Data Integration with Cytoscape

Samad Lotia

L'institut Pasteur

6 May 2009

The Problem

Given a biological problem:

- The "data integration" part: data from disparate sources
- Now that our data is "integrated", what do we do with it? We wish to draw conclusions from this data *collectively*

Why Integrate Data in the First Place?

- Many instruments for observing biological phænomena are noisy:
  - Gene expression profiles can very noisy if expression levels are low[1]
  - Up to 50% of yeast two–hybrid data are false[2]
- If one can observe the same hypothesis from a variety of instruments, this improves the chance of the hypothesis' validity

[1] Quackenbush, H. Weighing our measures of gene expression. *Mol. Syst. Biol.* 2, 63 (2006).

[2] Sprinzak, E., Sattath, S., & Margalit, H. How reliable are experimental protein–protein interaction data? *J. Mol. Biol.* 327, 919—923 (2003).

How Are We Going to Integrate Data?

- In biology we have components that interact or influence other components, that in turn influence other components, and so on.
  - Genes
  - Proteins
  - Pathways
- The essence of this talk: *Use a network abstraction as a basis for bringing together disparate data sources and to draw conclusions from them for biological problems*.

What Can We Do with a Network Abstraction?

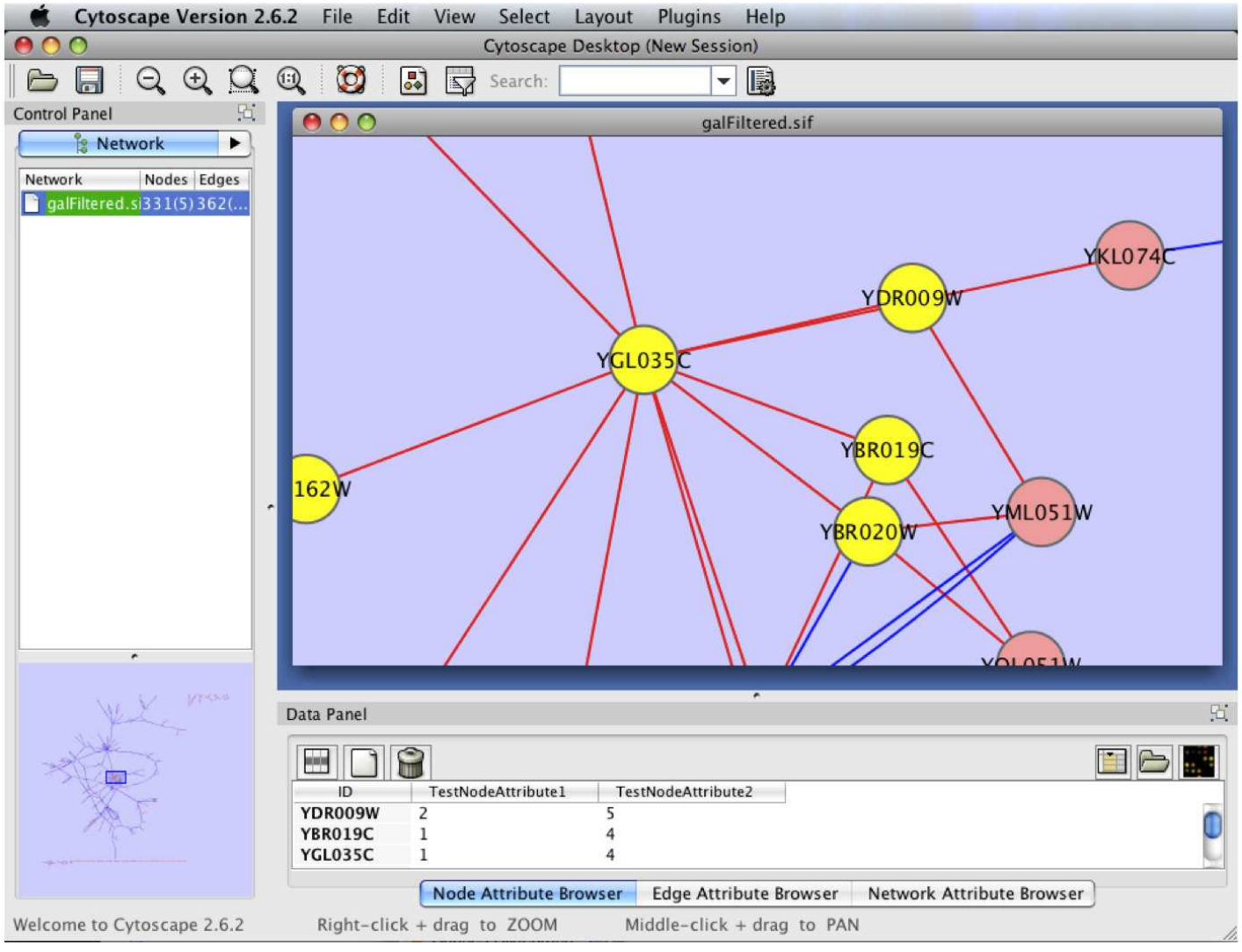
- By using networks to tackle biological problems, we can use Cytoscape to:
  - Bring together data from a variety of places
  - Visualize interactions of biological components
  - Analyze networks to determine what parts are biologically significant

