

Chapter XVIII

Module Finding Approaches for Protein Interaction Networks

Tero Aittokallio
University of Turku, Finland

ABSTRACT

This chapter provides an overview of the computational approaches developed for exploring the modular organization of protein interaction networks. A special emphasis is placed on the module finding tools implemented in three freely available software packages, VisANT, Cytoscape and MATISSE, as well as on their biomedical applications. The selected methods are presented in the broader context of module discovery options, ranging from approaches that rely merely on topological properties of the underlying network to those that take into account also other complementary data sources, such as the mRNA levels of the proteins. The author will also highlight some current limitations in the measured network data that should be understood when developing and applying module finding methodology, and discuss some key future trends and promising research directions with potential implications for clinical research.

INTRODUCTION

Recent advances in experimental technologies and computational methods have made it possible to measure and predict protein-protein interactions on a global scale (see Chapters I-VI and Shoemaker & Panchenko, 2007a, b). While the large-scale interaction datasets can provide an unprecedented glimpse into the cellular mechanisms underlying the behavior of various biological systems, the increasing sizes and densities of the protein interaction networks available today pose also many challenging computational problems. In particular, the inherent complexity of most biological processes and the large number of possible interactions involved can make it difficult to interpret and mine the interaction networks only by eye, even with the help of sophisticated visualization and layout tools available. Software packages that implement computational tools for more advanced network data mining can facilitate the explorative network analysis by identifying the key players and their interactions that contribute to the cellular processes of interest. This may allow e.g. to pinpoint errors in experimentally or computationally derived interaction links, identify proteins directly involved in the particular process, and to formulate hypotheses for follow-up experiments. It should be realized, however, that even if the rapidly developing complex

network theory has successfully been applied to analysis of various social and technological networks, such as the Internet and computer chips, its impact on studying molecular interaction networks is still an emerging area of research and therefore all these methods should be considered as experimental. At the moment, the computational tools are best used together with network visualization and analysis software that enable interactive and fully-controlled mining of the complex protein interaction networks.

One of the most fundamental properties found in many biological networks is their modular organization (Hartwell et al., 1999). Consequently, the decomposition of large networks into a hierarchy of possible overlapping sub-networks (so-called *modules*) has become a principal analytical approach to deal with the complexity of large cellular networks (Barabási & Oltvai, 2004). In protein interaction networks, a functional module refers to a group of physically connected proteins that work together to carry out a specific cellular function in a particular spatio-temporal context. A large number of computational tools of increasing complexity have recently been developed for investigating the modular organization of interaction networks. These tools cannot only identify whether a given network is modular or not, but also detect the modules and their inter-relationships in the underlying network. By relating the found sub-networks with complementary functional genomics or proteomics data, such as gene expression profiles from genome-wide microarray experiments or protein abundance measurements from mass-spectrometry-based assays, it is also possible to identify a hierarchy of connected groups of components that show coherent expression patterns. Such functionally organized modules cannot only emphasize the biological meaning of the modules discovered, but also allow us to gradually focus on the active subsystems of particular interest (*active modules*), which can lead to concrete hypotheses about the regulatory mechanisms and pathways most important for the given process (Ideker et al., 2002). Moreover, these modules can subsequently be used in predictive modeling studies, with the aim of suggesting new biological hypotheses, such as unexplored new interactions or the function of individual components, or even distinguishing different biomedical phenotypes (*discriminative modules*).

While several computational approaches to discovering functional modules in protein interactions networks are available to molecular biologists, and new ones are constantly being developed by the methodologists, there is no general consensus on how to choose between the different strategies and definitions. The choice of the network mining approach is an important aspect which depends both on the network and questions under the analysis. Therefore, basic understanding of the graph-theoretic concepts and algorithms behind the software packages is required so that the specific questions can be addressed and biologically meaningful interpretations can be made. This chapter describes several computational strategies developed for investigating the modularity of protein interaction networks, together with their relative merits and potential pitfalls. Since this topic has witnessed great progress lately, only representative examples of the different approaches can be presented. Among the multitude of algorithms and methods available in the literature and Internet, I describe here those features and options available in three popular software tools, VisANT, Cytoscape and MATISSE, which support the discovery of functionally coherent network modules. These publicly available software were selected because they come with a convenient graphical user interface making the methods easily accessible also to the end-users without programming skills. A special emphasis is placed on four specific tools implemented in these software, which represent typical computational strategies in the wide spectrum of module finding options, ranging from approaches that are based on topological network properties, such as nodes with a particular connectivity level (e.g. VisANT) or sub-networks that are highly connected (MCODE), to more sophisticated approaches that take into account also other related data sources, such as significant differences or co-expression patterns in the mRNA levels of the proteins (jActiveModules and JACS, respectively).

BACKGROUND AND MOTIVATION

Many advanced network analysis software complement the visual network exploration with options that facilitate the characterization of topological and functional properties of the underlying network (comparison of the features available in the popular software tools can be found from Chapter XVI and Suderman & Hallett, 2007; Zhang et al., 2007; Cline et al., 2007). Typical topological properties include, e.g., the *degree distribution* for determining hub nodes, *clustering coefficient* for detecting dense neighborhoods, and *shortest paths* for finding

17 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:

www.igi-global.com/chapter/module-finding-approaches-protein-interaction/5572

Related Content

Genome Subsequences Assembly Using Approximate Matching Techniques in Hadoop

Govindan Rajaand U. Srinivasulu Reddy (2017). *International Journal of Knowledge Discovery in Bioinformatics* (pp. 83-97).

www.irma-international.org/article/genome-subsequences-assembly-using-approximate-matching-techniques-in-hadoop/190794

Genetic Code and Stochastic Matrices

Sergey Petoukhovand Matthew He (2010). *Symmetrical Analysis Techniques for Genetic Systems and Bioinformatics: Advanced Patterns and Applications* (pp. 91-114).

www.irma-international.org/chapter/genetic-code-stochastic-matrices/37898

Drug Optimization for Cystic Fibrosis Patients Based on Disease Pathways Crosstalk

Shuting Linand Yifei Wang (2021). *International Journal of Applied Research in Bioinformatics* (pp. 1-11).

www.irma-international.org/article/drug-optimization-for-cystic-fibrosis-patients-based-on-disease-pathways-crosstalk/267820

GO-Based Term Semantic Similarity

Marco A. Alvarez, Xiaojun Qiand Changhui Yan (2013). *Bioinformatics: Concepts, Methodologies, Tools, and Applications* (pp. 93-104).

www.irma-international.org/chapter/based-term-semantic-similarity/76058

Classifier Ensemble Methods for Diagnosing COPD from Volatile Organic Compounds in Exhaled Air

Ludmila Ilieva Kuncheva, Juan Jose Rodríguez, Yasir Iftikhar Syed, Christopher O. Phillipsand Keir Edward Lewis (2012). *International Journal of Knowledge Discovery in Bioinformatics* (pp. 1-15).

www.irma-international.org/article/classifier-ensemble-methods-diagnosing-copd/77927