Ph.D. Thesis

Overlapping module structure of biological networks: computer programs and their applications

Máté Szalay-Bekő

Supervisors:

Prof. Péter Csermely
Semmelweis University, Department of Medical Chemistry

Dr. Balázs Papp
Hungarian Academy of Sciences, Biological Research Center,
Institute of Biochemistry

Doctoral School of Biology
University of Szeged, Faculty of Science and Informatics

SZEGEDI TUDOMÁNYEGYETEM TERMÉSZETTUDOMÁNYI ÉS INFORMATIKAI KAR

Szeged
2013

Introduction

In the last fifteen years, the rapid development of high-throughput technologies and the increased amount of available experimental data enabled the emergence of systematic informatics algorithms and methods analyzing complex biological systems. Network science brings interesting and new aspects to the field of biology by constructing graph models for many of these systems, like signaling or regulation processes of the cells. In these graphs the nodes represent biological entities (such as proteins or genes), while the links are modeling physical interactions or logical associations between these entities.

The goal of the research work showed in my thesis is to develop computer programs helping to understand complex biological systems trough the structural analysis of biological network models.

The real natural systems (like biological, social or even economical systems) are too successful and robust to have a random structure. Therefore, one of the key challenges of network science is to analyze and understand the structure and topology of these networks. In case of the most of natural networks, the interacting nodes tend to form modules, dense regions. For example, in the protein interaction network of a cell such module can be the functional group of proteins responsible for forming ribosomes, or involved in RNA splicing. It is a hard algorithmic task to find these modules in large networks containing thousands or ten thousands of nodes. More than a hundred significantly different methods have been published to solve this problem (Fortunato, 2010). However, most of these methods assign the given node into only one given module. More details of the module structure are revealed by the overlapping modularization methods, which are capable to assign a given node into multiple modules. Fuzzy methods are producing even more information by also determining the strength of module assignment. **In my thesis, I present the computer implementation of the ModuLand fuzzy modularization algorithms and I show the application of these bioinformatics tools on practical biological examples.**

Aims of the work

In my research work I formulated the following goals.

1. My basic goal was to develop such an user-friendly modularization computer tool, which can help to answer practical biological questions by analyzing the fuzzy module structure of complex biological network models.
2. My aim was to prove the applicability of the tool by analyzing the fuzzy module structure of the amino acid network describing the *Escherichia coli* Met-tRNS synthetase enzyme.
3. My aim was to prove the applicability of the tool by comparing the module structure of the *Buchnera aphidicola* and *Escherichia coli* metabolic networks.
4. My aim was to prove the applicability of the tool by analyzing the yeast protein interaction network, and by comparing the modular positions of the date and party hubs, which proteins show interesting expression dynamics behavior.

Methods

During my research work, I used the ModuLand fuzzy modularization algorithms defined by Dr. István Kovács. We published these algorithms in a joint publication (Kovács *et al.*, 2010). I also reused the first Linux implementations of these algorithms we created together with Robin Palotai, Gábor Szuromi and Balázs Zalányi. My own work (Szalay-Bekő *et al.*, 2012) was mainly to optimize these algorithms, to create a user-friendly graphical interface and to apply the program on numerous biological example.

To implement an easy-to-use bioinformatics tool, I integrated the ModuLand method into the widely used open source Cytoscape (Shannon *et al.*, 2003) program, which provides a network visualization and analysis framework running on Windows, Linux and Mac operating systems. The

Cytoscape ModuLand plug-in[1] was published in the international Bioinformatics journal (Szalay-Bekő *et al.*, 2012). During the implementation I used C++ and Java programming languages, I introduced several optimizations to reduce the run-time of the modularization, and implemented several measure calculation and visualization features to help the analysis of biological networks.

The ModuLand Cytoscape plug-in enables the users to visualize and analyze the module hierarchy of any undirected network. The tool can determine the central regions of the modules and can calculate different measures describing the structural position of each node in the network. For example in a protein interaction network, the user can check the module centrality measure to understand how central is the given protein in the cell. A different measure gives the ability to highlight the bridge proteins connecting different functional modules in the protein interaction network. The numerous different module structure related measures and reports can be exported easily from the program (e.g. using Excel file format).

