Network modeling, visualization and analysis I (Cytoscape)

Master en Bioinformática
2013, Valencia

Luz Garcia-Alonso
PhD at Computational Genomics Institute
Centro de Investigación Príncipe Felipe (CIPF)

Index

- Paper Review (Barabasi & Oltvait, 2004)
  - Why Network Biology?
  - Basic Network Nomenclature and Measures
  - Architectural Features of Cellular Networks
  - Motifs, Modules and Hierarchical Networks
  - Network Robusntess
- Cystocape
  - Why?
  - Loading Networks and Attributes
  - Visualize and Analyze Networks
  - Extended Functionality
  - Exporting & Saving Data
- Exercises
  - Installing Cytoscape and Plugins
  - Hands on Cytoscape: an Example
  - Exersices 1 and 2

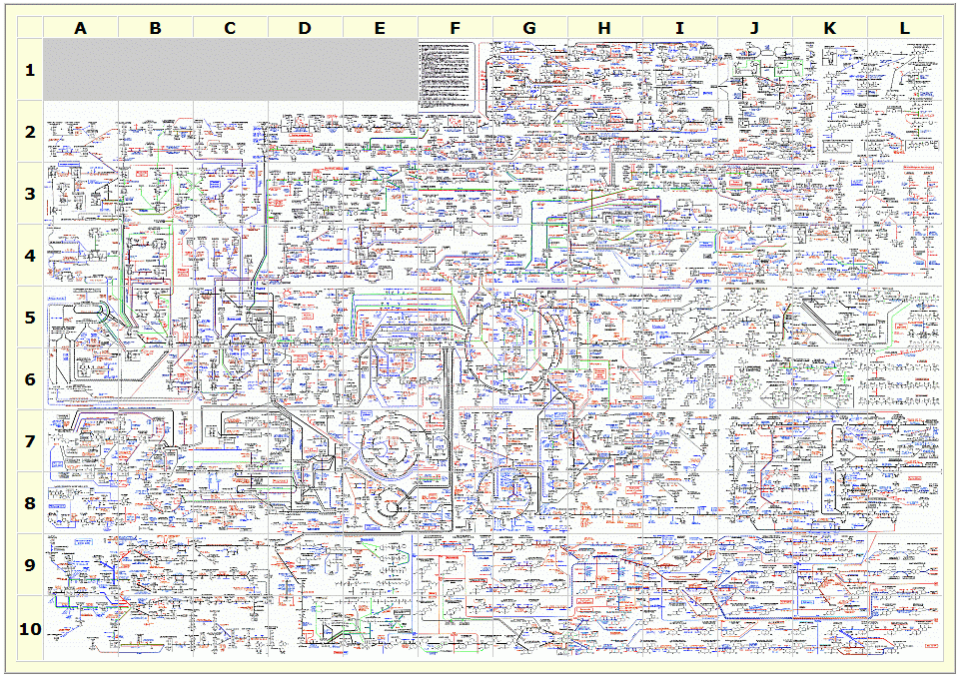
Paper Review (Barabasi & Oltvait, 2004)

NETWORK BIOLOGY: UNDERSTANDING THE CELL'S FUNCTIONAL ORGANIZATION

Albert-László Barabási & Zoltán N. Oltvai†*

A key aim of postgenomic biomedical research is to systematically catalogue all molecules and their interactions within a living cell. There is a clear need to understand how these molecules and the interactions between them determine the function of this enormously complex machinery, both in isolation and when surrounded by other cells. Rapid advances in network biology indicate that cellular networks are governed by universal laws and offer a new conceptual framework that could potentially revolutionize our view of biology and disease pathologies in the twenty-first century.

Paper Review (Barabasi & Oltvait, 2004)
Why Network Biology?



- Molecules alone are not sufficient to understand complex systems: **reductionist approach**
- Necessity to understand the relationships between molecules: **holistic approach**

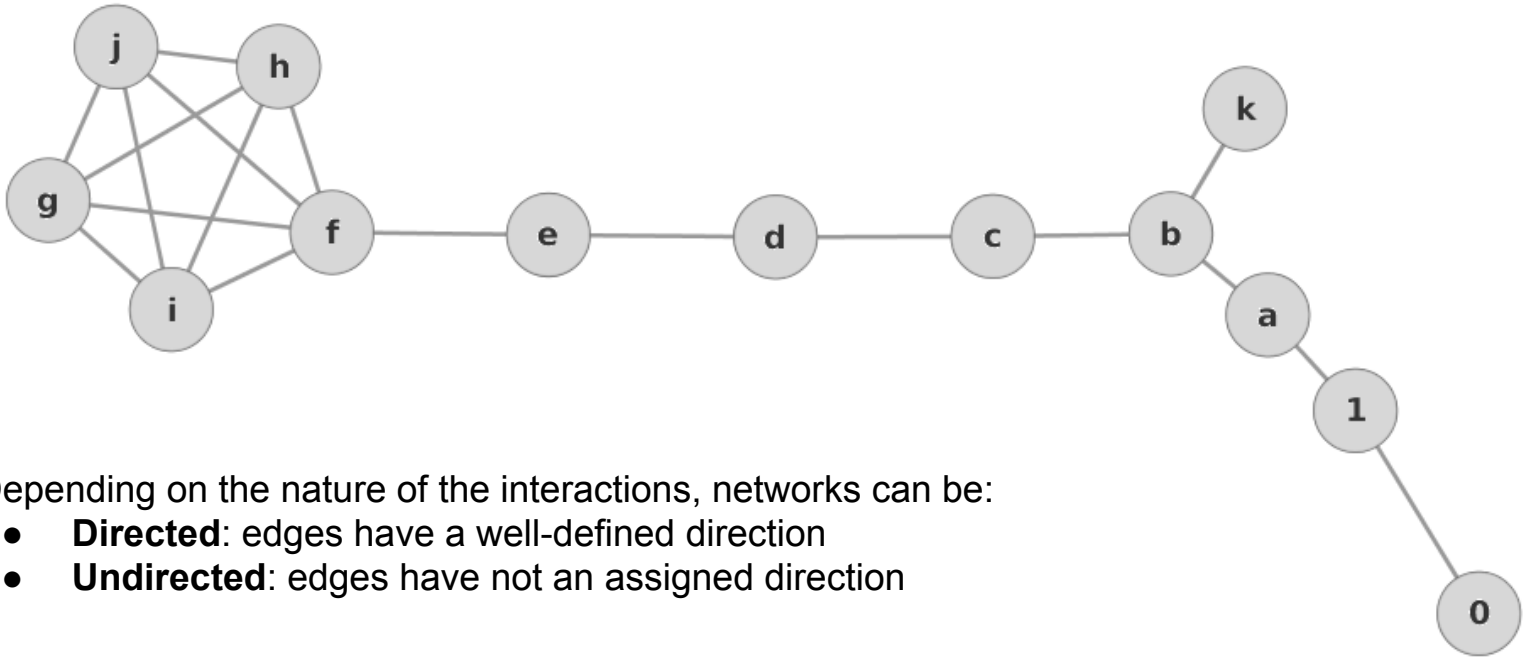
Paper Review (Barabasi & Oltvait, 2004)
Why Network Biology?

