Exploration of biological network centralities with CentiBiN

Björn H. Junker[1], Dirk Koschützki*[1] and Falk Schreiber[1]

[1] Department of Molecular Genetics, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Corrensstr. 3, 06466 Gatersleben, Germany

Email: Björn H. Junker - junker@ipk-gatersleben.de; Dirk Koschützki*- koschuet@ipk-gatersleben.de; Falk Schreiber - schreibe@ipk-gatersleben.de;

*Corresponding author

Abstract

Background: The elucidation of whole-cell regulatory, metabolic, interaction and other biological networks generates the need for a meaningful ranking of network elements. *Centrality analysis* ranks network elements according to their importance within the network structure and different centrality measures focus on different importance concepts. Central elements of biological networks have been found to be, for example, essential for viability.

Results: CentiBiN (Centralities in Biological Networks) is a tool for the computation and exploration of centralities in biological networks such as protein-protein interaction networks. It computes 17 different centralities for directed or undirected networks, ranging from local measures, that is, measures that only consider the direct neighbourhood of a network element, to global measures. CentiBiN supports the exploration of the centrality distribution by visualising central elements within the network and provides several layout mechanisms for the automatic generation of graphical representations of a network. It supports different input formats, especially for biological networks, and the export of the computed centralities to other tools.

Conclusions: CentiBiN helps systems biology researchers to identify crucial elements of biological networks. CentiBiN including a user guide and example data sets is available free of charge at http://centibin.ipk-gatersleben.de/. CentiBiN is available in two different versions: a Java Web Start application and an installable Windows application.

Background

The shift of biological research towards massively parallel techniques opens new opportunities but at the same time raises problems in deriving meaningful information out of the wealth of generated data. Such data might be represented as networks, in which the vertices (e.g. transcripts, proteins or metabolites) are linked by edges (correlations, interactions or reactions, respectively). Structural analysis of networks can lead to new insights into biological systems and is a helpful method for proposing new hypotheses. Several techniques for such structural analysis exist, such as the analysis of the global network structure (e.g. scale-free networks [1]), network motifs (i.e. small subnetworks which occur significantly more often in the biological network than in random networks [2]), network clustering (modularisation of the network into parts [3]) and network centralities [4]. Network centralities are used to rank elements of a network according to a given importance concept.

Ranking of network elements has been used to analyse biological networks in several cases. For example, it has been shown for metabolic networks that the most central metabolites are evolutionarily conserved [5]. Moreover, in the protein-protein interaction network of baker's yeast (*Saccharomyces cerevisiae*) it has been found that the centrality of a protein correlates with the essentiality of its gene, which was characterised by a high probability of a lethal effect observed upon knockout of this gene [6]. Recently, it has been observed that in transcript co-expression networks genes with high degree-centrality, that is, highly connected genes, tend to be essential and conserved [7].

The determination of central elements in biological networks will create new hypotheses that lead to more rational approaches in experimental design. If the important elements of a network are known, further experimental investigations cen be limited to them. Depending on the biological question, a vertex of a network might be of importance, for example, if it is connected to many other vertices or if the sum of the shortest path distances to all the other vertices is small, see Figure 1. For these different ranking concepts, a broad variety of centrality measures are available (see Table 1) which have been described in a recent review [8].

However, the use of centralities as a structural analysis method for biological networks is controversial and several centrality measures should be considered within an exploratory process [9]. To support such analysis and due to the complexity of both biological networks and centrality calculations, a tool is needed to facilitate these investigations. Here we present CentiBiN, an application for the calculation and visualisation of centralities for biological networks.

Implementation

The core of CentiBiN are newly implemented algorithms for centrality analysis and visualisation (e.g. most of the centrality measures, the graphical user interface, cleanup methods and some imports/exports such as DOT, Pajek `.vec` and TSV). It is based on JUNG, the Java Universal Network/Graph Framework, an open source library which can be downloaded from [10]. JUNG provides standard graph library functionality (e.g. data structures, imports/exports, layouts, graph generators and a few centrality algorithms). CentiBiN is written in Java. It requires an installation of the Java Runtime Environment Version 1.4.2 or later which is available from the Java download page [11]. It is available free of charge as a Java Web Start application and an installable Windows application including a user guide and example data sets. Depending on available main memory and the centrality algorithm used networks up to several tens of thousands vertices can be analysed. Large networks with several thousand vertices are not readable on a computer screen anymore and the drawing routine significantly slows down for such networks. Thus they are not displayed, but can nevertheless be analysed and exported. A corresponding threshold is configurable by the user.

Results and discussion

CentiBiN's major features are:

Computation of centralities CentiBiN supports a wide range of different centrality measures ranging from local measures (which only consider the direct neighbourhood of a vertex) to global measures. In total 17 centralities for undirected networks and 15 centralities for directed networks are available, see Table 2.