Results

The plug-in was successfully used for bioinformatics research projects many times by me, by the members of my research group and also by independent groups.

I used the plug-in to analyze the protein structure network of the Met-tRNA synthetase protein of the *Escherichia coli* bacteria (Szalay-Bekő *et al.*, 2012). The ModuLand plug-in revealed the interesting modular positions of the amino acids responsible for transferring the conformational changes between the catalytic center and the anticodon binding site of the enzyme. The five major sub-domains of Met-tRNA synthetase were well reflected by the five modules found by the ModuLand plug-in. Key amino acids of the most frequently used communication path either belonged to the module cores of the three

[1] The ModuLand plug-in can be downloaded from the following site: http://www.linkgroup.hu/modules.php

modules involved in transmission of conformational changes, or were special inter-modular nodes highlighted by the ModuLand plug-in.

Also with the help of the plug-in, together with my colleagues we compared the metabolic network of the symbiont *Buchnera aphidicola* and free-living *Escherichia coli* bacteria, and showed the differences of the module structures caused by the different levels of environmental stability. We found that the average size of modules and module overlaps are significantly higher in case of the *Buchnera* bacteria, while the *E. coli* module cores corresponded to significantly less metabolic functions than those of *Buchnera*. These results indicated that modules of the metabolic network of an organism from a variable environment (*E. coli*) are more specialized than metabolic network modules of an organism living in a constant environment (*Buchnera*). Our results are in agreement with earlier publications (Parter *et al.*, 2007; Mihalik and Csermely, 2011) where the authors also found more specialized modules in protein interaction or metabolic networks describing organisms experiencing more stress.

Using the ModuLand method family we analyzed (Kovács *et al.*, 2010) the yeast protein interaction network, where we were able to map main modules to different cell functions. We also analyzed the expression patterns of the protein interactions of hub proteins, having high number of connections. Based on the fuzzy module structure, we successfully separated the date hubs (typically expressing all of their connections together) from the party hubs (typically varying their connections in the different expression datasets). In this way we were able to predict expression dynamic behavior of proteins, analyzing only the statical module structure of protein interaction network.

Since I developed the ModuLand plug-in and the related publication in 2012, more than 150 researchers downloaded the program and used it for several internationally published research projects. For example, the ModuLand plug-in was used to analyze the system level effects of oxidative stress on yeast cells (Lehtinen *et al.*, 2013), while an other study used the plug-in to align protein interaction networks of different species based on their module structures (Wang and Gao, 2012). In a third publication (Sharma *et al.*, 2013) the authors used the ModuLand plug-in to analyze the 20 functional modules in the protein interaction network built around the Sirtuin enzyme family.

Based on these initial verifications of the use of the program, the ModuLand Cytoscape plug-in (and the other fuzzy modularization methods) will help to answer several interesting questions in the fields of biology and pharmacology in the future. Beside the previously mentioned examples, the plug-in can be used to predict cross-talk regions in signaling networks, to identify possible drug target proteins, or even to help in the structured visualization of complex biological systems.

List of own publications

2nd of November, 2013.[2]

Cumulative impact factors: 50

Independent citations: 187

Previous name used in publications before marriage: **Szalay, M.S.**

Publications related to the thesis

1. Szalay-Bekő, M., Palotai, R., Szappanos, B., Kovács, I.A., Papp, B., Csermely, P. (2012) ModuLand plug-in for Cytoscape: determination of hierarchical layers of overlapping modules and community centrality. *Bioinformatics*, **28**, 2202-2204, IF: 5.3 http://arxiv.org/abs/1111.3033 – 6 independent citations
2. Kovács, I.A., Palotai, R., Szalay, M.S., Csermely, P. (2010) Community landscapes: a novel, integrative approach for the determination of overlapping network modules. *PloS ONE* **7**, e12528, IF: 3.7 www.arxiv.org/abs/0912.0161 – 26 independent citations

[2] To calculate the impact factors and the number of independent citations, I used the *Web of Science* online tool. (http://www.webofknowledge.com)

