Employing Power Graph Analysis to Facilitate Modeling Molecular Interaction Networks

Momchil Nenov[1], Svetoslav Nikolov[1,2*]

[1]*Department of Biomechanics*
Institute of Mechanics – Bulgarian Academy of Sciences
Acad. G. Bonchev Str., Bl. 4, 1113 Sofia, Bulgaria
E-mail: m.nenov@imbm.bas.bg

[2]*Department of Mechanics*
University of Transport
158 Geo Milev Str., 1574 Sofia, Bulgaria
E-mail: s.nikolov@imbm.bas.bg

[*]*Corresponding author*

Received: January 09, 2015 **Accepted: March 26, 2015**

Published: April 01, 2015

Abstract: *Mathematical modeling is used to explore and understand complex systems ranging from weather patterns to social networks to gene-expression regulatory mechanisms. There is an upper limit to the amount of details that can be reflected in a model imposed by finite computational resources. Thus, there are methods to reduce the complexity of the modeled system to its most significant parameters. We discuss the suitability of clustering techniques, in particular Power Graph Analysis as an intermediate step of modeling.*

Keywords: *Power graph analysis, Molecular interaction networks, Dynamic system.*

Introduction

Molecular interaction networks such as cell signaling pathways and gene expression regulatory networks frequently involve tens of individual elements in a dynamic relationship with each other. This makes a descriptive approach insufficient for their understanding and prediction. The complexity of such systems however can make mathematical modeling time and resource intensive.

In computational systems biology a fundamental problem is the construction of biochemical reaction system models that can effectively predict cellular behavior. Given the high level of redundancy inherent to biological networks, an intermediate step which employs a clustering algorithm to reduce redundancy in the network would provide a consistent method to reduce the complexity of the model as a whole, esp. in non-trivial cases.

A potential candidate for such a method is *power graph* (PG) analysis, which introduces both lossless clustering algorithm and a new visualization paradigm to highlight functional inter-dependencies in molecular networks [2].

The term network here is used as a synonym for graph $G=(N, E)$, which can be defined as an order pair (N, E) of a set $N=\{n_0, n_1, ..., n_i\}$ of nodes and a set E of edges, note that each edge is an unordered pair of nodes, i.e. $E \subseteq \{\{x, y\} | x, y \in N\}$. Thus, two nodes $A, B \in N$ are adjacent if and only if there exists an edge $\{A, B\} \in E$, in which case A and B are neighbors.

A PG $G' = (N', E')$ is a graph defined on the power set $N' \subseteq P(N)$ of *power nodes* connected to each other by *power edges*: $E' \subseteq N' \times N'$ [8]. Hence PGs are defined on the power sets of nodes and edges of the graph *G*.

Recently, several examples of how Power Graph Analysis (PGA) can be used to visualize large biological networks (e.g., protein-protein interactions and gene co-expressions) have been documented in the published literature on systems biology and functional genomics. For example, Praneenararat and co-workers [7] developed NaviClusterCS, which enables researchers to interactively navigate large biological networks of ~100000 nodes in a "Google Maps-like" manner in the Cytoscape environment.

Similarly, numerous groups have used PGA to analyze network structure, and thereby to compare false positive and false negative noise levels in protein-protein interaction networks, as well as to identify functional modules in protein interaction networks [3, 6, 9]. The approach of PGA is useful because it provides a formal way to reduce the complexity of the molecular interaction networks and interpret new findings in the context of known biological processes.

A dynamic system can be thought of as a graph, where the vertices of the graph are its parameters and the *power edges* represent the interactions between them.

The simplest case of two vertices (*A* and *B*) with a *power edge* between them can be expressed with 2 ordinary differential equations:

$$\begin{aligned} \frac{dA}{dt} &= f(B), \\ \frac{dB}{dt} &= g(A). \end{aligned} \tag{1}$$

Three interconnected vertices would result in
$A - B, A - C, B - C$, i.e. (2)

$$\frac{dA}{dt} = f(B, C),$$

$$\frac{dB}{dt} = g(A, C), \tag{3}$$

$$\frac{dC}{dt} = h(A, B).$$

The number of equations equals the number of vertices, *n*, and the number of parameters per equation equals the number of edges connected to that vertex. The higher the density of the graph, the more computationally expensive it will be to model the system it describes, *worst case* being $O(n(n-1))$.

Redundancy is the presence of more than one path from vertex *A* to vertex *B*. High redundancy is a characteristic feature of biological networks among others. A reduction in redundancy leads to a reduction in the density of the graph and thus the complexity of the mathematical model.

There are two strategies to achieve that goal:

- Reducing the graph to a tree, by eliminating loops.
- Grouping several vertices together.

The first method is *lossy*, meaning data is actually lost during the transformation and is typically used to minimize storage space, although it might lead to an increase in search times. Feed forward/back loops are an important feature of biological networks and eliminating them would alter the nature of the system fundamentally, thus making it an unlikely candidate as a way to reduce complexity of the model without affecting its results, but it can be used for other purposes, such as reducing noise in order to find groups of interacting molecules, etc.