- Complex biological systems may be abstractly represented and analyzed as computable **networks**.
- Useful for discovering **emergent properties** in large data sets (better than tables or matrixes).
- Can represent:

Type	Relationship Represented	Technology
Protein-protein interaction networks	Protein Binding	Y2H, Protein Arrays
Gene regulatory networks	DNA-Protein Interactions / Gene Coexpression	ChIP-chip, ChIP-seq
Metabolic networks	Metabolic Reactions	Large-Scale Ezymology
Signaling networks	Cascade Signaling Interactions	Integration of previous networks
Neuronal Networks	Neuron Axon Connections	Neuroimaging
Species interaction networks	Associations Between Individuals	Spatio-Temporal interactions

Paper Review (Barabasi & Oltvait, 2004)
Basic Network Nomenclature

Nodes represent units in the network, while **edges** represent the interactions between the units. Networks define a sequence of edges connecting nodes (**paths**).



Depending on the nature of the interactions, networks can be:

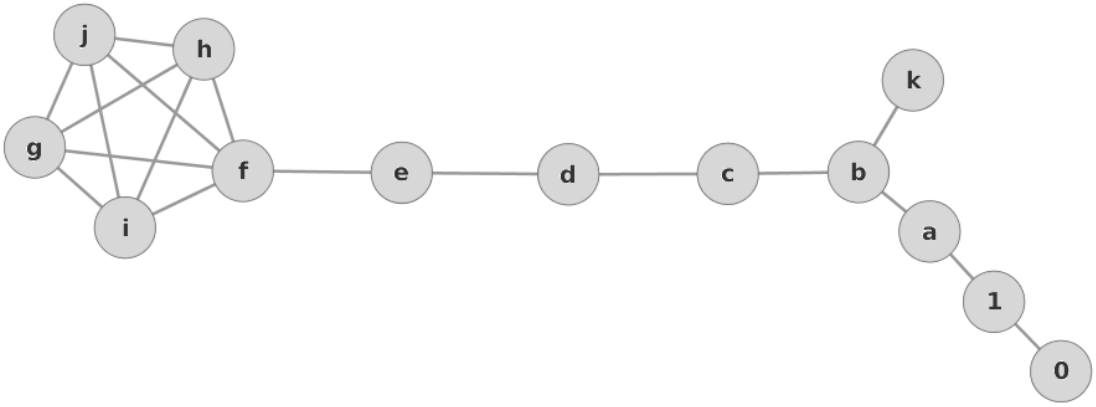
- **Directed**: edges have a well-defined direction
- **Undirected**: edges have not an assigned direction

Paper Review (Barabasi & Oltvait, 2004)
Architectural Features of Cellular Networks

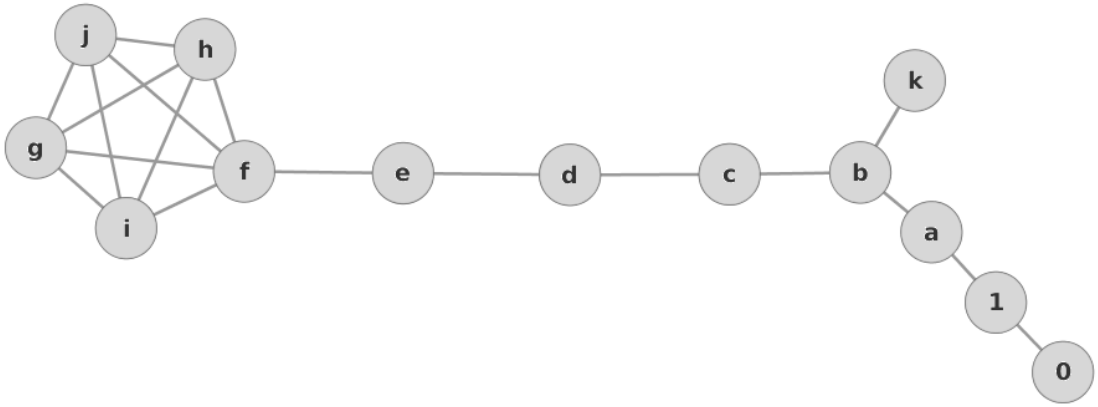
The **topology** of the network characterizes the system (global topology parameters) and its individual components (local topology parameters). Network topology parameters are easily compared.

Local Topology Parameters

- **Degree**: the number of direct interactors. Hub is a node with an exceptional high degree.
- **Betweenness**: the frequency with which a protein lies on the shortest path between other proteins.
- **Closeness**: the average distance to all other proteins.
- **Clustering Coefficient**: the degree to which nodes in a graph tend to cluster together.



Paper Review (Barabasi & Oltvait, 2004)
Architectural Features of Cellular Networks

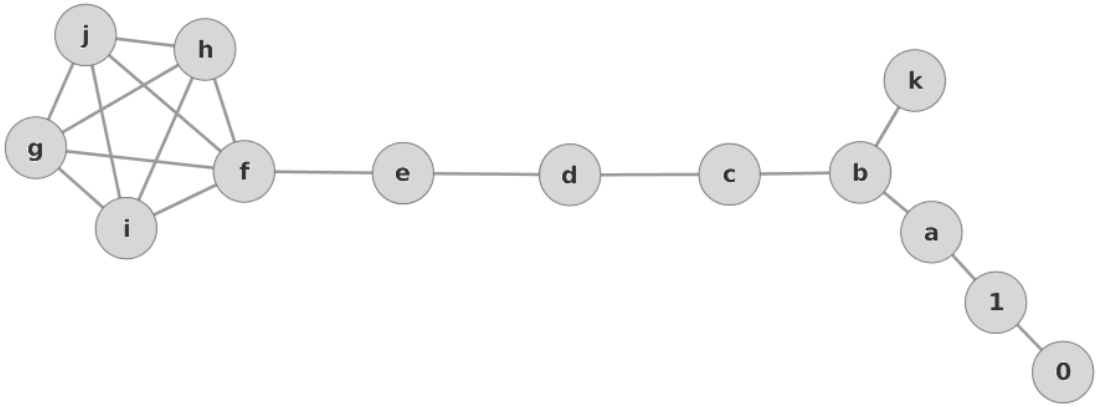


Degree: the number of direct interactors.
Betweenness: the frequency with which a protein lies on the shortest path between other proteins.
Closeness: the average distance to all other proteins.
Clustering Coefficient: the degree to which nodes in a graph tend to cluster together.

Questions

- What protein has the highest Degree?
- What protein has the highest Betweenness?
- What proteins have the highest Closeness?
- What proteins have the highest Clustering Coefficient?

Paper Review (Barabasi & Oltvait, 2004)
Architectural Features of Cellular Networks



Degree: the number of direct interactors.
Betweenness: the frequency with which a protein lies on the shortest path between other proteins.
Closeness: the average distance to all other proteins.
Clustering Coefficient: the degree to which nodes in a graph tend to cluster together.

