.com/pmoris/miles-subgraph-miner under the MIT license. MILES can be run from the command line or via a GUI on any operating system equipped with Java SE Runtime Environment 8 or higher.

5 Conclusion

MILES is a versatile tool tailored to uncover informative subgraphs in a network that are enriched in specific nodes of interest. It provides an extension to existing enrichment analysis methodologies, by integrating network information in order to discover novel and biologically relevant patterns. As such, it offers researchers an additional solution to analyse this increasingly prevalent complex type of biomolecular data.

Acknowledgements

This work was supported by the FWO [1141217N to PMo] and the University of Antwerp [BOF DOCPRO to DB]. Many thanks to Manuel Gerardo Orellana Cordero for his contributions to the Java code base.

Conflict of interest: none declared.

References

Bartholomeus, E., De Neuter, N., Meysman, P., Suls, A., Keersmaekers, N., Elias, G., Jansens, H., Hens, N., Smits, E., Van Tendeloo, V., Beutels, P., Van Damme, P., Ogunjimi, B., Laukens, K., and Mortier, G. (2018). Transcriptome profiling in blood before and after hepatitis B vaccination shows significant differences in gene expression between responders and non-responders. *Vaccine*, **36**(42), 6282–6289.

Benson, A. R., Gleich, D. F., and Leskovec, J. (2016). Higher-order organization of complex networks. *Science*, **353**(6295), 163–166.

Franz, M., Lopes, C. T., Huck, G., Dong, Y., Sumer, O., and Bader, G. D. (2015). Cytoscape.js: A graph theory library for visualisation and analysis. *Bioinformatics*, page btv557.

Hočevar, T. and Demšar, J. (2014). A combinatorial approach to graphlet counting. *Bioinformatics*, **30**(4), 559–565.

Meysman, P., Saeys, Y., Sabaghian, E., Bittremieux, W., van de Peer, Y., Goethals, B., and Laukens, K. (2016). Mining the Enriched Subgraphs for Specific Vertices in a Biological Graph. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, pages 1–1. 00000.

Mrzic, A., Meysman, P., Bittremieux, W., Moris, P., Cule, B., Goethals, B., and Laukens, K. (2018). Grasping frequent subgraph mining for bioinformatics applications. *BioData Mining*, **11**(1).

Przulj, N. (2007). Biological network comparison using graphlet degree distribution. *Bioinformatics*, **23**(2), e177–e183.

Xifeng Yan and Jiawei Han (2002). gSpan: Graph-based substructure pattern mining. In *Proceedings of the 2002 IEEE International Conference on Data Mining.*, pages 721–724. IEEE Comput. Soc.