Alternatively, grouping *vertices* together (clustering) can be either *lossy* or *lossless*, depending on the particular algorithm. PGA is both a clustering algorithm and a way to represent graphs visually intended to make sense of biological networks [8]. It groups *vertices* into sets called *power nodes* based on common neighbors and is *lossless*. Individual *edges* are represented by links between sets of nodes and called *power edges*.

Two sets of vertices where each individual vertex is connected to all vertices in the other set is represented by an element called a *bi-clique* and visualized as 2 *power nodes* connected by a *power edge*. A set of interconnected vertices is called a *clique* and is visualized as a single *power node* with a *power edge* loop. A *star* is a special case of *bi-clique* where one of the sets consists of a single *vertex*.

Software

We created a software program as a test-bed to assess the suitability of Power Graph Analysis (and potentially other clustering algorithms) as a tool to reduce the complexity of biological systems and identify significant parameters for the purposes of mathematical modeling.

Features

The program consists of a command-line tool, written in C [http://www.imbm.bas.bg/lib/files.php], which can be sub-divided into 3 logical parts.

1. Input: Provides a framework for the actual parsers dealing with specific formats in the form of libraries. Constructs a graph in memory.
2. Clustering algorithm: Turns the above *graph* into a PG (as of the time of writing).
3. Output: Creates a visual representation of the *graph* and writes it as an image file.

Dependencies

- The Cairo graphics library is used for image generation.
- C99 compatible compiler is required to compile the sources.

Usage

- The program accepts line-delimited lists of tab-delimited ASCII tokens to its standard input. Support for more formats can be added through libraries.
- The output of the program is an image file in one of several formats.
- For command line options, see the documentation.

Todo

The tool itself is a proof-of-concept with rudimentary, at this point, features. Considering there are more mature alternatives available it remains to be seen whether our team (or someone else) will find it useful enough to continue its development. One area which can be improved is adding more parsers to deal with different data formats such as those produced by Producer Price Index (PPI) databases etc.

Alternatives and other similar tools

A command-line tool and a Cytoscale plug-in are provided by the Technical University of Dresden at [http://www.biotec.tu-dresden.de/research/schroeder/powergraphs].

Results

Two examples of the output from the program are shown in Figs. 1 and 2. As the first example, we use the autocatalytic (positive feedback) growth of a protein module (Myc and E2Fs), which is inhibited by miRNA – 17-92 [1]. It is known that the miRNA-17-92 cluster is a polycistronic gene located in human chromosome 13, ORF 25, located at 13q31-q32. The cluster consists of 7 mature miRNAs, namely, miRNA-17-5p, miRNA-17-3p, miRNA-18a, miRNA-19a, miRNA-20a, miRNA-19b, and miRNA-92-1. Gene expression data shows an over-expression of miRNA-17-92 in different tumors, including lung, colon, breast, prostate, stomach and pancreas cancer [4, 5].

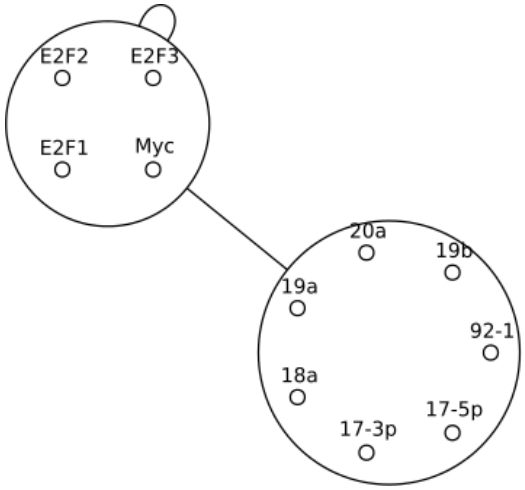


Fig. 1 Output of the program when fed the Myc/E2F/mir-17-92 network consisting of 11 elements as described in [1]. Result shows two sets, one self-referential clique, containing all proteins and another one, containing miRNAs. It is comparable to the model proposed in the source article by Aguda et al. [1] if the undirected nature of the PG is taken into account.

Fig. 1 demonstrates *cliques* and *bi-cliques*, but contains no nested *power nodes*, which we felt was an important feature that warranted its own example. The following image does not reflect an actual biological network, but one that was intended specifically to show nested *power nodes* while keeping the size of the resulting image suitable for embedding in this article.