Plotting the distribution of centrality values The distribution of centrality values and a histogram of centrality values can be displayed, see Figure 2. Several of these diagrams can be opened simultaneously and allow the easy comparison of the centrality distributions.

Visualisation and navigation within the network From the list of centrality values several vertices can be selected. These are highlighted in the network and can therefore be easily located. Additionally CentiBiN supports zoom and pan functionalities to navigate within a displayed network. The underlying graph library offers five different layout algorithms for networks, reaching from simple

circular to more advanced force directed layouts [12, 13]. Depending on the network, one or the other layout method results in a better visualisation.

Cleaning up a network Depending on the centrality measure to be applied, the network has to fulfil certain preconditions. These can be simplicity, connectedness, and loop-freeness. Therefore, several algorithms are implemented for transforming a network into the required form. These are the removal of all loops (edges from one vertex to itself), the removal of all vertices that are not part of the largest connected component (giant strong component), the removal of all parallel edges, and the transformation of the network into an undirected or directed form.

Reading and writing networks and centralities Networks can be loaded out of four different file formats: the Pajek `.net` file format [14], a text file containing an adjacency matrix representation of a network, the GraphML file format [15] and the TSV-files provided by the Database of Interacting Proteins (DIP) [16]. It is possible to store networks in the Pajek `.net` file format, in a text file containing a representation of the adjacency matrix, and in the DOT file format used by Graphviz [17]. To support further analysis of computed centralities they can be saved either in the Pajek `.vec` file format or in a TSV format. These files can be imported in other applications, such as R [18].

Generation of random networks The generation of random networks based on five different algorithms (provided by JUNG), such as Kleinberg's small-world generator [19] and the Barabási-Albert scale-free generator [20], is available. These networks can be analysed and visualised and may be used as reference models.

A typical example for networks which can be analysed are protein-protein interaction networks. All interactions between proteins of an organism can be represented as a network. Several databases contain information about such protein-protein interactions. We chose the Database of Interacting Proteins (DIP) [16] from which interaction data can be imported into CentiBiN. Figure 3 shows screen-shots of the tool with interaction data from *Escherichia coli* and *Mus musculus* (mouse). The most import proteins according to the Current-Flow betweenness [21] are highlighted.

Besides CentiBiN, several other software tools support the analysis of networks with centralities. Many of them are described in detail in the work of Huisman and van Duijn [22]. The focus of their comparison lies on software for social network analysis, an area where interactions between individuals are analysed. As

the original concept of centralities can be traced back into this field of science, software packages for social network analysis often provide methods for centrality analysis. Some of the tools evaluated by Huisman and van Duijn are commercial (UCINET, NetMiner), have a text based front end (STRUCTURE) or are software systems for advanced statistical modelling of social networks (StOCNet). Furthermore, only some of them have more than a few centralities implemented (e.g. MultiNet, Visone, Pajek). Most of the tools specifically designed for the analysis of biological networks (e.g. Cytoscape [23], Osprey [24]) do not support centrality analysis so far. The distribution plot (see Figure 2) is similar to the plot available in VisANT [25]. Compared to CentiBiN none of the available systems covers this extensive number of different centrality measures. Moreover, CentiBiN supports direct access to biological data, allows the visualisation of the network and the centralities together, has a simple input file format, and is available free of charge. For the next release we plan to integrate several methods for comparison of centralities with biological information. For that it is foreseen to integrate mechanisms to mark vertices according to given information or to correlate centralities with experimental results. Additionally we plan to support more input formats, for example PSI's Molecular Interaction XML Format [26] and to implement advanced visualisation methods.

Conclusions

CentiBiN is a tool for the computation and exploration of *centralities in biological networks*. With CentiBiN it is possible to infer information about the "importance" of an element in a biological network based on different importance concepts. We have applied this for protein-protein interaction networks. We are convinced that CentiBiN provides valuable help to systems biologists in the generation of hypotheses from large-scale data sets.

Availability and requirements

- **Project name:** CentiBiN
- **Project home page:** http://centibin.ipk-gatersleben.de/
- **Operating system(s):** Platform independent
- **Programming language:** Java

- **Other requirements:** Java 1.4.2 or higher, 256MB RAM recommended
- **License:** The CentiBiN application is available free of charge.
- **Any restrictions to use by non-academics:** none

Authors' contributions

All authors participated in the design of the system. DK implemented the system. DK and BHJ drafted the manuscript. All authors read, revised and approved the final version and were listed in alphabetical order.

