Bioinformatics Advance Access published April 8, 2011

CyClus3D: a Cytoscape plugin for clustering network motifs in integrated networks

Pieter Audenaert[1*], Thomas Van Parys[2,3*], Florian Brondel[1], Mario Pickavet[1], Piet Demeester[1], Yves Van de Peer[2,3] and Tom Michoel[4†]

[1]Department of Information Technology, Ghent University – IBBT, Gaston Crommenlaan 8/201, B-9050 Gent, Belgium; [2]Department of Plant Systems Biology, VIB and [3]Department of Plant Biotechnology and Genetics, Ghent University, Technologiepark 927, B-9052 Gent, Belgium; [4]Freiburg Institute for Advanced Studies, University of Freiburg, Albertstraße 19, D-79104 Freiburg, Germany

Associate Editor: Dr. Joaquin Dopazo

ABSTRACT

Summary: Network motifs in integrated molecular networks represent functional relationships between distinct data types. They aggregate to form dense topological structures corresponding to functional modules which cannot be detected by traditional graph clustering algorithms. We developed CyClus3D, a Cytoscape plugin for clustering composite 3-node network motifs using a 3-dimensional spectral clustering algorithm.

Availability: Via the Cytoscape plugin manager or http://bioinformatics.psb.ugent.be/software/details/CyClus3D.

Contact: tom.michoel@frias.uni-freiburg.de

1 INTRODUCTION

In systems biology, the cell is modeled as an integrated network with multiple types of interactions, *e.g.* protein-protein, protein-DNA, protein-metabolite or genetic interactions (Zhu *et al.*, 2007). Cellular functions are carried out by independently functioning units called modules (Hartwell *et al.*, 1999), which, in graph-theoretical terms, correspond to clusters of densely connected nodes, and a multitude of algorithms have been developed to identify such clusters in undirected graphs (Fortunato, 2010). A major problem remains how to harness the multi-layered information contained in different interaction networks in order to identify biologically more realistic topological modules. In the naive Bayes approach, multiple interaction types are overlayed to create a single integrated association network which can be clustered by traditional means (Lee *et al.*, 2004). In the SAMBA approach, heterogeneous data types are merged in a single bipartite gene-property graph in which modules are defined as dense subgraphs (Tanay *et al.*, 2004). While SAMBA has the advantage of preserving the identity of each interaction type, information is inevitably lost by representing complex networks as bipartite graphs.

We developed CyClus3D, a Cytoscape (Shannon *et al.*, 2003) plugin for the identification of modules in integrated networks which uses network motifs to query a 3-dimensional spectral clustering algorithm. Network motifs are frequently occuring subgraphs in regulatory (Shen-Orr *et al.*, 2002) or integrated networks (Yeger-Lotem *et al.*, 2004; Yu *et al.*, 2006), which aggregate to form topological modules (Kashtan *et al.*, 2004; Zhang *et al.*, 2005). Each network motif defines a relationship between heterogeneous data types, with a distinct information-processing role or functional interpretation (Shen-Orr *et al.*, 2002; Zhang *et al.*, 2005; Zhu *et al.*, 2007). Hence, CyClus3D identifies modules composed of multiple interaction types which reflect regulatory, signaling or compensatory pathway mechanisms in addition to the stable protein complexes found by traditional clustering algorithms.

2 METHODS

2.1 Network motif clustering algorithm

We consider a system modeled by N types of pairwise interactions which may be directed or undirected. For a given 3-node network motif whose edges can be of any type, we denote the list of all motif instances as a 3-dimensional array T with $T_{ijk} = 1$ if the system contains a motif between nodes (i, j, k), and 0 otherwise. We define a motif cluster by three sets of nodes (X_1, X_2, X_3) with an aggregation score

$$\mathcal{S}(X_1, X_2, X_3) = \frac{\sum_{i\in X_1, j\in X_2, k\in X_3} T_{ijk}}{|X_1|^{1/p}|X_2|^{1/p}|X_3|^{1/p}}, \quad (1)$$

