Author's response to reviews

Title: Using graph theory to analyze biological networks

Authors:

Georgios A Pavlopoulos (pavlopou@embl.de)
Charalampos N Moschopoulos (mosxopul@ceid.upatras.gr)
Theodoros Soldatos (maria.secrier@embl.de)
Maria Secrier (thsoldat@googlemail.com)
Sophia Kossida (skossida@bioacademy.gr)
Reinhard Schneider (rschneid@embl.de)
Pantelis G Bagos (pbagos@ucg.gr)

Version: 3 Date: 18 February 2011

Author's response to reviews: see over
Dear Editors,

Please find attached our revised manuscript entitled: “Using graph theory to analyze biological networks”. We would like to thank the reviewers and the editors for their time and effort in evaluating this manuscript and for their valuable contributions to the improvement of this work. Modifications have been made according to the reviewers’ suggestions, as described below.

Reviewer 1

Minor Essential revisions addressed:

It seems that special characters are missing in the prepared version. Section "Data Structures" first paragraph "which requires ....", Section "Network Models" first paragraph "have degree k is .....", Section "Cluster Analysis" in description of Median linkage, in Discussion in second paragraph "it has been shown that .....". Section "Cluster Analysis" expression "Minkowski metric" "K" change to "k" in the summation operator. Also, it is not a good practice to have "d" for dimensionality and in the same expression "d(i,j)" for distance.

The reviewer noted a series of changes that should be made in order to have more formal and accurate descriptions of formulas and concepts. Special characters have been introduced accordingly and some characters have been changed as suggested:

In the section "Data Structures", in the Adjacency matrix paragraph, "which requires \( n=|V| \)" has been replaced by “which requires \( \Theta(|V|) \)” (page 10).

In the “Discussion” section, in the second paragraph for the formulation “it has been shown that \( \frac{N_{\text{reg}}}{N_{\text{tot}}} N \) for prokaryotes and \( \frac{N_{\text{reg}}}{N_{\text{tot}}} N^{0.3} \) for eukaryotes”, the following has been added for clarification of what N is: “where N is the network size” (page 35).
In the section "Cluster Analysis" expression "Minkowski metric", "K" has been changed to "k" in the summation operator. "d" standing for dimensionality was changed to “D”, such that it doesn't cause confusion with the distance measure "d(i,j)" in the same expression. The formula now looks like this:

\[ d(i,j) = \left( \sum_{k=1}^{D} (x_{i,k} - x_{j,k})^p \right)^{1/p} \]  

(page 28).

**Discretionary revisions addressed:**

In Section "Cluster Analysis" description of Markov Clustering is not clear enough in order to be understood and used by non-experts. The same comment is true also for the application of Pearson's correlation coefficient for detecting assortative networks in Section "Network topology". The work is a good review of the graph methods but illustrative examples, in which they have been successfully applied, especially with details of their applications, are missing. This fact strongly limits the usefulness of the work.

The reviewer was concerned about the lack of clarity in the description of Markov Clustering. The reviewer is right in pointing this out and we appreciate the chance to reformulate the explanation of the concept in a clearer manner, as described below (more references to relevant papers have also been added, pages 31-32):

**Markov Clustering** [134] (MCL) algorithm is a fast and scalable unsupervised clustering algorithm based on simulation of stochastic flow in graphs. The MCL algorithm can detect cluster structures in graphs by a mathematical bootstrapping procedure which takes into account the connectivity properties of the underlying network. The process deterministically computes the probabilities of random walks through a graph by alternating two operations: expansion, and inflation of the underlying matrix. The principle behind it is that random walks on a graph are likely to get stuck within dense subgraphs rather than move between dense subgraphs via sparse connections. In other words, higher length paths are more often encountered between nodes in the same cluster than between nodes within different clusters, such that the probabilities between nodes in the same complex will typically be higher in expanded matrices. Clusters are identified by alternating expansion and inflation until the graph is partitioned into subsets so that there are no longer paths between these subsets [135, 136].