Questions

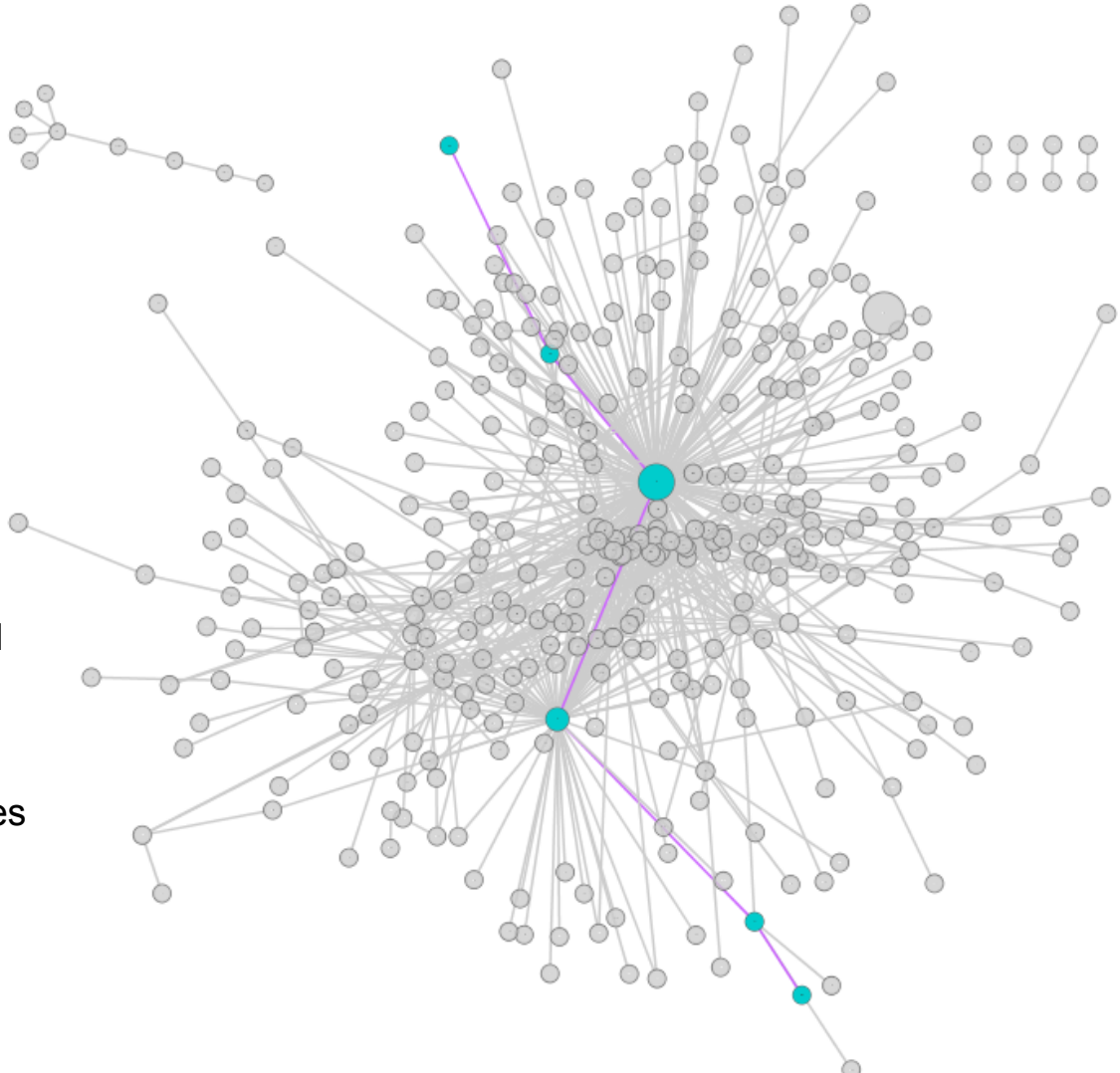
- What protein has the highest Degree?
- What protein has the highest Betweenness?
- What proteins have the highest Closeness?
- What proteins have the highest Clustering Coefficient?

ID	BetweennessCentrality	ClosenessCentrality	ClusteringCoefficient	Degree
f	0.46153846	0.30952381	0.6	5
g	0.0	0.25490196	1.0	4
k	0.0	0.26	0.0	1
c	0.53846154	0.36111111	0.0	2
d	0.53846154	0.36111111	0.0	2
1	0.15384615	0.23214286	0.0	2
e	0.51282051	0.34210526	0.0	2
a	0.28205128	0.2826087	0.0	2
b	0.6025641	0.34210526	0.0	4
h	0.0	0.25490196	1.0	4
i	0.0	0.25490196	1.0	4
j	0.0	0.25490196	1.0	4
0	0.0	0.19117647	0.0	1

Paper Review (Barabasi & Oltvait, 2004)
Architectural Features of Cellular Networks

Global Toplogy Parameters

- **Components**: Set of nodes connected between them but disconnected to the other nodes.
- **Shortest Path:** The path with the smallest number of edges between two nodes
- The largest shortest path is the network **diameter**
- **Degree Distribution:** Counts the number of nodes with degree k and gives the probability that a selected node has exactly k edges. The degree distribution allows us to distinguish between different classes of networks.



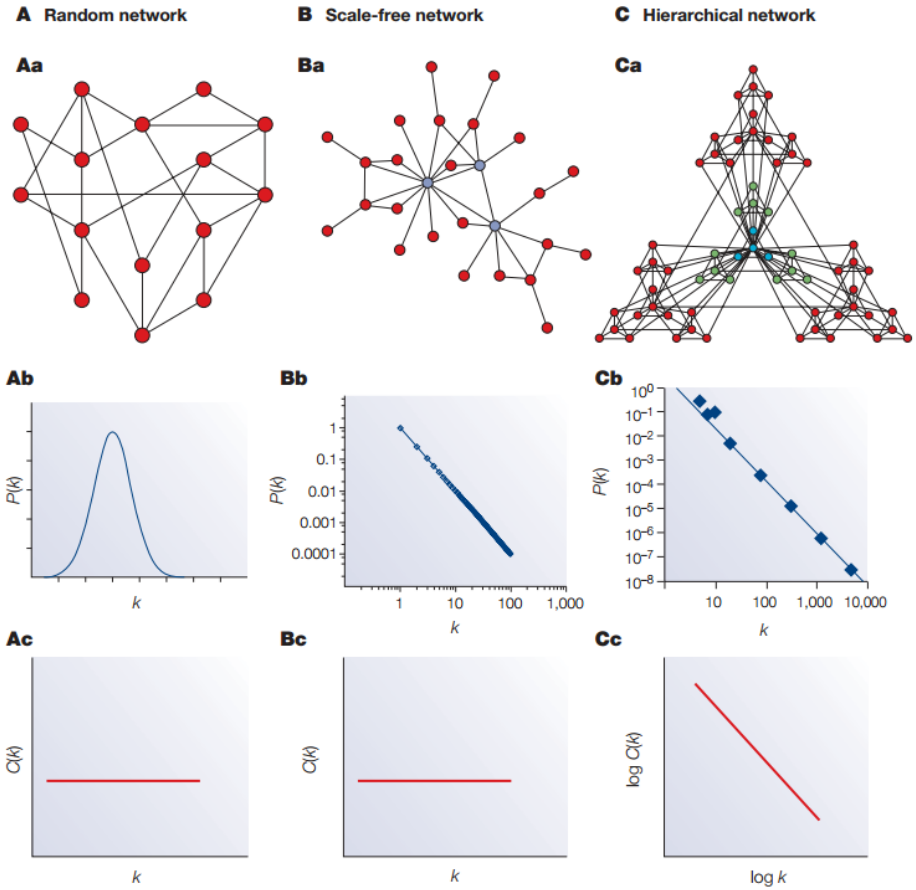
Paper Review (Barabasi & Oltvait, 2004)
Architectural Features of Cellular Networks

Network models The degree distribution allows us to **distinguish** between different classes of networks

The **Erdös–Rényi (ER) model** of a random network, which creates a graph with randomly placed links.