Publications independent from the thesis

Original research papers

1. Fazekas, D., Koltai, M., Türei, D., Módos, D., Pálfy, M., Dúl, Z., Zsákai, L., Szalay-Bekő, M., Lenti, K., Farkas, I.J., Vellai, T., Csermely, P., Korcsmáros, T. (2013) SignaLink 2 - a signaling pathway resource with multi-layered regulatory networks. *BMC Systems Biology* **7**(1):7, IF: 3 – 1 independent citation
2. Korcsmáros, T. Szalay, M.S., Rovó, P., Palotai, R., Fazekas, D., Lenti, K., Farkas, I.J. Csermely, P., Vellai, T. (2011) Signalogs: orthology-based identification of novel signaling pathway components in three metazoans. *PLoS ONE* **8**, e19240, IF: 3.7 – 2 independent citations
3. Korcsmáros, T., Farkas, I.J., Szalay, M. S., Rovó, P., Fazekas, D., Spiró, Z., Böde, C., Lenti, K., Vellai, T., Csermely, P. (2010) Uniformly curated signaling pathways reveal tissue-specific cross-talks, novel pathway components, and drug target candidates. *Bioinformatics* **26**, 2042-2050, IF: 5.3 www.signalink.org -- 10 independent citations
4. Wang, S., Szalay, M.S., Zhang, C., Csermely, P. (2008) Learning and innovative elements of strategy update rules expand cooperative network topologies. *PLoS ONE* **3**, e1917, IF: 3.7 www.arxiv.org/0708.2707 -- 21 independent citations

Patents

1. Korcsmáros T., Szalay-Bekő M., Palotai R., Szuromi G., Fazekas D., Dunai Zs. (2011) Procedure and computer system to simulate effect of drug metabolites. Hungarian patent application, P1100368
2. Szalay, M., Stanojevic, O., Farkas, L. (2010) Automatic use of behavioral information for promotional purposes in communications system. International Ericsson PCT patent application, PCT/SE2010/051312

3. Kovács, I.A., Csermely, P., Szalay, M.S., Korcsmáros, T. (2006) Method for analyzing the fine structure of networks. International PCT patent application, PCT/IB2007/05047

Review papers

1. Farkas, I.J., Korcsmáros, T., Kovács, I.A., Mihalik, Á., Palotai, R., Simkó, G.I., Szalay, K.Z., Szalay-Bekő, M., Vellai, T., Wang, S., Csermely, P. (2011) Network-based tools in the identification of novel drug-targets. *Science Signaling* **4**, pt3, IF: 7.6 -- 3 independent citations
2. Palotai, R. Szalay, M.S., Csermely, P. (2008) Chaperones as integrators of cellular networks: changes of cellular integrity in stress and diseases. *IUBMB Life* **60**, 10-18, arxiv.org/0710.1622, IF: 2.8 -- 24 independent citations
3. Korcsmáros, T., Szalay, M.S., Böde. C., Kovács, I.A., Csermely, P. (2007) How to design multi-target drugs: Target-search options in cellular networks. *Expert Op. Drug Discov*. **2**, 799-808, arxiv.org/q-bio.MN/0703010, IF: 2.3 -- 20 independent citations
4. Böde. C., Kovács, I.A., Szalay, M.S., Palotai, R. Korcsmáros, T., Csermely, P. (2007) Network analysis of protein dynamics. *FEBS Lett.* **581**, 2776-2782, arxiv.org/q-bio.BM/0703025, IF: 3.6 -- 51 independent citations
5. Szalay, M.S., Kovács, I.A., Korcsmáros, T., Böde. C., Csermely, P. (2007) Stress-induced rearrangements of cellular networks: consequences for protection and drug design. *FEBS Lett.* **581**, 3675-3680, arxiv.org/q-bio.MN/0702006, IF: 3.6 -- 18 independent citations
6. Korcsmáros, T., Kovács, I.A., Szalay, M.S., Csermely, P. (2007) Molecular chaperones: the modular evolution of cellular networks. *J. Biosci.* **32**, 441-446, arxiv.org/q-bio.MN/0701030, IF: 1.8 -- 14 independent citations