Acknowledgements

The German Ministry of Education and Research (BMBF) supported part of this work under grants 0312706A and 0313115. We like to thank the developers of the JUNG library for their excellent work, Michael Telgkamp for his work on the implementation and Daniel Fleischer for his help in implementing two centrality algorithms.

References

1. Albert R, Barabási AL: **Statistical Mechanics of Complex Networks**. *Reviews of Modern Physics* 2002, **74**:47–97.
2. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U: **Network Motifs: Simple Building Blocks of Complex Networks**. *Science* 2002, **298**(5594):824–827.
3. Holme P, Huss M, Jeong H: **Subnetwork hierarchies of biochemical pathways**. *Bioinformatics* 2003, **19**(4):532–538.
4. Wuchty S, Stadler PF: **Centers of complex networks**. *J Theor Biol* 2003, **223**:45–53.
5. Fell DA, Wagner A: **The small world of metabolism**. *Nat Biotechnol* 2000, **18**:1121–1122.
6. Jeong H, Mason SP, Barabási AL, Oltvai ZN: **Lethality and centrality in protein networks**. *Nature* 2001, **411**:41–42.

7. Bergmann S, Ihmels J, Barkai N: **Similarities and Differences in Genome-Wide Expression Data of Six Organisms**. *PLoS Biol* 2004, **2**:85–93.

8. Koschützki D, Lehmann KA, Peeters L, Richter S, Tenfelde-Podehl D, Zlotowski O: **Centrality Indices**. In *Network Analysis: Methodological Foundations*, Volume 3418 of *LNCS Tutorial*. Edited by Brandes U, Erlebach T, Springer 2005:16–61.

9. Koschützki D, Schreiber F: **Comparison of Centralities for Biological Networks**. In *Proc. German Conf. Bioinformatics (GCB'04)*, Volume P-53 of *LNI* 2004:199–206.

10. **JUNG, the Java Universal Network/Graph Framework** [**http://jung.sourceforge.net**].

11. **Java Web Page** [**http://www.java.com/**].

12. Kamada T, Kawai S: **An algorithm for drawing general undirected graphs**. *Information Processing Letters* 1989, **31**:7–15.

13. Fruchterman TMJ, Reingold EM: **Graph Drawing by Force-directed Placement**. *Software - Practice and Experience* 1991, **21**(11):1129–1164.

14. Batagelj V, Mrvar A: **Pajek - Analysis and Visualization of Large Networks**. In *Graph Drawing Software*. Edited by Jünger M, Mutzel P, Berlin: Springer 2004:77–103.

15. Brandes U, Eiglsperger M, Herman I, Himsolt M, Marshall MS: **GraphML Progress Report: Structural Layer Proposal**. In *Proceedings of the 9th International Symposium on Graph Drawing*, Volume 2265 of *Lecture Notes in Computer Science*. Edited by Mutzel P, Jünger M, Leipert S, Berlin: Springer 2002:501–512.

16. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D: **The Database of Interacting Proteins: 2004 update**. *Nucleic Acids Res* 2004, **32**:D449–D451.

17. **Graphviz - Graph Visualization Software** [**http://www.graphviz.org/**].

18. R Development Core Team: **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria 2005, [http://www.R-project.org]. [ISBN 3-900051-07-0].

19. Kleinberg JM: **Navigation in a small world**. *Nature* 2000, **406**:845.

20. Barabási AL, Albert R: **Emergence of scaling in random networks**. *Science* 1999, **286**:509–512.

21. Brandes U, Fleischer D: **Centrality Measures Based on Current Flow**. In *Proc. 22nd Symp. Theoretical Aspects of Computer Science (STACS '05)*, Volume 3404 of *Lecture Notes in Computer Science (LNCS)*, Springer-Verlag 2005:533–544.

22. Huisman M, van Duijn MA: **Software for Social Network Analysis**. In *Models and Methods in Social Network Analysis*. Edited by Carrington PJ, Scott J, Wasserman S, New York: Cambridge University Press 2004:270–316.

23. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks**. *Genome Res* 2003, **13**(11):2498–2504.

24. Breitkreutz BJ, Stark C, Tyers M: **Osprey: a network visualization system**. *Genome Biol* 2003, **4**(3).

25. Hu Z, Mellor J, Wu J, Yamada T, Holloway D, DeLisi C: **VisANT: data-integrating visual framework for biological networks and modules**. *Nucleic Acids Res* 2005, **33**(Web Server issue):W352 – W357.