Scale-free networks are characterized by a power-law degree distribution; the probability that a node has k edges follows $P(k) \sim k^{-\gamma}$, where γ is the degree exponent.

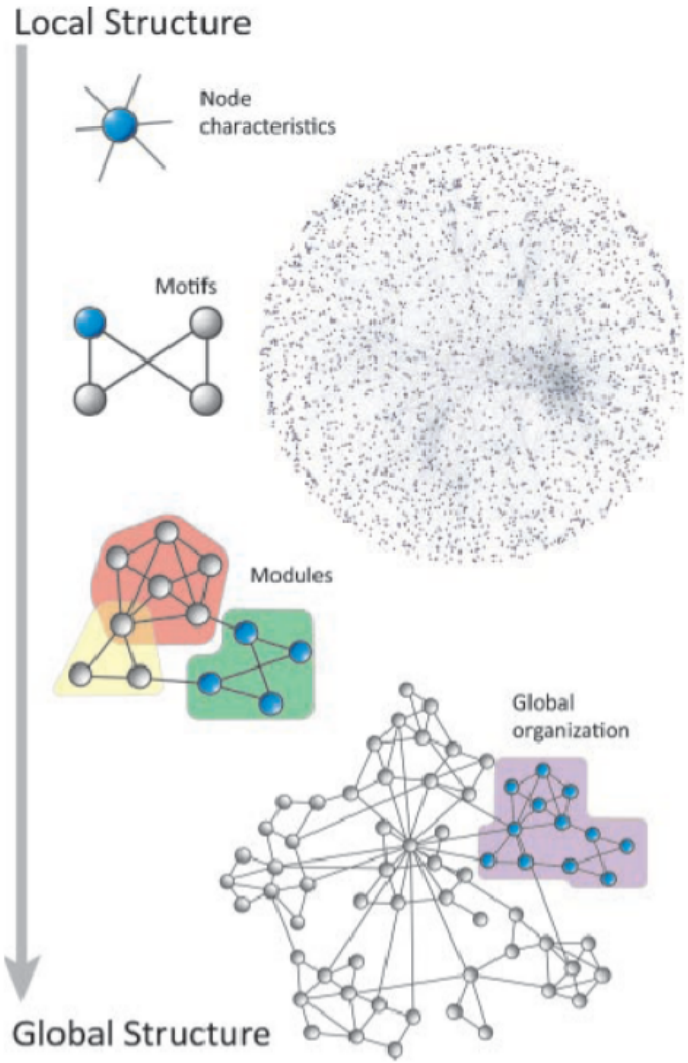
Hierarchical networks integrates a scale-free topology with an inherent modular structure by generating a network that has a power-law degree distribution scaling the clustering coefficient (Hard to see in noisy data)



Paper Review (Barabasi & Oltvait, 2004)
Motifs, modules and hierarchical Networks

Cellular functions are likely to be carried out in a highly modular manner. The levels of organization of complex networks:

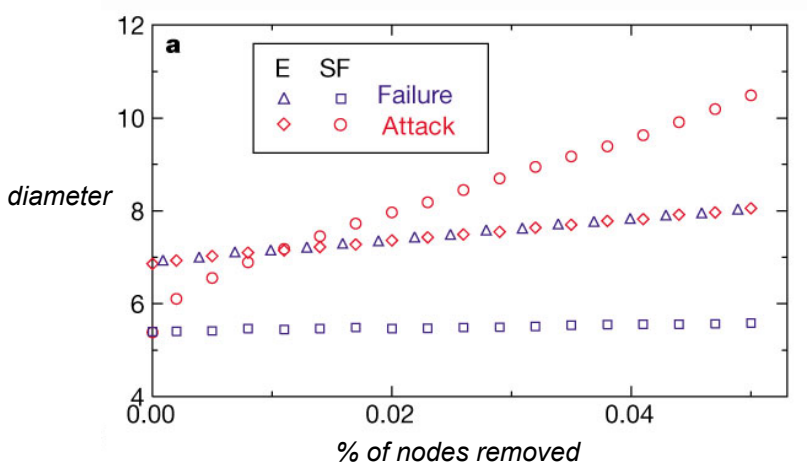
- **Hierarchy** describes how the various structural elements are combined
- Groups of highly interconnected nodes are called **modules** (or communities)
- Each module can reduced to a set of triangles (**motifs**). A high degree of triangles is reflected by the clustering coefficient (the signature of a netwrk potential modularity)