7. Kovacs, I.A., Szalay, M.S., Csermely, P. (2005) Water and molecular chaperones act as weak links of protein folding networks: energy landscape és punctuated equilibrium changes point towards a game theory of proteins. http://arxiv.org/abs/q-bio.BM/0409030, *FEBS Lett.* **579**, 2254-2260, IF: 3.6 -- 21 independent citations

Independent papers cited in this thesis booklet

1. Fortunato,S. (2010) Community detection in graphs. *Physics Reports*, **486**, 75–174.
2. Lehtinen,S., Marsellach,F.X., Codlin,S., Schmidt,A., Clément-Ziza,M., Beyer,A., Bähler,J., Orengo,C., Pancaldi,V. (2013) Stress induces remodelling of yeast interaction and co-expression networks. *Molecular Biosystems*, **9**, 1697–1707.
3. Mihalik,Á., Csermely,P. (2011) Heat shock partially dissociates the overlapping modules of the yeast protein-protein interaction network: a systems level model of adaptation. *PLoS Computational Biology*, **7**, e1002187.
4. Parter,M., Kashtan,N., Alon,U. (2007) Environmental variability and modularity of bacterial metabolic networks. *BMC Evolutionary Biology*, **7**, 169.
5. Shannon,P., Markiel,A., Ozier,O., Baliga,N.S., Wang,J.T., Ramage,D., Amin,N., Schwikowski,B., Ideker,T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, **13**, 2498–2504.
6. Sharma,A., Costantini,S., Colonna,G. (2013) The protein-protein interaction network of the human Sirtuin family. *Biochim. Biophys. Acta*, **1834**, 1998-2009.
7. Wang,B., Gao,L. (2012) Seed selection strategy in global network alignment without destroying the entire structures of functional modules. *Proteome Science*, **10**, S16.

Acknowledgments

I thank all the help to the University of Szeged, to the doctoral school of Biology, to the PhD commission and to the organizers of my PhD defense.

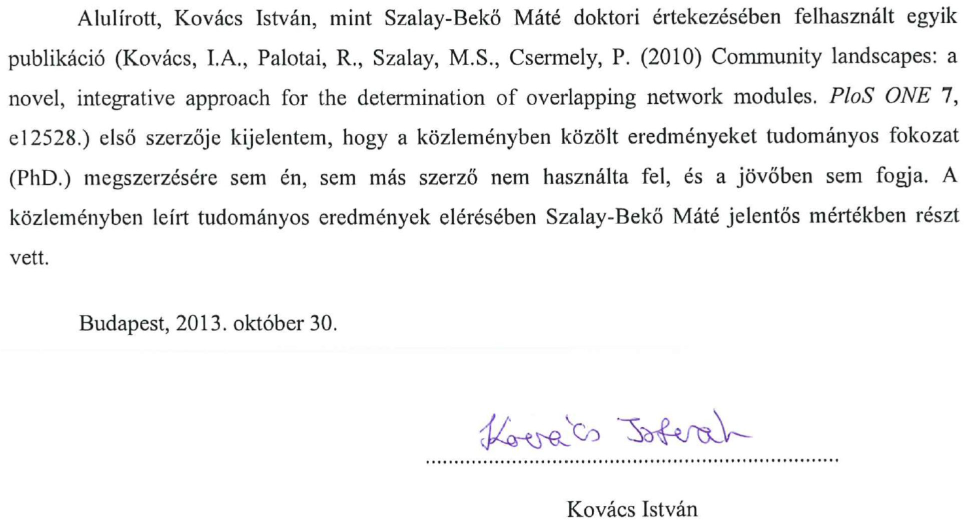
I also thank the help of my supervisors. I am especially thankful to **Professor Péter Csermely**, who was helping me in the last ten years in many different ways (as a president, mentor, chief, supervisor and especially as a friend). I also wish to thank to **Dr. Balázs Papp** all the helps regarding my research works and administrative obligations during my PhD.