26. Hermjakob H, Montecchi-Palazzi L, Bader G, Wojcik J, Salwinski L, Ceol A, Moore S, Orchard S, Sarkans U, von Mering C, Roechert B, Poux S, Jung E, Mersch H, Kersey P, Lappe M, Li Y, Zeng R, Rana D, Nikolski M, Husi H, Brun C, Shanker K, Grant SGN, Sander C, Bork P, Zhu W, Pandey A, Brazma A, Jacq B, Vidal M, Sherman D, Legrain P, Cesareni G, Xenarios I, Eisenberg D, Steipe B, Hogue C, Apweiler R: **The HUPO PSI's Molecular Interaction format - a community standard for the representation of protein interaction data**. *Nat Biotechnol* 2004, **22**:177–183.

27. Jacob R, Koschützki D, Lehmann KA, Peeters L, Tenfelde-Podehl D: **Algorithms for Centrality Indices**. In *Network Analysis: Methodological Foundations*, Volume 3418 of *LNCS Tutorial*. Edited by Brandes U, Erlebach T, Springer 2005:62–82.

28. Harary F, Hage P: **Eccentricity and centrality in networks**. *Social Networks* 1995, **17**:57–63.

29. Sabidussi G: **The Centrality Index of a Graph**. *Psychometrika* 1966, **31**:581–603.

30. Valente TW, Foreman RK: **Integration and radiality: measuring the extent of an individual's connectedness and reachability in a network**. *Social Networks* 1998, **1**:89–105.

31. Slater PJ: **Maximin Facility Location**. *Journal of National Bureau of Standards* 1975, **79B**:107–115.

32. Shimbel A: **Structural Parameters of Communication Networks**. *Bulletin of Mathematical Biophysics* 1953, **15**:501–507.

33. Freeman LC: **A Set of Measures of Centrality Based Upon Betweeness**. *Sociometry* 1977, **40**:35–41.

34. Katz L: **A new status index derived from sociometric analysis**. *Psychometrika* 1953, **18**:39–43.

35. Bonacich P: **Factoring and Weighting Approaches to Status Scores and Clique Identification**. *Journal of Mathematical Sociology* 1972, **2**:113–120.

36. Hubbell CH: **In Input-Output Approach to Clique Identification**. *Sociometry* 1965, **28**:377–399.

37. Bonacich P: **Power and Centrality: A Family of Measures**. *American Journal of Sociology* 1987, **92**(5):1170–1182.

38. Page L, Brin S, Motwani R, Winograd T: **The PageRank Citation Ranking: Bringing Order to the Web**. Tech. rep., Stanford Digital Library Technologies Project 1998.

39. Kleinberg JM: **Authoritative Sources in a Hyperlinked Environment**. *Journal of the ACM* 1999, **46**(5):604–632.

Figures

Figure 1 - Different centrality measures rank vertices differently

The most important vertices according to the degree-centrality (red), that is, where a vertex of a network is central if it is highly connected, and the Closeness centrality (blue), that is, where a vertex is central if the sum of the shortest path distances to all the other vertices is small.

Figure 2 - Plots generated with CentiBiN

The distribution of the degree centrality and a histogram of the closeness centrality for a random network.

Figure 3 - Two screen-shots showing centrality analysis with CentiBiN

The networks represent protein-protein interactions in *Escherichia coli* (top) and *Mus musculus* (bottom) according to the Database of Interacting Proteins (DIP), release 2005-01-26. The top ranking proteins according to the Current-Flow Betweenness centrality [21] are highlighted.

Tables

Table 1 - Definitions for the centrality measures

Let $G = (V, E)$ be an undirected or directed, (strong) connected graph with $n = |V|$ vertices; $\deg(v)$ denotes the degree of the vertex v in an undirected graph; $\text{dist}(v, w)$ denotes the length of a shortest path between the vertices s and t; σ_{st} denotes the number of shortest paths from s to t and $\sigma_{st}(v)$ the number of shortest path from s to t that use the vertex v. Let A be the adjacency matrix of the graph G. For a more detailed description and further references please see [8, 27]. Abbreviations used: S.-P.: shortest path, C.-F.: current flow.