Paper Review (Barabasi & Oltvait, 2004)
Network Robusntess

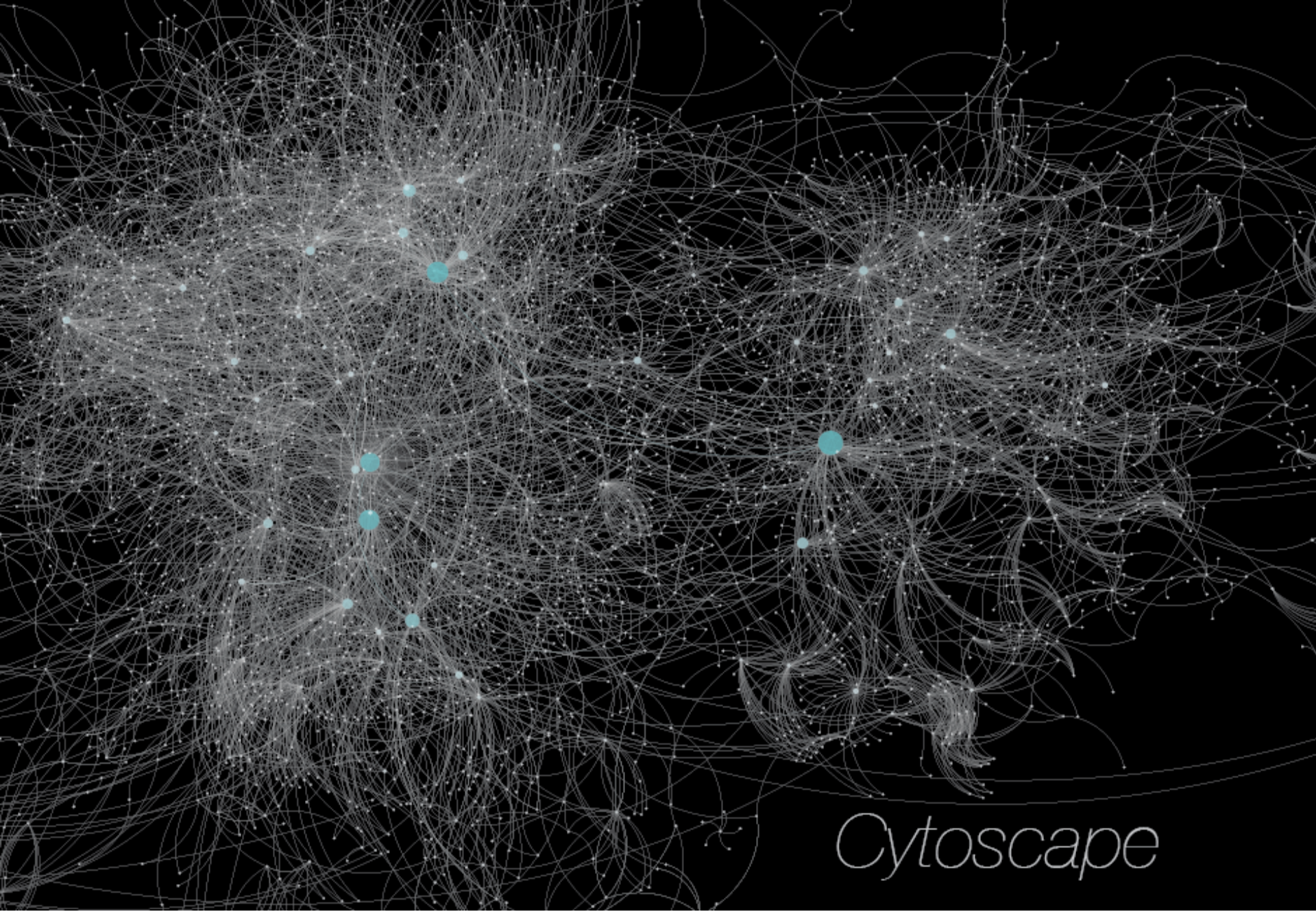
Scale-free topology provides **robustness** to biological networks. Robustness referes to the systems ability to respond to changes in the external conditions or internal organization while maintaining relatively normal behaviour.

Albert et al. (2000) compared the robustens of both random network and scale free networks to random node removal and differenciate between Failure and Attack.



Posterior studies showed that genes that are more “central” are more likely to:

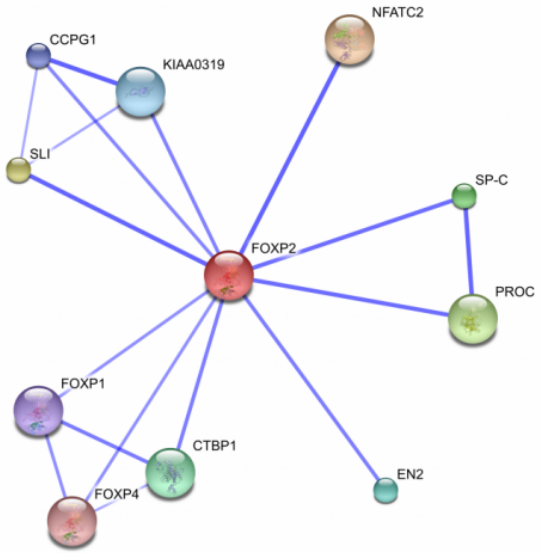
- be lethal when knocked out (Jeong et al. 2001)
- evolve more slowly (Fraser et al. 2002; Krylov et al. 2003; Hahn et al. 2004)


Cytoscape

Cytoscape
www.cytoscape.org

Software for Visualization, Integration & Analysis of Molecular Networks:

- Free & open source software application (LGPL license)
- Written in Java: can run on Windows, Mac, & Linux
- Developed by a consortium: UCSD, ISB, Agilent, MSKCC,
- Pasteur, UCSF, Unilever, UToronto
- Active community: mailing lists, annual conferences
- 10,000s users, 3000 downloads/month
- Extensible through **plugins** developed by third parties

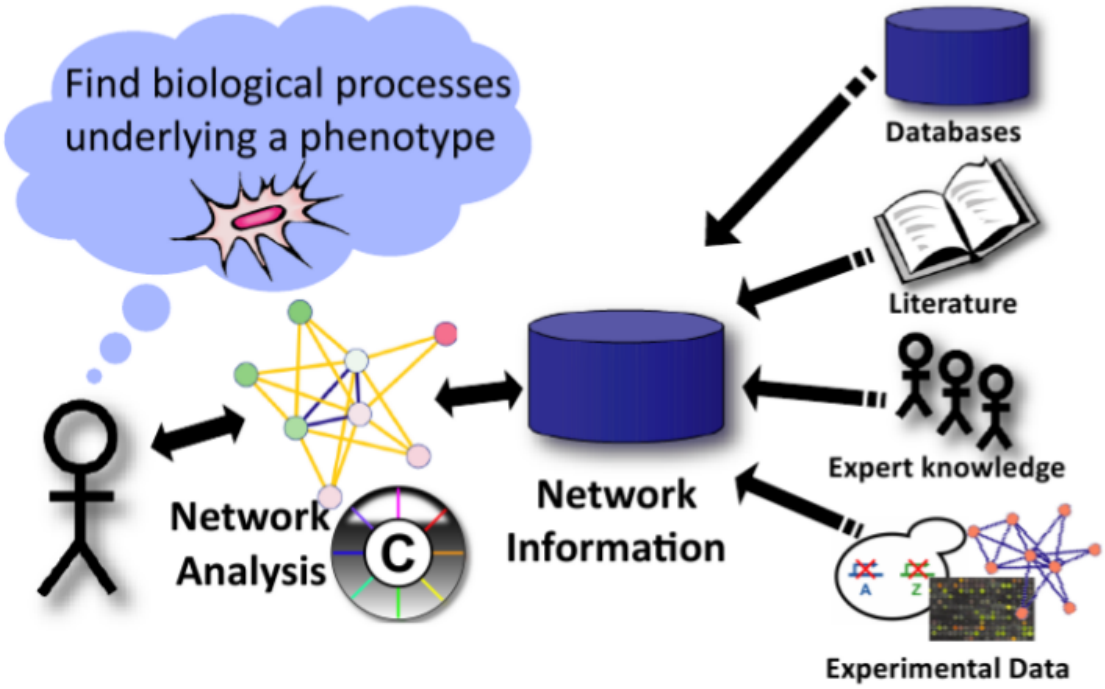


Cytoscape: a software environment for integrated models of biomolecular interaction networks
P Shannon, A Markiel, O Ozier, NS Baliga... - Genome ..., 2003 - genome.cshlp.org
Abstract **Cytoscape** is an open source software project for integrating biomolecular interaction networks with high-throughput expression data and other molecular states into a unified conceptual framework. Although applicable to any system of molecular **...**
Cited by 3617 Related articles All 15 versions Cite