The same concern was expressed about the Pearson correlation coefficient. We acknowledge that the description we had used could have confused the reader, so
A more straightforward way is to use the Pearson’s Correlation Coefficient (PCC), which quantifies the correlation or linear dependence between two variables (in this case, the degrees of two nodes). In other words, it measures to which extent one variable increases/decreases as the other increases. PCC (r-value) between two nodes is defined as the covariance of the two nodes divided by the product of their standard deviations. For the entire network, the assortativity coefficient is the measure of how assortative or disassortative a network is overall. If \( M \) is the number of edges, and \( x_i \) and \( y_i \) the degrees of the vertices at either ends of edge \( i \), the assortativity coefficient \( r \) is calculated as follows:

\[
r = \frac{M^{-1} \sum_i x_i y_i - \left( M^{-1} \sum_i \frac{1}{2} (x_i + y_i) \right)^2}{M^{-1} \sum_i \frac{1}{2} (x_i^2 + y_i^2) - \left( M^{-1} \sum_i \frac{1}{2} (x_i + y_i) \right)^2}, \text{ with } i = 1 \ldots M[102].
\]

This is equivalent to the Pearson correlation coefficient of the degrees at either ends of an edge. The range of the \( r \)-values is between +1 and −1, \( r < 0 \) corresponding to a disassortative network whereas \( r > 0 \) to an assortative one.

The lack of more illustrative examples of where graph theory has been applied in biology has been pointed out, and we have addressed this issue by including more examples in several sections of the paper, with corresponding references, as well as information about useful software tools, where appropriate, as it can be seen in the highlighted changes. Please see the response to reviewer 2, point 2 for more details on this.

**Reviewer 2**

1. **There are numerous small language issues that should be addressed to improve readability; for example within the Metabolic and Biochemical Networks section:**
   "occurring within a cell in different time points" should be "at different time points"
   "the main role in metabolic network is played..." should be "the main role within a metabolic network"
   "Often, enzymes are dependend on other cofactors" should be "Often, enzymes are dependant on other cofactors"
1. The reviewer pointed out a series of small language issues, providing suggestions of how to reformulate some statements in order to improve the readability of the text. We have altered our text accordingly in the sections mentioned:
    "occurring within a cell in different time points" was changed to "at different time points" (page 4)
    "the main role in metabolic network is played... " was changed to "the main role within a metabolic network" (page 5)
    "Often, enzymes are dependend on other cofactors" was changed to "Often, enzymes are dependant on other cofactors" (page 5)
Many additional typos have been corrected and some rephrasing has been performed to improve readability, as it can be seen in the revised version.

2. Much of the article explains graph theory and concepts, many of which could be found in text books on graph and network theory. I believe the utility of this article could be dramatically improved by adding one or both of the following aspects:

   A. Expand each section with literature references to specific examples of the graph theory concepts as they apply to biological data. For example, cliques have been used to identify new functional groups from gene expression data. Citing examples where the concept is used to ask a specific question would help readers relate these concepts to their own work.

2. The reviewer suggested expanding the graph theory sections with examples of applications to biological data or reference to software tools and techniques related to the described concepts. We have added such information in various sections of the paper.

   A. Several sections have been expanded with references to papers where graph theory concepts are applied:

In “Graph Theory and Definitions”: 
**Bipartite graph** is an undirected graph $G = (V,E)$ in which $V$ can be partitioned into 2 sets $V_1$ and $V_2$ such that $(u,v) \in E$ implies either $u \in V_1$ and $v \in V_2$ OR $v \in V_1$ and $u \in V_2$. Applications of this type of graph to visualization or modelling of biological networks range from representation of enzyme-reaction links in metabolic pathways to ontologies or ecological connections, as discussed in [61] or [62]. (page 8)

The total connectivity of a network is defined as $C = \frac{E}{N(N-1)}$ where $E$ is the number of edges and $N$ the total number of nodes. The connectivity structure of biological networks is often informative with respect to reaction interplay and reversibility, compounds that structure the network, like in metabolism, or trophic relationships, like in food-web networks. Such connectivity profiles can be detected based on mixture models using software like MixNet [63]. (page 9)

In “Network properties”:
- for graph density:
  