where $|X|$ is the number of nodes in X and $p > 1$ will act as an (inverse) resolution parameter. To maximize $\mathcal{S}$, we first determine the best rank-1 approximation to T, *i.e.* find real-valued vectors (x_1, x_2, x_3) maximizing

$$\mathcal{R}(x_1, x_2, x_3) = \frac{\sum_{ijk} T_{ijk}\, x_{1,i} x_{2,j} x_{3,k}}{\|x_1\|_p \|x_2\|_p \|x_3\|_p},$$

where $\|x\|_p = (\sum_i x_i^p)^{1/p}$ is the p-norm of x. A maximizer of $\mathcal{R}$ is found by solving the Euler-Lagrange equations

$$\lambda\, x_{1,i}^{p-1} = \sum_{jk} T_{ijk}\, x_{2,j} x_{3,k}, \quad (2)$$

subject to the constraint $\|x_1\|_p = 1$ and similarly for the other dimensions (De Lathauwer *et al.*, 2000). The solutions (x_1, x_2, x_3) are interpreted as cluster membership weight vectors and converted to a motif cluster by taking a suitable threshold on the weights. It can be shown that the optimal

*These authors contributed equally to this work

†To whom correspondence should be addressed

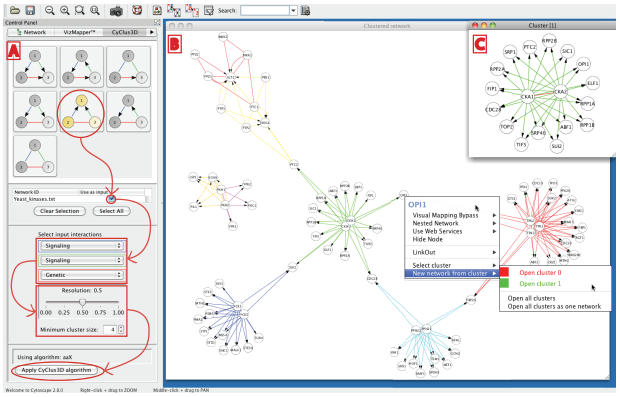

Fig. 1. CyClus3D screenshot with the workflow (A), a multi-cluster network with a node properties menu (B) and a single-cluster network (C).

threshold is the one which minimizes $\|x_m - u_{X_m}\|_p$, where $u_{X,i} = |X|^{-1/p}$ for $i \in X$ and 0 otherwise (see Supplementary Material). Having thus found a high-scoring motif cluster, we remove it from the list of motif instances T and repeat the procedure until no more instances remain. The best rank-1 approximation to the motif index array plays the same role as the dominant eigenvector of a network adjacency matrix and our algorithm can be understood as a generalization of 2-dimensional spectral clustering algorithms (Inoue and Urahama, 1999).

2.2 Implementation

Network motifs are often invariant under the permutation of some of their nodes. Thus, motif instances need to know their inherent symmetries, *e.g.* to efficiently determine the equality of two instances. We generated the motif symmetry groups offline and used a code generator to generate Java classes which are equipped with optimized methods for comparing and storing motifs. To locate all motif instances, we developed a motif finder which works on the principle of motif extensions. It allows quick pruning of branches in the search tree and is significantly faster than other subgraph matching algorithms (see Supplementary Material). To solve eqs. (2) we implemented a power algorithm (De Lathauwer *et al.*, 2000). The Java classes for network motif enumeration and clustering are independent of the Cytoscape visualisation classes and can be plugged into other network analysis and visualisation environments as well.

