BIOINFORMATICS APPLICATIONS NOTE

Vol. 00 no. 00 2005
Pages 1–2

CFinder: Locating cliques and overlapping modules in biological networks

Balázs Adamcsek[1], Gergely Palla[2], Illés J. Farkas[2], Imre Derényi[1], and Tamás Vicsek[1,2,*]

[1]Department of Biological Physics, Eötvös University, [2]Biological Physics Research Group of the Hungarian Academy of Sciences, Pázmány P. stny. 1A, H-1117 Budapest, Hungary

ABSTRACT

Summary: Most cellular tasks are performed not by individual proteins, but by groups of functionally associated proteins, often referred to as modules. In a protein assocation network modules appear as groups of densely interconnected nodes, also called communities or clusters. These modules often overlap with each other and form a network of their own, in which nodes (links) represent the modules (overlaps). We introduce CFinder, a fast program locating and visualizing overlapping, densely interconnected groups of nodes in undirected graphs, and allowing the user to easily navigate between the original graph and the web of these groups. We show that in gene (protein) association networks CFinder can be used to predict the function(s) of a single protein and to discover novel modules. CFinder is also very efficient for locating the cliques of large sparse graphs.

Availability: CFinder (for Windows, Linux, and Macintosh) and its manual can be downloaded from http://angel.elte.hu/clustering.

Contact: cfinder@angel.elte.hu

1 INTRODUCTION

High-throughput experimental techniques, e.g., protein-protein interaction (PPI) and mRNA expression methods, have largely advanced our knowledge about the functioning of the cell. *Gene (protein) association networks* integrate the broadest possible set of evidence – including high-throughput data – on protein linkages: they provide an integrated *list of binary interactions* (von Mering *et al*., 2005; Salwinski *et al*., 2004) and allow the discovery of previously uncharacterised cellular systems (Date and Marcotte, 2003). One major goal of current research efforts is to elucidate how the observed behaviours of an entire cell can be understood in terms of the interactions of its protein modules. To identify such modules, a common approach is to search for groups of densely interconnected nodes in the cell's protein association network (Bader and Hogue, 2003; Rives and Galitski, 2003). Note, however, that modules strongly overlap. According to the CYGD database (Guldener *et al*., 2005), in *Saccharomyces cerevisiae* the number of proteins in known protein complexes (modules where the participating proteins physically interact at the same time) vs. the sum of the sizes of these complexes is $2750/8932$. Thus, most protein modules probably share many of their proteins with other modules.

We introduce CFinder, a platform-independent, stand-alone application locating overlapping groups of densely interconnected nodes in graphs, and illustrate its use on the network of gene associations in the yeast genome. We decided to maintain CFinder as an independent program (as opposed to a package plugin), because it can be employed by potential users belonging to diverse fields including, in addition to bioinformatics, economics or sociology.

*To whom correspondence should be addressed.

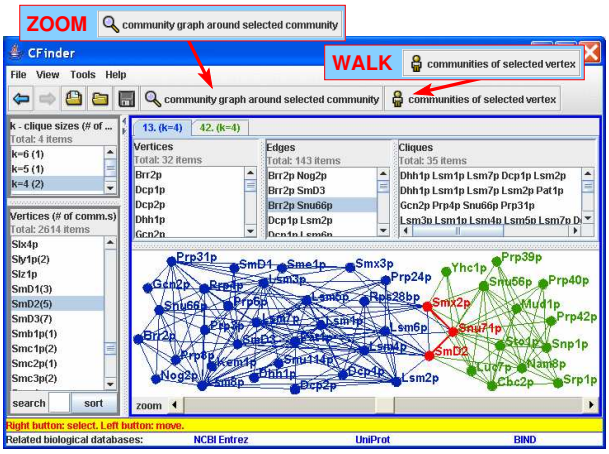

Fig. 1. (colour online) Modules of the protein SmD2 in the DIP (Database of Interacting Proteins) "yeast core" data set as shown by the *Vertices* view of CFinder. The two modules are coloured blue and green. Overlaps, i.e., proteins and links participating in more than one module are red. Enlarged on the top are two special buttons enabling the navigation between the original network (a part of which is displayed) and the web of its modules.