  **Dense** is a graph where $|E| \approx |V|^2$. It has been argued that biological networks are generally sparsely connected, as this confers an evolutionary advantage for preserving robustness. This has been observed for a series of organisms: the transcriptional regulatory networks of *S. cerevisiae*, *E. coli*, *D. melanogaster* all have connectivity densities lower than 0.1 [64]. (page 12)

- for cliques:

  Detection and analysis of these structures has found many biological applications: identifying groups of consistently co-expressed genes in microarray datasets, finding cis regulatory motifs or matching three-dimensional structures of molecules [68, 69]. (pages 14-15)

- for clustering coefficient:

  Biological networks have a significantly higher average clustering coefficient compared to random networks, which proves their modular nature. Indeed, many cellular processes are governed by a subsets of biomolecules that form an interaction module. Since cellular processes are linked, the modules tend to be linked as well, but the linking molecules are often few, such that the module overlap is quite low [72, 73]. (page 15)

- for network motifs:

  However, some motifs have been found to be associated with optimized biological functions, like in the case of positive and negative feedback loops, oscillators or bifans [73]. (page 16)

In “Network Centralities and Node Ranking”:
- **Degree centrality**:

  Scale-free networks tend to contain hubs. The removal of such central nodes has great impact on the topology of the network. It has been shown that biological networks tend to be robust against random perturbations, but disruption of hubs often leads to system failure [77, 78]. (page 17)
- **Closeness centrality:**

Closeness centrality has been used to identify the top central metabolites in genome-based large-scale metabolic networks [79], to compare unicellular and multicellular eukarya, to rank pathways and obtain a perspective on the evolution of metabolic organization [80]. A decrease in closeness centrality of components has been observed as a consequence of increased distance between pathways throughout evolution [80]. It has been chosen as the best centrality measure that can be used to extract the metabolic core of a network [81]. (pages 17-18)

- **Betweenness centrality:**

Proteins with high betweenness centralities have been termed “bottlenecks”, for their role as key connector proteins with essential functional and dynamic properties [73], for example metabolites that control the flux between two big metabolic modules. Calculation of this centrality measure is discussed in [82] and [83] and their properties within the PPI network of yeast are detailed in [84]. (page 18)

- **Eigenvector centrality:**

In biology this centrality measurement has been used, among others, to identify synthetic genetic interactions [85], gene-disease associations [86] or network hubs [77]. (page 19)

- **Eccentricity centrality:**

In biological networks, proteins or other bioentities with high eccentricity are easily functionally reachable by other components of the network, and thus can readily perceive changes in concentration of other enzymes or molecules they are linked to. In contrast, those proteins that have lower eccentricities will often play a marginal functional role in the system [87]. (page 19)

- **Subgraph centrality:**

Subgraph centrality analysis has been used to study essential proteins in proteomic maps [77], to compute the degree of folding of protein chains [88], to understand the molecular structure of drug-like compounds [89] or to zoom into the topological environment of certain nodes in PPI networks of several organisms [90]. (page 20)

- **Matching index:**

The matching index is often used to cluster different components of a biological network according to some property. For instance, it has been used to describe spatial growth in brain networks during development [91] or to predict the connectivity of primate cortical networks [92]. (page 20)

- **further aspects:**

Further centrality measurements and their application to the study of PPIs in yeast are introduced in [85]. A discussion about how centrality correlates with lethality in biological networks can be found in [93]. The coupling between centrality and essentiality has also been investigated in several eukaryotic
protein networks [94]. It is very often the case that studies of a particular network involve the analysis and comparison of several centrality measures, for instance to study pleiotropy in human genetic diseases [87], to compare PPI and transcriptional regulation networks [95] or to test hub essentiality [77].

In “Network Topology”:

- **Scale-free:**
  Many biological networks also have scale-free properties, with nodes representing bioentities and edges the interactions between them (like proteins that interact physically or metabolites that take part in the same reaction) [73, 93, 100].
  An interesting analysis of most of these properties in various PPI, metabolic or transcriptional networks of several organisms (S.cerevisiae, H.pylori, C.elegans) can be found in [100].

- **Disassortative:**
  This is characteristic to most molecular interaction networks, where hubs have the tendency to link to nodes with fewer interaction partners rather than to other hubs [103, 104]. Newman [102] discusses this property for protein interaction networks, neural networks and food webs.