3 APPLICATION

To illustrate the workflow of CyClus3D (postfix for 3-Dimensional Clustering in Cytoscape), we imported an integrated network of physical, genetic and signaling interactions between kinases and phosphatases in yeast (Breitkreutz *et al.*, 2010; Fiedler *et al.*, 2009) (data available as Supplementary Material). In the CyClus3D control panel (Fig. 1A), a query motif and one or more input networks are selected, interaction types are assigned to each edge and a value for the resolution parameter $r = 1/p$ (cfr. Methods) and the minimal number of motif instances in a cluster are set. An edge type is inferred to be directed if the edge in the motif it is assigned to is directed. The resolution parameter allows to vary the typical size and density of a cluster. At low r, the aggregation score is maximized by large sets of loosely connected motifs, while at high values, high-scoring motif clusters are small and dense. In our experience, the intermediate value $r = 0.5$ balances size and density and is recommended as a starting value (see Supplementary Material).

After running the algorithm, CyClus3D opens a new network containing all clustered motifs. For instance, Fig. 1B shows all clusters of genetically interacting, copointing kinases (with the settings of Fig. 1A). By right clicking on a node of interest, we can create new networks for the clusters containing this node, while through the CyClus3D entry in the Plugins menu, new networks can be created for all clusters. By default, edges in multi-cluster networks are colored by their cluster membership ('Cluster View', Fig. 1B), while in single-cluster networks they are colored by interaction type, with the colors matching the edge assignments in the control panel ('Interaction View', Fig. 1C). Via the VizMapper panel, the user can easily switch between these two visual styles. Multiple motifs can be clustered sequentially and newly found clusters either are added to or replace the existing clustered network (to add them, all query motifs must be formed from subsets of the same three edge types and the Interaction View will be updated to the latest edge assignment).

By integrating heterogeneous types of molecular interaction data, CyClus3D identifies modules which reflect regulatory, signaling or compensatory functions which are not found by clustering each network in isolation (Zhang *et al.*, 2005). The underlying algorithms are highly efficient and allow further extension. In particular, future versions will extend CyClus3D towards higher-dimensional motifs, with applications in the domain of network alignment and comparison.

ACKNOWLEDGMENT

This research was supported by grants from the IWT (SBO-BioFrame), IUAP P6/25 (BioMaGNet) and Ghent University (MRP "Bioinformatics: from nucleotides to networks").

REFERENCES

Breitkreutz, A. *et al.* (2010). A global protein kinase and phosphatase interaction network in yeast. *Science*, **328**, 1043.

De Lathauwer, L. *et al.* (2000). On the best rank-1 and rank-($r_1, r_2, \ldots, r_n$) approximations of higher-order tensors. *SIAM J Matrix Anal Appl*, **21**, 1324.

Fiedler, D. *et al.* (2009). Functional organization of the *S. cerevisiae* phosphorylation network. *Cell*, **136**, 952.

Fortunato, S. (2010). Community detection in graphs. *Phys Rep*, **486**, 75.

Hartwell, L. H. *et al.* (1999). From molecular to modular cell biology. *Nature*, **402**, C47.

Inoue, K. and Urahama, K. (1999). Sequential fuzzy cluster extraction by a graph spectral method. *Pattern Recogn Lett*, **20**, 699.

Kashtan, N. *et al.* (2004). Topological generalizations of network motifs. *Phys Rev E*, **70**, 031909.

Lee, I. *et al.* (2004). A probabilistic functional network of yeast genes. *Science*, **306**, 1555.

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

Shen-Orr, S. S. *et al.* (2002). Network motifs in the transcriptional regulation network of escheria coli. *Nat Genet*, **31**, 64.

Tanay, A. *et al.* (2004). Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *PNAS*, **101**, 2981.

Yeger-Lotem, E. *et al.* (2004). Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction. *PNAS*, **101**, 5934.

Yu, H. *et al.* (2006). Design principles of molecular networks revealed by global comparisons and composite motifs. *Genome Biol*, **7**, R55.

Zhang, L. V. *et al.* (2005). Motifs, themes and thematic maps of an integrated *Saccharomyces cerevisiae* interaction network. *J Biol*, **4**, 6.

Zhu, X. *et al.* (2007). Getting connected: analysis and principles of biological networks. *Genes & Dev*, **21**, 1010.