Some Background on Cytoscape

- Free & open source software application—LGPL license
- Written in Java—can run on Windows, Mac, & Linux
- Core development made by:
  - Agilent Technologies (Santa Clara, California, USA)
  - Eötvös Loránd University (Budapest, Hungary)
  - Institute for Systems Biology (Seattle, Washington, USA)
  - Memorial Sloan–Kettering Cancer Center (New York, New York, USA)
  - National Center for Integrative Biomedical Informatics (Ann Arbor, Michigan, USA)
  - L'institut Pasteur (Paris, Île-de-France, France)
  - University of Toronto (Toronto, Ontorio, Canada)
  - University of California, San Diego (La Jolla, California, USA)
  - University of California, San Francisco (San Francisco, California, USA)
  - Unilever (London, UK)

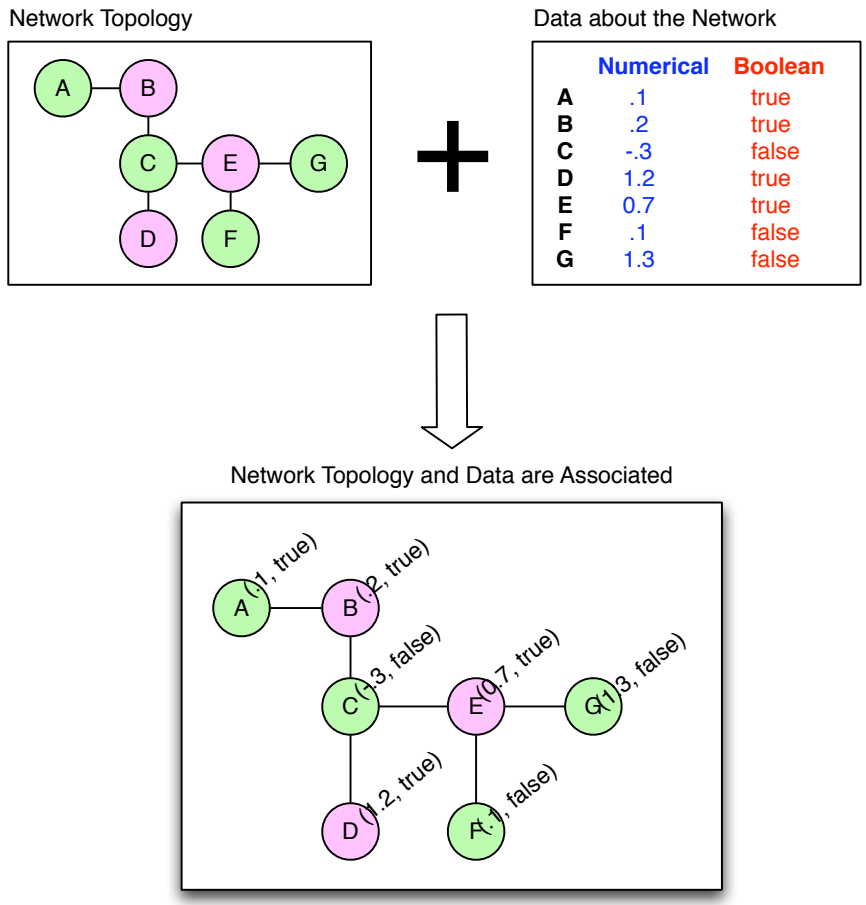
Cytoscape's Core Functionality

- Visualizing networks



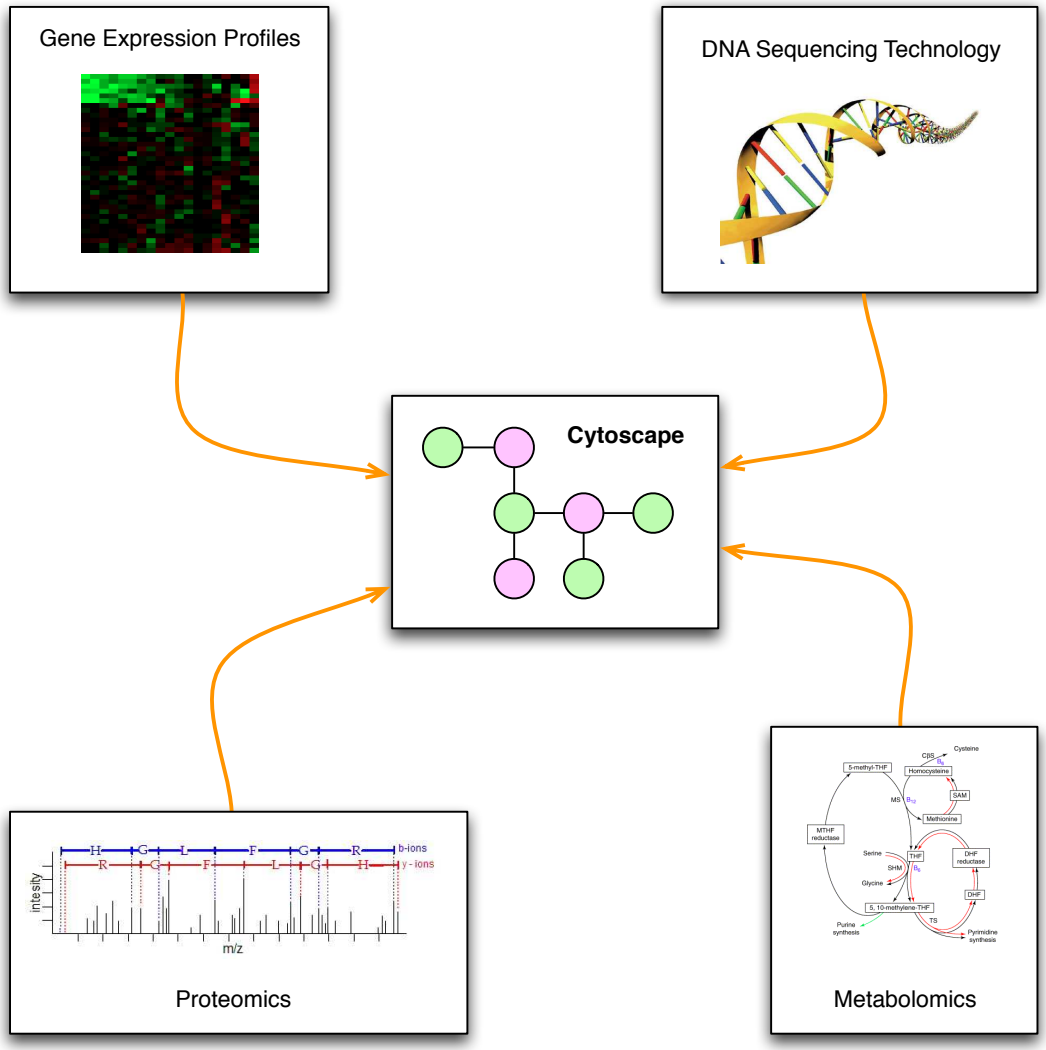
Cytoscape's Core Functionality

- Nodes and edges can have *attributes* associated with them
- Attributes and corresponding nodes and edges are *dynamically* bound