Integration of biological networks and gene expression data using **Cytoscape**
MS Cline, M Smoot, E Cerami, A Kuchinsky... - Nature protocols, 2007 - nature.com
Abstract **Cytoscape** is a free software package for visualizing, modeling and analyzing molecular and genetic interaction networks. This protocol explains how to use **Cytoscape** to analyze the results of mRNA expression profiling, and other functional genomics and **...**
Cited by 682 Related articles All 12 versions Cite

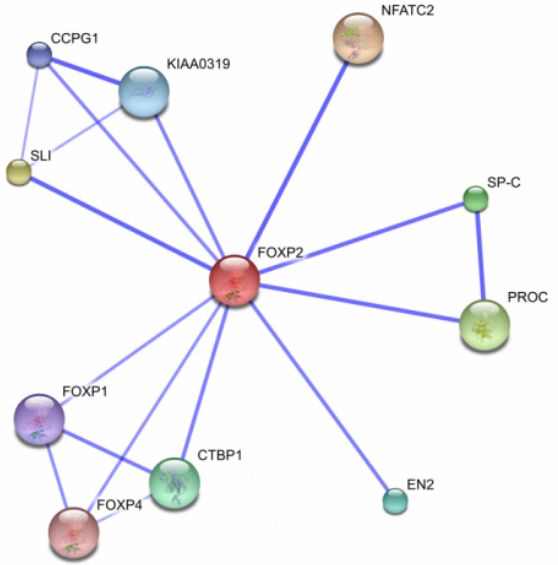
Cytoscape
Why?

Allows the integration of multiple data types.



Cytoscape
A Typical Cytoscape Workflow

1. Load Networks (Import network data into Cytoscape)
2. Load Attributes (Get data about networks into Cytoscape)
3. Analyse and Visualise Networks
4. Prepare for Publication



Cytoscape
Loading networks

- Supported network file formats: SIF, GML, XGMML, BioPAX, PSI-MI 1 & 2.5, SBML Level 2, KGML
- Use import network from table:
  - Excel file
  - Comma or tab delimited text
- Use import network from web services
  - IntAct (EMBL EBI)
  - Reactome
  - Pathway commons (collection of sources)
  - NCBI Entrez EUtilities
- Use a plugin from the Network Inference category
  - AgilentLiteratureSearch

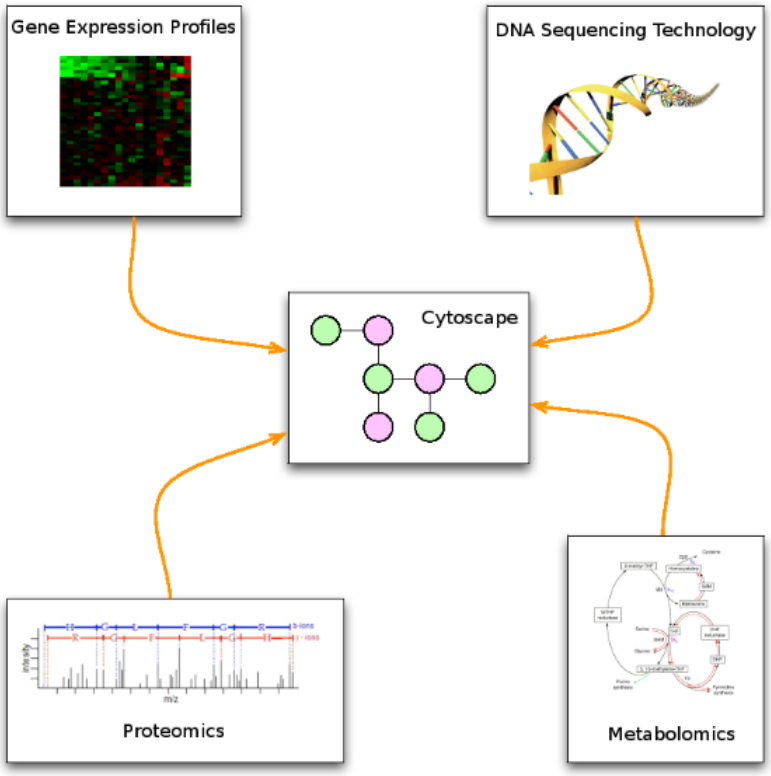
Cytoscape
Loading networks

Tips & Tricks: Network View

- When you open a large network, you will not get a view by default
- To improve interactive performance, Cytoscape has the concept of ”Levels of Detail”
- Some visual attributes will only be apparent when you zoom in
- The level of detail for various attributes can be changed in the preferences
- To see what things will look like at full detail: View → Show Graphics Details

Cytoscape
Loading Attributes

Nodes and edges can have **attributes** associated with them:



- Cytoscape supports multiple data types: Numbers, Text, Logical, Lists..
- Upload:
  - table exel file
  - tab delimited text
  - expression matrix
- Import attribute from web services
  - NCBI Entrez Gene
  - Ensembl Biomart
- Create attributes manually in the attribute browser

Cytoscape
Loading Attributes

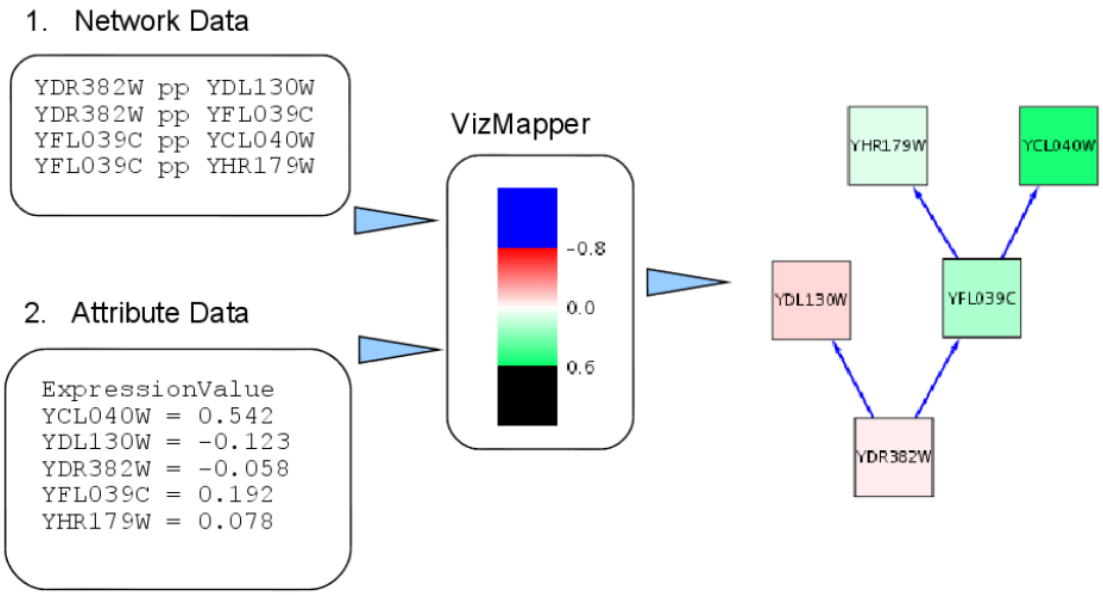
Tips & Tricks: Root Graph and Sessions

- You can have multiple graphs but ”There is one graph to rule them all...”
- The networks in Cytoscape are all ”views” on a single graph.
- Changing the attribute for a node in one network will also change that attribute for a node with the same ID in all other loaded networks
- There is no way to ”copy” a node and keep the same ID
- Make a copy of the session