Name	**Definition**	**Remarks**	**Ref**
Degree	$\mathcal{C}_{deg}(v) := \deg(v)$	For directed graphs in- and out-degree is used.	
Eccentricity	$\mathcal{C}_{ecc}(v) := \frac{1}{\max\{\mathrm{dist}(v,w):\, w \in V\}}$		[28]
Closeness	$\mathcal{C}_{clo}(v) := \frac{1}{\sum_{w \in V} \mathrm{dist}(v,w)}$		[29]
Radiality	$\mathcal{C}_{rad}(v) := \frac{\sum_{w \in V} (\Delta_G + 1 - \mathrm{dist}(v,w))}{n-1}$	Δ_G is the diameter of the graph G, defined as the maximum distance between any two vertices of G.	[30]
Centroid Value	$\mathcal{C}_{cen}(v) := \min\{\mathrm{f}(v,w):\, w \in V \setminus \{v\}\}$	Where $f(v,w) := \gamma_v(w) - \gamma_w(v)$ and $\gamma_v(w)$ denotes the number of vertices that are closer to v than to w.	[31]
Stress	$\mathcal{C}_{str}(v) := \sum_{s \neq v \in V} \sum_{t \neq v \in V} \sigma_{st}(v)$		[32]
S.-P. Betweenness	$\mathcal{C}_{spb}(v) := \sum_{s \neq v \in V} \sum_{t \neq v \in V} \delta_{st}(v)$	$\delta_{st}(v) := \frac{\sigma_{st}(v)}{\sigma_{st}}$	[33]
C.-F. Closeness	$\mathcal{C}_{cfc}(v) := \frac{n-1}{\sum_{t \neq v} p_{vt}(v) - p_{vt}(t)}$	Where $p_{vt}(t)$ equals the potential difference in an electrical network.	[21]
C.-F. Betweenness	$\mathcal{C}_{cfb}(v) := \frac{1}{(n-1)(n-2)} \sum_{s,t \in V} \tau_{st}(v)$	Where $\tau_{st}(v)$ equals the fraction of electrical current running over vertex v in an electrical network.	[21]
Katz Status	$\mathcal{C}_{katz} := \sum_{k=1}^{\infty} \alpha^k (A^T)^k \vec{1}$	Where α is a positive constant.	[34]
Eigenvector	$\lambda \mathcal{C}_{eiv} = A\mathcal{C}_{eiv}$	The eigenvector to the dominant eigenvalue of A is used.	[35]
Hubbell index	$\mathcal{C}_{hbl} = \vec{E} + W\mathcal{C}_{hbl}$	Where $\vec{E}$ is some exogenous input and W is a weight matrix derived from the adjacency matrix A.	[36]
Bargaining	$\mathcal{C}_{brg} := \alpha(I - \beta A)^{-1} A\vec{1}$	Where α is a scaling factor and β is the influence parameter.	[37]
PageRank	$\mathcal{C}_{pr} = dP\mathcal{C}_{pr} + (1-d)\vec{1}$	Where P is the transition matrix and d is the damping factor.	[38]
HITS-Hubs	$\mathcal{C}_{hubs} = A\mathcal{C}_{auths}$	Assuming $\mathcal{C}_{auths}$ is known.	[39]
HITS-Authorities	$\mathcal{C}_{auths} = A^T\mathcal{C}_{hubs}$	Assuming $\mathcal{C}_{hubs}$ is known.	[39]

Table 2 - Centrality measures implemented in CentiBiN

Definitions for these measures can be found in Table 1.

Centrality	Type	Directed graphs	Undirected graphs
Degree	Neighbourhood based	Yes[1]	Yes
Eccentricity	Distance based	Yes	Yes
Closeness	Distance based	Yes	Yes
Radiality	Distance based	Yes	Yes
Centroid	Distance based	Yes	Yes
Stress	Shortest-Path based	Yes	Yes
S.-P. Betweenness	Shortest-Path based	Yes	Yes
C.-F. Betweenness	Current-Flow based	No[2]	Yes
C.-F. Closeness	Current-Flow based	No[2]	Yes
Katz Status	Feedback based	Yes	Yes
Eigenvector	Feedback based	Yes	Yes
Hubbell index	Feedback based	Yes	Yes
Bargaining	Feedback based	Yes	Yes
PageRank	Feedback based	Yes	Yes
HITS-Hubs	Feedback based	Yes	Yes[3]
HITS-Auths	Feedback based	Yes	Yes[3]
Closeness-vitality	Vitality based	Yes	Yes