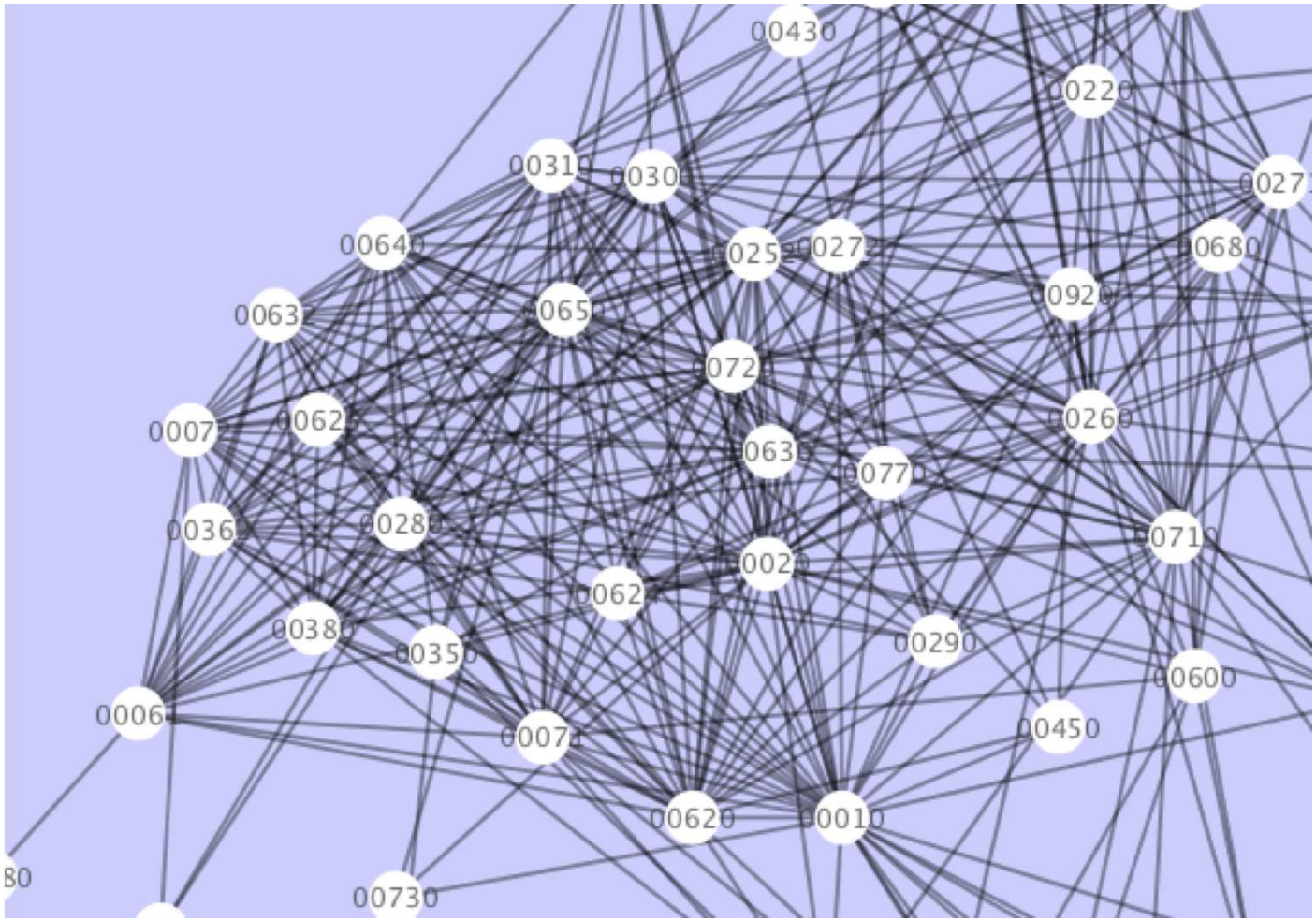
Cytoscape's Core Functionality

- Agnostic semantics
- We can overlay data from a variety of instruments on top of a network