Cytoscape
Visualize Networks

Data Mapping

Mapping of data values associated with graph elements onto graph visuals

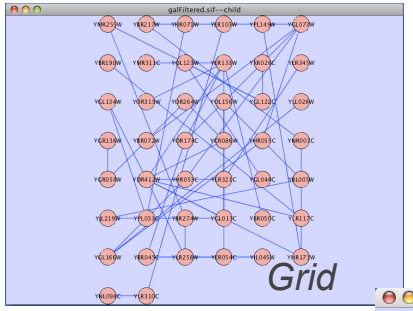


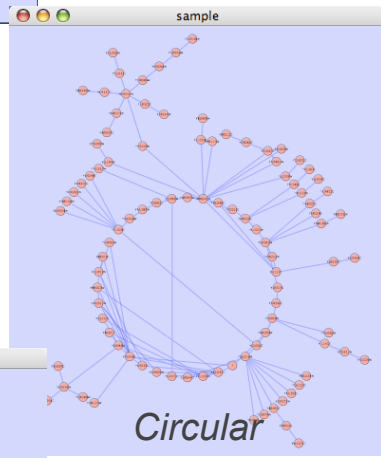
- Visual attributes
  - Node fill colour, border colour, border width, size, shape, opacity, label
  - Edge type, colour, width, ending type, ending size, ending colour
- Mapping types
  - Passthrough (labels)
  - Continuous (numeric values)
  - Discrete (categories)

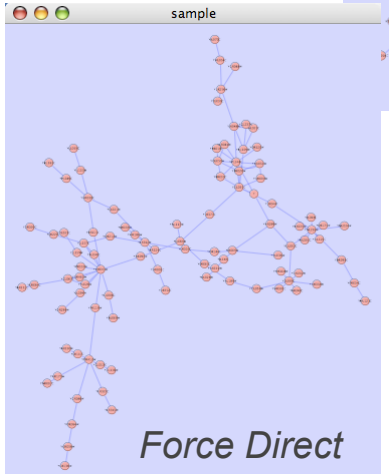
Cytoscape
Visualize Networks

Layouts 15 algorithms available through plugins

- Simple: grid, not very informative
- Hierarchical: layout data as a tree or hierarchy (works best when there are no loops)
- Circular (Radial): arrange nodes around a circle, could use node attributes to govern position, e.g. degree sorted
- Force-Directed: simulate edges as springs (nodes repel and edges pull), good for up to 500 nodes
- Multi-layer layouts: partition graph, layout each partition then layout partitions







Cytoscape
Visualize Networks

Filtering & Editing Data

- QuickFind nodes and edges
- Filtering according the value in an attribute
- Build complex filters using AND, OR, NOT relations
- Define topological filters (considers properties of near-by nodes)

Cytoscape
Analyze Networks

Network topology statistics such as node centrality, betweenness, degree distribution of nodes, clustering coefficient, shortest path between nodes and robustness of the network to the random removal of single nodes.

Modularity refers to the identification of sub-networks of interconnected nodes that might represent molecules physically or functionally linked that work coordinately to achieve a specific function.

Motif analysis is the identification of small network patterns (or subgraphs) that are over-represented when compared with a randomized version of the same network. Discrete biological processes such as regulatory elements are often composed of such motifs.

Network alignment and comparison tools can identify similarities between networks (such as common subgraphs) and have been used to study evolutionary relationships between protein networks of organisms.

Cytoscape
Extended Functionality

Cytoscape extends its functionality with plugins:

- Developed by third parties
- Listed at http://chianti.ucsd.edu/cyto_web/plugins/
- Major categories:
  - Online databases import
  - Graph Analisis
  - Genomics Integrated Analysis
  - Clustering
  - Visualization
  - Network Generation
- Usually available through
  - Plugin Manager in the toolbar
  - Download from the plugins's websites

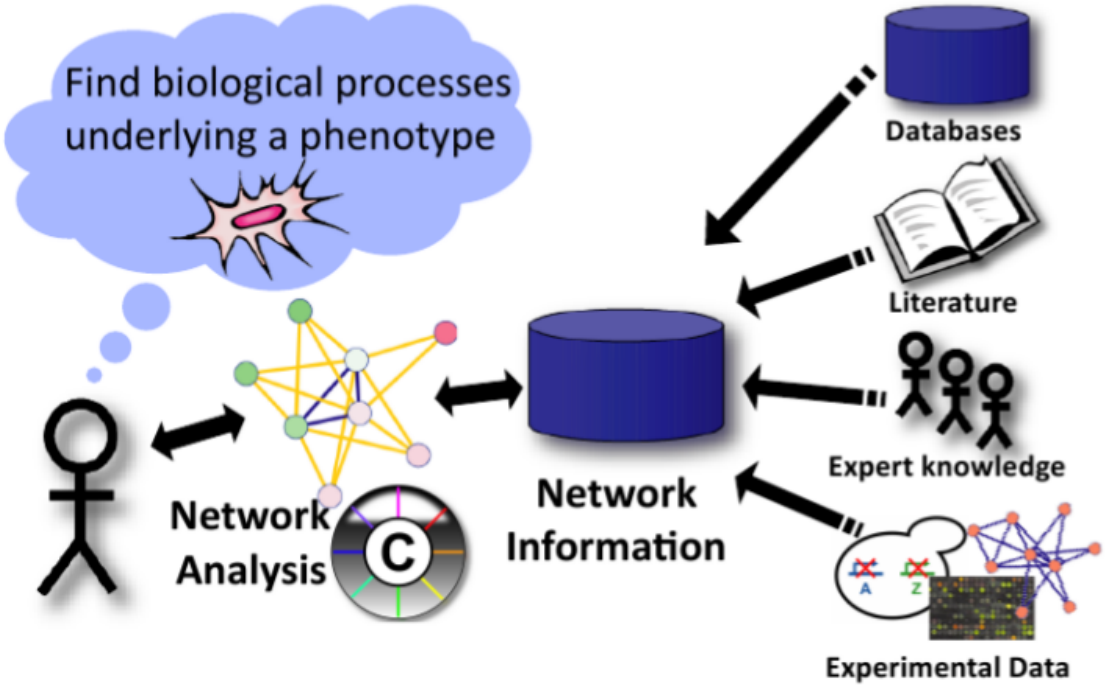
Cytoscape
Exporting & Saving Data

- Sessions save pretty much everything:
  - Networks
  - Properties
  - Visual styles
  - Screen
- Export networks in different formats: SIF, GML, XGMML, BioPAX, PSI-MI 1 & 2.5
- Publication quality graphics in several formats: PDF, EPS, SVG, PNG, JPEG, and BMP
- Export Session to HTML for Web (SessionForWeb plugin)

Cytoscape
Some Useful Links

- Documentation: http://www.cytoscape.org/documentation_users.html
- Plugins: http://chianti.ucsd.edu/cyto_web/plugins/
- Mailing lists: http://www.cytoscape.org/community.html
- Presentations & Tutorials: http://cytoscape.wodaklab.org/wiki/Presentations

Cytoscape
Ready?