I thank **Dr. Tamás Korcsmáros** for the numerous professional advices regarding my research work and my PhD defense. I also wish to thank the creative community and the many years of common thinking to the **LinkGroup research group** lead by Prof. Péter Csermely. Many of them helped me during the biological analyzes and informatics work described in my thesis. I thank all the help to **Dr. István Kovács**, who defined the algorithms of the ModuLand method family. I am also thankful to **Robin Palotai**, **Gábor Szuromi** and **Balázs Zalányi**, who helped to implement the first linux version of the ModuLand methods. I also thank to **Balázs Szappanos** the help regarding the comparison of the metabolic networks of *B. aphidicola* és *E. coli* bacteria.

I thank to the **Ericsson Hungary Ltd.** for the paid leave they provided me to finish my PhD studies and for taking over some parts of the related fees.

Last but mainly, I thank to my family for supporting me in my PhD studies during the last few years, and tolerating my overtimes I spent on research work beside my daily jobs. I thank the support to my parents, my siblings and especially to my wife and son, **Melinda** and **Márk**!

Statements of co-authors
(in Hungarian)



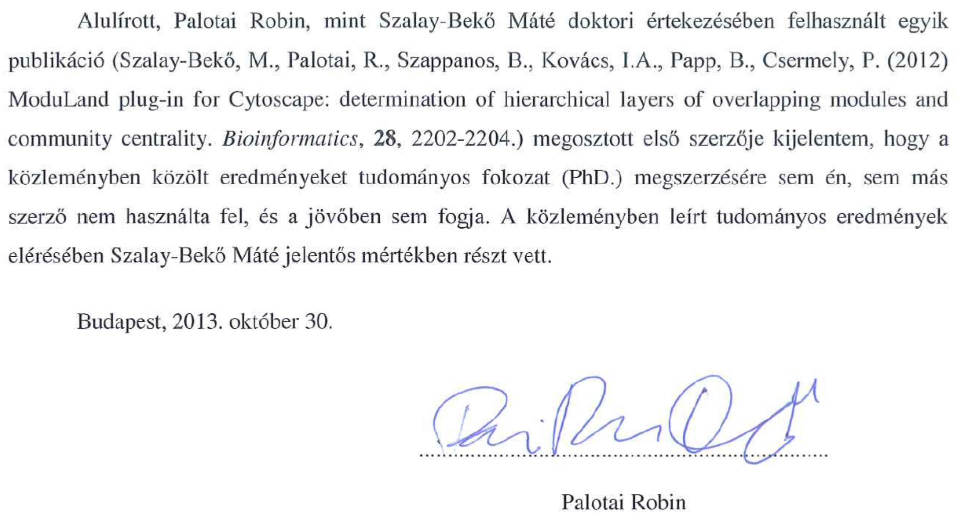
Alulírott, Kovács István, mint Szalay-Bekő Máté doktori értekezésében felhasznált egyik publikáció (Kovács, I.A., Palotai, R., Szalay, M.S., Csermely, P. (2010) Community landscapes: a novel, integrative approach for the determination of overlapping network modules. *PloS ONE* **7**, e12528.) első szerzője kijelentem, hogy a közleményben közölt eredményeket tudományos fokozat (PhD.) megszerzésére sem én, sem más szerző nem használta fel, és a jövőben sem fogja. A közleményben leírt tudományos eredmények elérésében Szalay-Bekő Máté jelentős mértékben részt vett.

Budapest, 2013. október 30.

Kovács István

....

Kovács István



Alulírott, Palotai Robin, mint Szalay-Bekő Máté doktori értekezésében felhasznált egyik publikáció (Szalay-Bekő, M., Palotai, R., Szappanos, B., Kovács, I.A., Papp, B., Csermely, P. (2012) ModuLand plug-in for Cytoscape: determination of hierarchical layers of overlapping modules and community centrality. *Bioinformatics*, **28**, 2202-2204.) megosztott első szerzője kijelentem, hogy a közleményben közölt eredményeket tudományos fokozat (PhD.) megszerzésére sem én, sem más szerző nem használta fel, és a jövőben sem fogja. A közleményben leírt tudományos eredmények elérésében Szalay-Bekő Máté jelentős mértékben részt vett.

Budapest, 2013. október 30.

....

Palotai Robin