[1] both in- and out-degree

[2] not defined for directed graphs

[3] HITS-Hubs and HITS-Auths give identical results for undirected graphs

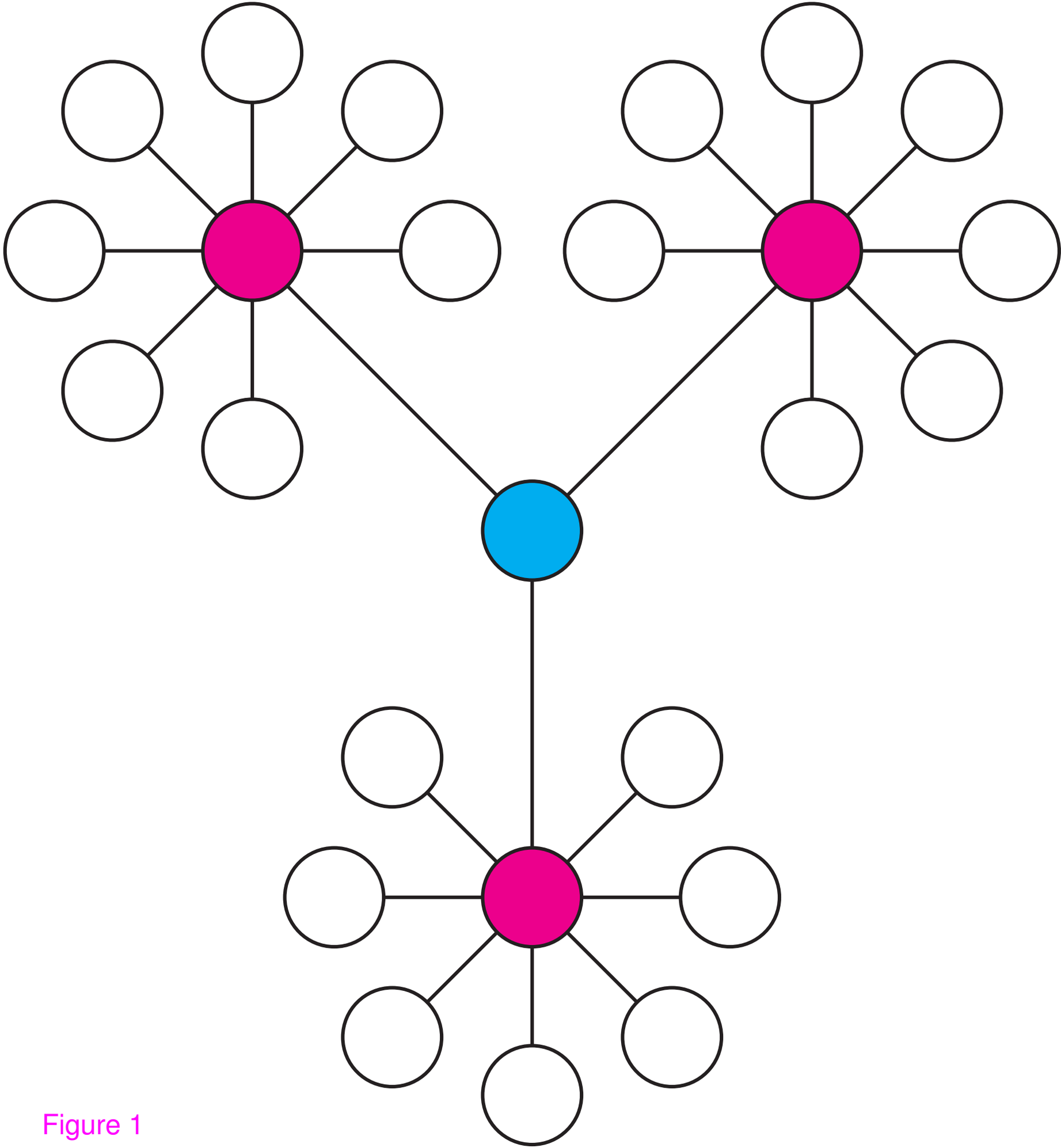
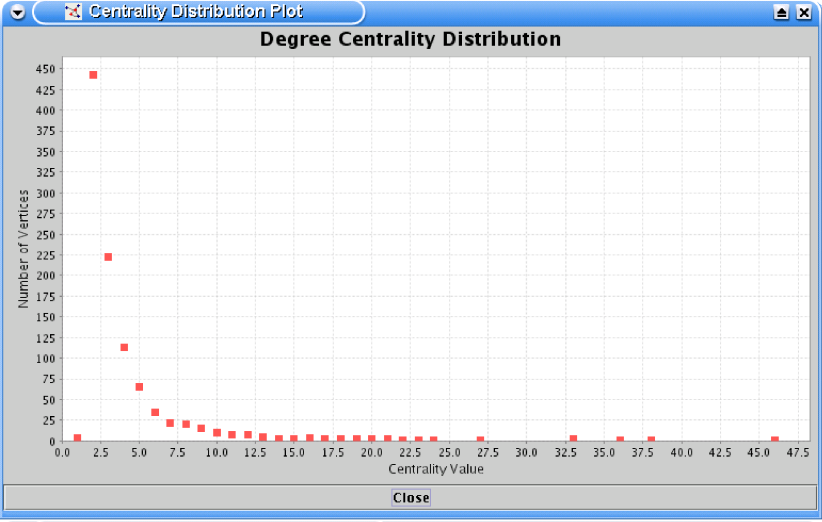


Figure 1


Centrality Distribution Plot
Degree Centrality Distribution
Number of Vertices
Centrality Value
Close