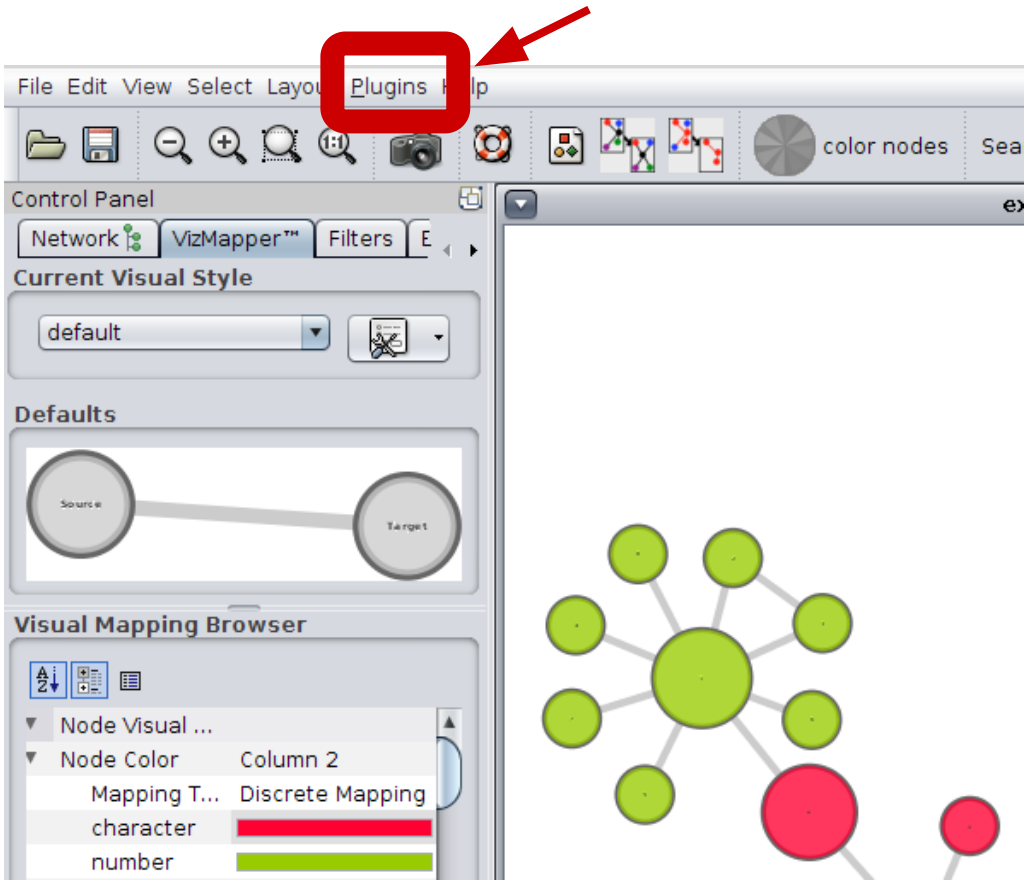


Cytoscape
Install Cytoscape

http://www.cytoscape.org/download.html

1. Register and Accept Conditions
2. Go to Older Versions
3. Download Version 2.8.2: *cytoscape-2.8.2.tar.gz*
4. Run in the terminal
    a. $ tar -xvf cytoscape-2.8.2.tar.gz
    b. $ ./cytoscape.sh

Cytoscape
Install Cytoscape Plugins



Search and Install

- allegroMCODE
- Network Analysis
- MultiColoredNodes

Hands on
An Example

Download example.sif, example.attr, example:expressionMat.txt
Network is in sif format (2 columns plain text, every row represents an edge)

Steps:

1. Upload the networks (*File* > *Import* > *Import Network from Table*)
2. PLEASE Customize view: change node, edge ans backgroud color (*VizMapper* > *Defaults*)
3. Apply the Force Directe Layout (*Layout* > *Cytoscape Layouts*)
4. Upload Node Attributes (*File* > *Import* > *Import Attributes from Table*)
5. Upload Expression (File > Import > Attribute/Expression Matrix)
6. Import gene HGNC symbol from Biomart (*File* > *Import* > *Import data from Biomart*)
7. Change Node Label according the gene symbol (*VizMapper* > *Node Color* > *HGNC symbol*)
8. Change Node Shape according attribute “type” from example.attr (*VizMapper* > *Node Color* > *type*)
9. Analyze Topology (*Plugins* > *Network Analysis* > *Analyze Network*)
10. Analyze Modularity (*Plugins* > *allegroMCODE* > *Analyze*)
11. Color Nodes by expression Value (*Plugins* > *MulticoloredNodes* > *Absolute MulticoloredNodes*)

Questions:

1. What degree distribution follows the network?
2. Which is the diameter, number of components and the average path length of every network?
3. Which node has more betweenness, closeness, degree and clustering coefficient?
4. Is there any module?
5. What can you say after mapping the exprssion values onto the network?

Exercises
1. Topology Study

Download the 3 networks: network1.sif, network2.sif, network3.sif

Steps:

1. Upload the networks (*File* > *Import* > *Import Network from Table*)
2. Customize view: change node, edge ans backgroud color (*VizMapper* > *Defaults*)
3. Apply the Force Directe Layout (*Layout* > *Cytoscape Layouts*)
4. Analyze Topology (*Plugins* > *Network Analysis* > *Analyze Network*)
5. Analyze Modularity (*Plugins* > *allegroMCODE* > *Analyze*)

Questions:

1. What degree distribution follows every network? Which degree distribution model do they fit the best?
2. Which is the diameter, number of components and the average path length of every network?
3. Which network would be more damaged after removing the node 24?
4. Which network is more modular (have more highly scored clusters)?

Exercises
2. Data Integration

Imagine you are working for a wet lab that is pretty interested in the apoptosis changes between two conditions. You're provided of:

- a gene expression matrix with three replicates for 2 conditions *apoptosisAssay_expressionMat.txt*
- a protein interactome *protein_interactome.sif*
- a protein vs Gene Ontology for Biological Process annotation *protein_GObp.annot*

Integrate gene expression data with the protein interactome focusing on apoptosis related genes (filtering by GO) and represent it using Cytoscape. Export the final network view that you obtain and comment it (like a paper figure footer). For the FINAL plot, change the node labels in order to show HGNC symbols.

Notes

1. Expression matrix contains fold chages relatives to an external control. You DO NOT need to perform any analysis. 0 means that the expression is equal to the control and negative values mean the expression is below the expression in the external control.
2. Array probes have already been merged and translated to ENSP identifiers. Also protein interactome and protein-GO annotation are in ENSP identifiers.