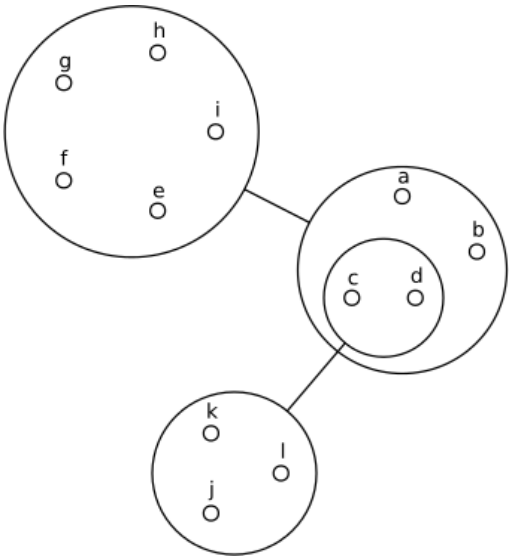


Fig. 2 Example of an abstract network of 12 original elements
showing nesting of *power nodes*

Conclusion

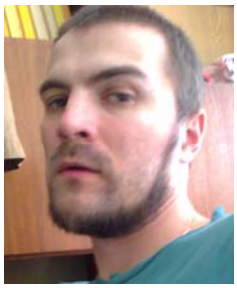
For the purposes of modeling biological networks, such as signaling pathways or gene expression regulatory networks the algorithm identifies groups of elements which can be abstracted, thus reducing the complexity of the model. However, an intermediate human-assisted step is needed to translate the input parameters from the uncompressed network into actual parameters for the PG model, because PGA has no notion of the nature of the interactions and it works with undirected graphs. This is not an easy task. A possible future feature of this program might be the additional ability to also take the nature of the interactions into account, in effect extending the paradigm of PGA.

References

1. Aguda B., Y. Kim, M. Piper-Hunter, A. Friedman, C. Marsh (2008). MicroRNA Regulation of a Cancer Network: Consequences of the Feedback Loops Involving miR-17-92, E2F and Myc, PNAS, 105(50), 19678-19683.
2. Ahners S. (2014). Generalised Power Graph Compression Reveals Dominant Relationship Patterns in Complex Networks, Scientific Reports, 4, 4385.
3. McGuffin M. (2012). Simple Algorithms for Network Visualization: A Tutorial, Tsinghua Science and Technology, 17(4), 383-398.
4. Nikolov S., J. Vera, U. Schmitz, O. Wolkenhauer (2011). A Modelling Based Strategy to Investigate the Role of MicroRNA Regulation in Cancer Signalling Networks, Theory in Biosciences, 130(1), 55-69.
5. Petrocca F., R. Visone, M. R. Onelli, M. H. Shah, M. S. Nicoloso, I. de Martino, D. Iliopoulos, E. Pilozzi, C. G. Liu, M. Negrini, L. Cavazzini, S. Volinia, H. Alder, L. P. Ruco, G. Baldassarre, C. M. Croce, A. Vecchione (2008). E2F1-regulated MicroRNAs Impair TGFbeta-dependent Cell-cycle Arrest and Apoptosis in Gastric Cancer, Cancer Cell, 13(3), 272-286.
6. Pinkert S., J. Schultz, J. Reichardt (2010). Protein Interaction Networks – More than Mere Modules, PLoS Computational Biology, 6(1), e1000659.
7. Praneenararat T., T. Takagi, W. Iwasaki (2012). Integration of Interactive, Multi-Scale Network Navigation Approach with Cytoscape for Functional Genomics in the Big Data Era, BMC Genomics, 13(Suppl. 7), S24.

8. Royer L., M. Reimann, B. Andreopulos, M. Schroeder (2008). Unravelling Protein Networks with Power Graph Analysis, PloS Computational Biology, 4, e1000108.
9. Royer L., M. Reimann, A. Stewart, M. Schroeder (2012). Network Compression as a Quality Measure for Protein Interaction Networks, PLoS ONE, 7(6), e35729.

Assist. Momchil Nenov, Ph.D. Student
E-mail: m.nenov@imbm.bas.bg



Momchil Nenov received a M.Sc. degree in Genetics from St. Kl. Ohridski University in Sofia, Faculty of Biology in 2010. Currently, he is an Assistant in the Department of Biomechanics at the Institute of Mechanics – Bulgarian Academy of Sciences. His interests include modeling and simulation of biological systems, specifically cell signaling and expression control pathways.

Assoc. Prof. Svetoslav G. Nikolov, D.Sc., Ph.D.
E-mail: s.nikolov@imbm.bas.bg

Svetoslav Nikolov's research and educational interests are in the fields of mathematical modelling, nonlinear dynamics and bifurcation analysis of systems in Cell biology. He received a M.Sc. degree in Mechanical Engineering from the Technical University of Sofia, Bulgaria, in 1994, a Ph.D. degree – from the Institute of Mechanics and Biomechanics – Bulgarian Academy of Sciences, in 1999, and a D.Sc. degree – from the University of Transport, Sofia, in 2013. Since 2005 he has been an Associate Professor at the Institute of Mechanics and Biomechanics and since 2012 – an Associate Professor at the University of Transport, Sofia, Bulgaria.