In “Network Models”:

- **Watts and Strogatz:**
  This type of topology characterizes many biological networks, like metabolic networks where it often happens that paths of few (three-four) reactions link most metabolites. As a consequence, local changes in metabolite concentration local perturbations in these networks will propagate throughout the entire network.

**B. Expand each section with software tools or techniques that would allow the type of analysis mentioned. It would be extremely useful to know how to conduct a specific type of analysis on a particular graph. These additions would provide a sense of the current state of biological network analysis. Essentially, as a reader, I’m excited to hear about all the possible ways to approach graph-based datasets, but I have no idea how to conduct this type of analysis (beyond writing my own code), or what has been tried for specific types of data before.**

**B. Software tools that can be used to analyze various graph theory properties on a network of interest have been mentioned:**

In “Network Properties”, for cliques:
Several tools have been developed for clique identification, like Clique Finder [70] within the Arabidopsis Co-expression Tool server or MiClique [68]. Bioconductor [71] provides a large collection of software for clique analysis. (page 15)

In “Network Centralities and Node Ranking”:
Tools that have implemented functionality for exploring the different types of centralities previously mentioned in biological networks and not only are CentiBiN [96], Visone [97], Pajek [98], VisANT [99]. In most of the cases, however, only a limited selection of centrality measures is available. (page 21)

In “Cluster Analysis and Visualization”:
Concerning the visualization of networks and the availability of the clustering techniques and their chaotic parameterizations, unfortunately today there are no such platforms or visualization tools that are able to integrate a variety of such algorithms and the implementation of such tools emerges [143]. Platforms that share clustering algorithms are the Network Analysis Tool (NEAT) [144] and jClust [145] but they are still poor in the variety of methods they offer. Software like ArrayCluster [146] and MCODE [60] is often used in analysis of gene expression profiles and coexpression detection. Many visualization tools [143] such as Medusa [147], Cytoscape [148], Pajek [98] and many others [143] visualize networks in both 2D and 3D, but very few of them like Arena3D [149] try to bridge the gap between clustering analysis and visualization. (page 33)

Additionally, several other comments and references have been introduced in order to improve the flow of the text and its informative capacity. We do not detail these here, but they are highlighted in the revised version.

Reviewer 3
Pavlopoulos and colleagues address an important and actual issue. The review is well organized and well written; there is a nice balance between theory and examples associated to a large but adequate number figures that makes the text easier to understand and pleasant to read.

b) page 7: "...Figure 2 below", please delete below since the final format is not ready yet
c) page 9: "has tuning time complexity..", \( \theta(N^3) \) should be substituted with \( O(N^3) \)
d) page 10: clique definition should be moved before the paragraph "Clustering Coefficient" since in the latter, the term clique is introduced
e) page 11: In Degree Centrality, add ref to Figure 5 when writing about hubs
f) page 15: please add references to Network Models descriptions
The reviewer pointed out a series of language and concept revisions to improve the clarity and accuracy of the text. We have addressed them as following:

a) ‘R’ was described as the set of real numbers and it was replaced by the proper mathematical character: “w: E → , where denotes the set of all real numbers” (page 8).

b) “below” has been removed (page 11).

c) The substitution has been performed as suggested: “Floyd’s algorithm has running time complexity O(N^3)” (page 14).

d) The clique definition was moved before the “Cluster ing Coefficient” paragraph (page 14).

e) Reference to Figure 5 added: “Nodes with very high degree centrality are called hubs since they are connected to many neighbors (see Figure 5)” (page 17).

f) References added to network models descriptions (pages 25-26).

g) Introduction before linkage list description added: “In the following, we describe the different methods used to calculate distances between clusters in hierachical clustering. “ (page 29).

h) References illustrating such examples are shown in 142, 143

i) Added “Bipartite” in the legend of Figure 1 (page 55).

j) Typo corrected in Figure 3 legend (page 55).

k) “Probes” have been replaced by “genes” in the legend of Figure 8 (page 59).
With the comments of the reviewers addressed, we would be grateful if you could consider this manuscript for publication in your journal.

We would like to thank you for your attention and for reviewing the current study.

Looking forward to your decision.

Yours sincerely,

The authors.