Generic graph visualisation and analysis programs (Batagelj and Mrvar, 1998) are frequently used for the layout and structural analysis of networks. Recent bioinformatics software platforms (Shannon *et al*., 2003), on the other hand, enable the user to integrate many different types of data, e.g., PPI, expression levels, and annotation information. CFinder reads a list of binary interactions, performs a search for dense subgraphs (groups), and – unlike several currently used algorithms (Newman, 2004) – it allows for any node to belong to more than one group. Due to its algorithm and implementation, CFinder is efficient for networks with millions of nodes and, as a byproduct of its search, the full clique overlap matrix of the network is determined. Below we will show that in gene association networks CFinder's results can be used to predict novel modules and novel individual protein functions.

2 OVERVIEW OF CFINDER

The *input of CFinder* is a file containing strings and numbers ordered into three columns; in each row the first two strings correspond to the two end points of a link and the third item is the weight of this link.

The computational core of CFinder was implemented in C++, while the visualisation and analysis components were written in Java. The *search algorithm* uses the Clique Percolation Method (CPM, see Derényi *et al*., 2005) to locate the *k-clique percolation clusters* of the network that we interpret

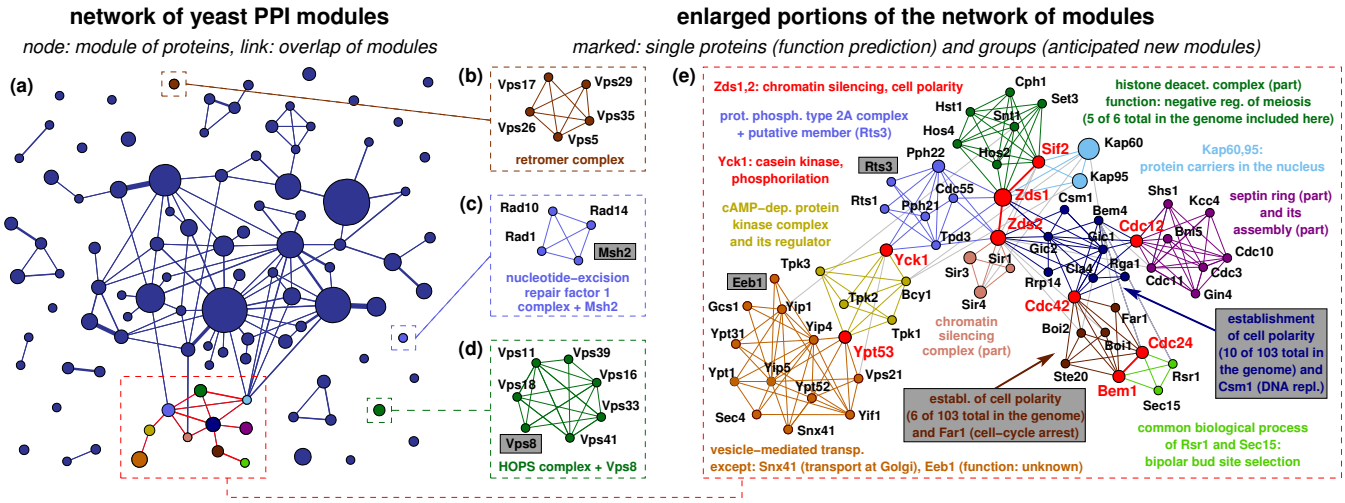

Fig. 2. (colour online) **(a)** The network of modules mapped by CFinder in the DIP "yeast full" data set ($k = 4$). **(b-d)** In addition to locating known complexes, CFinder often groups together a known complex with one additional protein, allowing the improvement of the functional annotation of that protein (Msh2, Vps8). **(e)** Zooming into the network of modules and adding Gene Ontology (GO) annotation terms (i) produces a *detailed and well-structured layout* of the original network of proteins, (ii) provides *characterisation for individual proteins* (Eeb1, Rts3) and (iii) *predicts new modules* (dark blue and brown, see text).