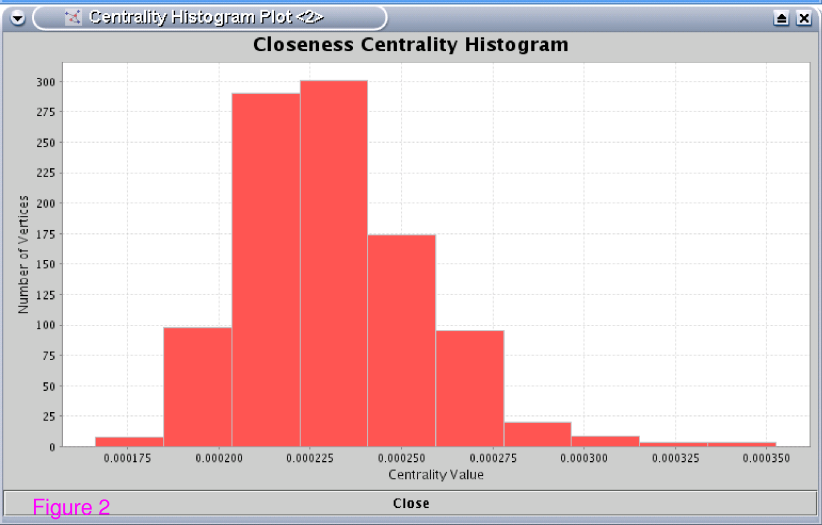
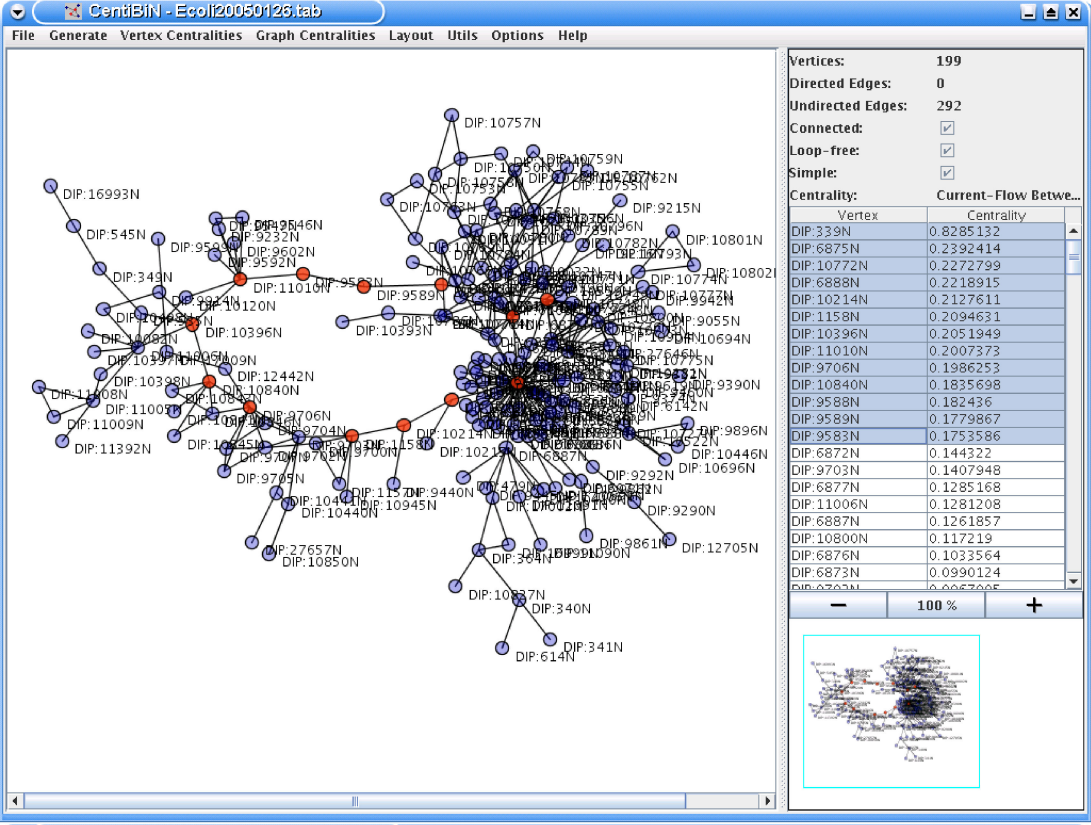

Centrality Histogram Plot <2>
Closeness Centrality Histogram
Number of Vertices
Centrality Value
Close

Figure 2


CentiBiN - Ecoli20050126.tab
File Generate Vertex Centralities Graph Centralities Layout Utils Options Help
Vertices: 199
Directed Edges: 0
Undirected Edges: 292
Connected:
Loop-free:
Simple:
Centrality: Current-Flow Betwe...
Vertex Centrality
DIP:339N 0.8285132
DIP:6875N 0.2392414
DIP:10772N 0.2272799
DIP:6888N 0.2218915
DIP:10214N 0.2127611
DIP:1158N 0.2094631
DIP:10396N 0.2051949
DIP:11010N 0.2007373
DIP:9706N 0.1986253
DIP:10840N 0.1835698
DIP:9588N 0.182436
DIP:9589N 0.1779867
DIP:9583N 0.1753586
DIP:6872N 0.144322
DIP:9703N 0.1407948
DIP:6877N 0.1285168
DIP:11006N 0.1281208
DIP:6887N 0.1261857
DIP:10800N 0.117219
DIP:6876N 0.1033564
DIP:6873N 0.0990124
100 %

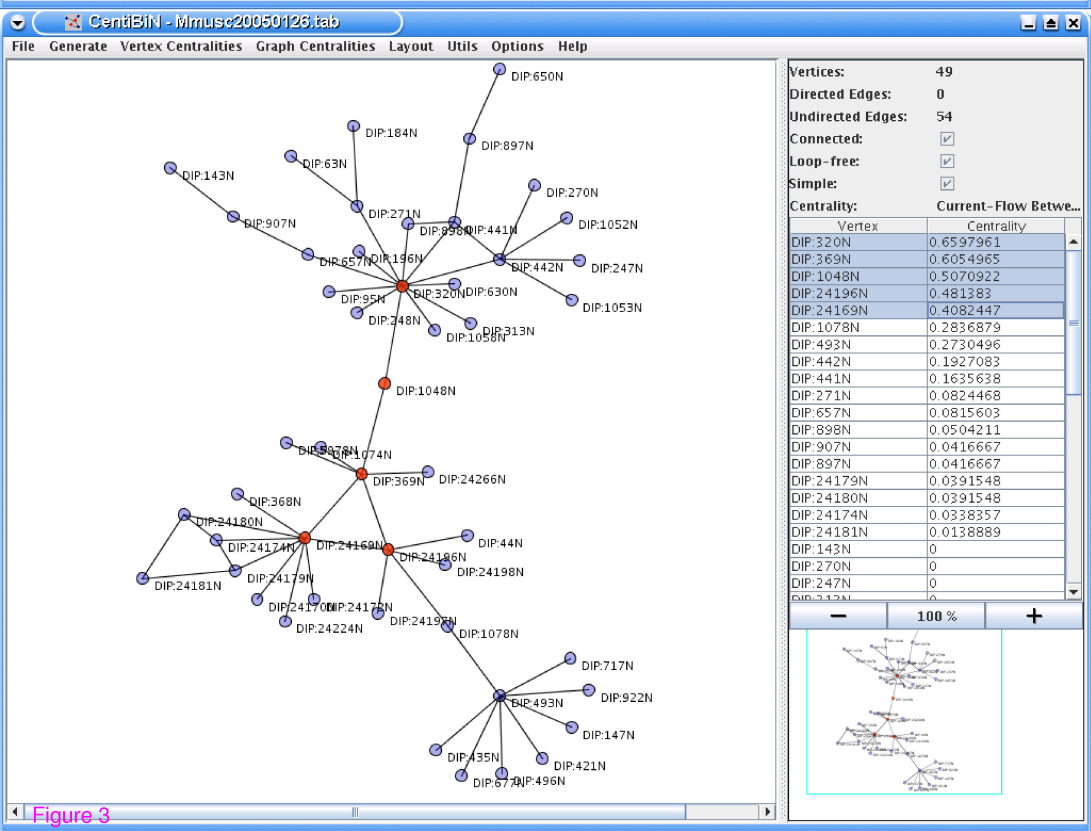

CentiBiN - Mmusc20050126.tab
File Generate Vertex Centralities Graph Centralities Layout Utils Options Help
Vertices: 49
Directed Edges: 0
Undirected Edges: 54
Connected:
Loop-free:
Simple:
Centrality: Current-Flow Betwe...
Vertex Centrality
DIP:320N 0.6597961
DIP:369N 0.6054965
DIP:1048N 0.5070922
DIP:24196N 0.481383
DIP:24169N 0.4082447
DIP:1078N 0.2836879
DIP:493N 0.2730496
DIP:442N 0.1927083
DIP:441N 0.1635638
DIP:271N 0.0824468
DIP:657N 0.0815603
DIP:898N 0.0504211
DIP:907N 0.0416667
DIP:897N 0.0416667
DIP:24179N 0.0391548
DIP:24180N 0.0391548
DIP:24174N 0.0338357
DIP:24181N 0.0138889
DIP:143N 0
DIP:270N 0
DIP:247N 0
100 %

Figure 3