as modules. A k-clique is a complete subgraph on k nodes ($k = 3, 4, \ldots$), and two k-cliques are said to be adjacent, if they share exactly $k - 1$ nodes. A k-clique percolation cluster consists of (i) all nodes that can be reached via chains of adjacent k-cliques from each other and (ii) the links in these cliques. Note that larger values of k correspond to a higher stringency during the identification of dense groups and provide smaller groups with a higher density of links inside them. For both local and global analyses in a network, we suggest using such a value of k (typically between 4 and 6) that provides the user with the richest group structure (see Palla *et al*., 2005). In the presence of link weights CFinder can apply lower and upper cutoff values to keep only the set of connections judged to be significant by the user.

The *user interface* of CFinder offers several views of the analysed network and its module structure. As an example, Fig. 1 shows the modules of the protein Pwp2 in the DIP "yeast core" network (Salwinski *et al*., 2004) at clique size $k = 4$. Alternative views currently available in CFinder are "Communities" (displaying the identified modules), "Cliques", "Stats" (statistics of, e.g., module and overlap sizes) and "Graph of communities". The special buttons "forward", "back", "zoom" and "walk" allow a quick navigation between the views. A wide variety of visualisation settings can be adjusted in the "Tools" menu.

Figure 2 displays the network of modules produced by CFinder ($k = 4$) in the DIP "yeast full" data set. In the complete map (a) each node represents a module, the area of a node is proportional to the number of proteins in the corresponding module, and the width of a link is proportional to the number of proteins shared by the two modules. Panel (b) shows a previously known complex identified by CFinder. Panels (c) and (d) both display a known complex grouped together with one additional protein (Msh2 and Vps8, respectively), leading to an improved functional annotation of that protein. In panel (e) Eeb1 (function currently unknown) is grouped together with proteins participating in vesicle-mediated transport, thus, *we predict this to be a key function of Eeb1*. Proteins in the marked dark blue and brown groups of panel (e) cooperate on the establishment of cell polarity, a function performed by a total of 103 proteins in the cell. (Please, see colour figure online.) We anticipate that these two groups are *biologically meaningful, novel modules* within that larger set of 103 proteins. [Gene names and annotations were handled with Perl tools, e.g., GO::TermFinder (Boyle *et al*., 2004).]

3 ACKNOWLEDGEMENTS

The authors acknowledge funding from the Hungarian Sci. Res. Fund, OTKA (Grants No. D048422 , F047203, and T049674).

REFERENCES

Bader,G.D. and Hogue,C.W.V. (2003) An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, **4**, 2.

Batagelj,V. and Mrvar,A. (1998) Pajek – program for large network analysis. *Connections*, **21**, 47-57.

Boyle,E.I. *et al*. (2004) GO::TermFinder–open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, **20**, 3710-3715.

Date,S.V. and Marcotte,E.M. (2003) Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. *Nat. Biotechnol.*, **21**, 1055-1062.

Derényi,I. *et al*. (2005) Clique percolation in random networks. *Phys. Rev. Lett.*, **94**, 160202.

Guldener,U. *et al*. (2005) CYGD: the Comprehensive Yeast Genome Database. *Nucleic Acids Res.*, **33**, D364-D368.

von Mering,C. *et al*. (2005) STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res.* **33**, D433-D437.

Newman,M.E.J. (2004) Detecting community structure in networks. *Eur. Phys. J. B*, **38**, 321-330.

Palla,G. *et al*. (2005) Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, **435**, 814-818.

Rives,A.W. and Galitski,T. (2003) Modular organization of cellular networks. *Proc. Natl. Acad. Sci. U S A*, **100**, 1128-1133.

Salwinski,L. *et al*. (2004) The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.*, **32**, D449-451.

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