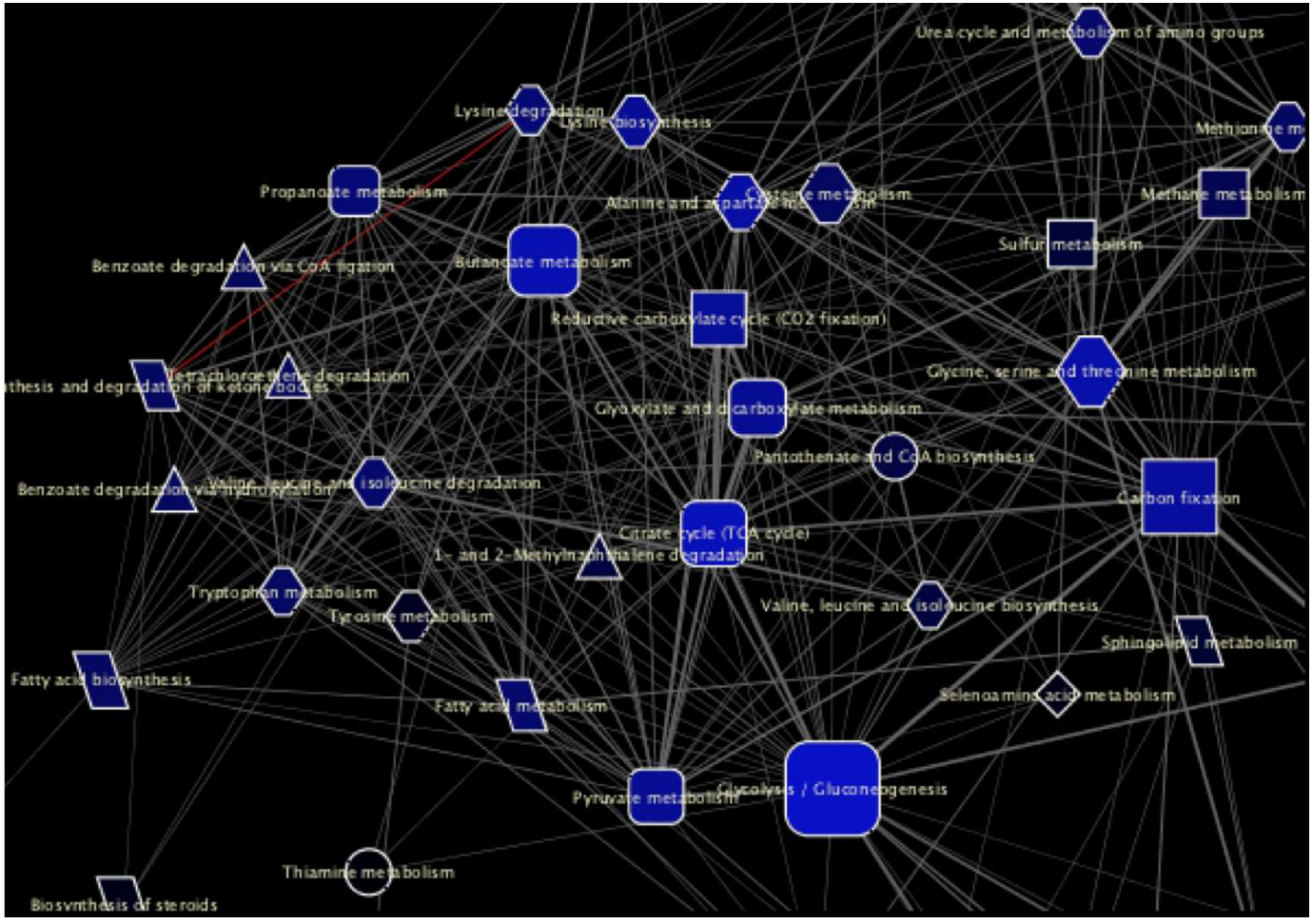
Cytoscape's Core Functionality

- VizMapper: map network *attributes* to visual properties of the network
- A network without a visual mapping:



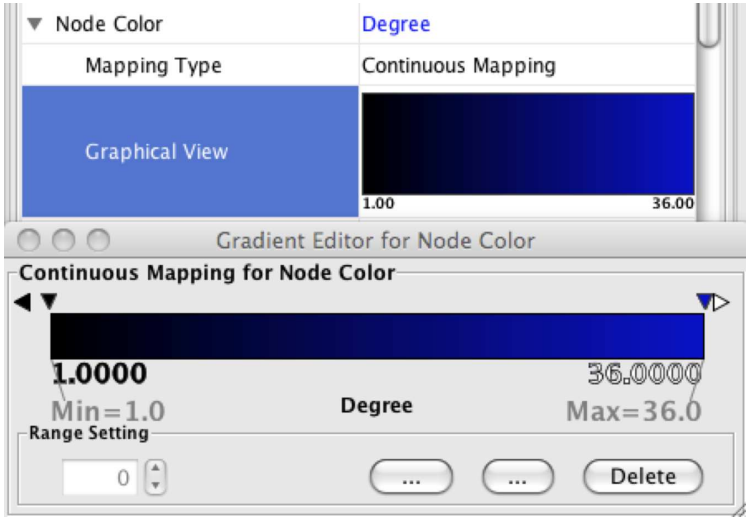
Cytoscape's Core Functionality

- Same network but with a visual mapping:



Cytoscape's Core Functionality

- VizMapper's continuous mapping:

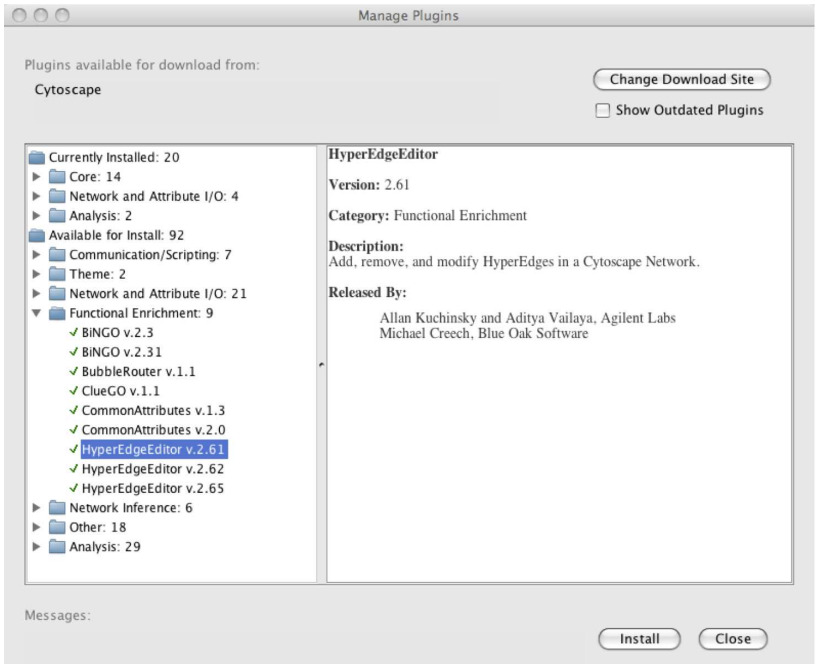


- VizMapper's discrete mapping:

Node Shape	Category
Mapping Type	Discrete Mapping
Amino Acid Metabolism	HEXAGON
Biosynthesis of Secondary ...	TRIANGLE
Carbohydrate Metabolism	ROUND_RECT
Energy Metabolism	RECT
Glycan Biosynthesis and M...	OCTAGON
Lipid Metabolism	PARALLELOGRAM
Metabolism of Cofactors a...	ELLIPSE
Metabolism of Other Amin...	DIAMOND
Nucleotide Metabolism	HEXAGON
Xenobiotics Biodegradatio...	TRIANGLE

Cytoscape's Extended Functionality

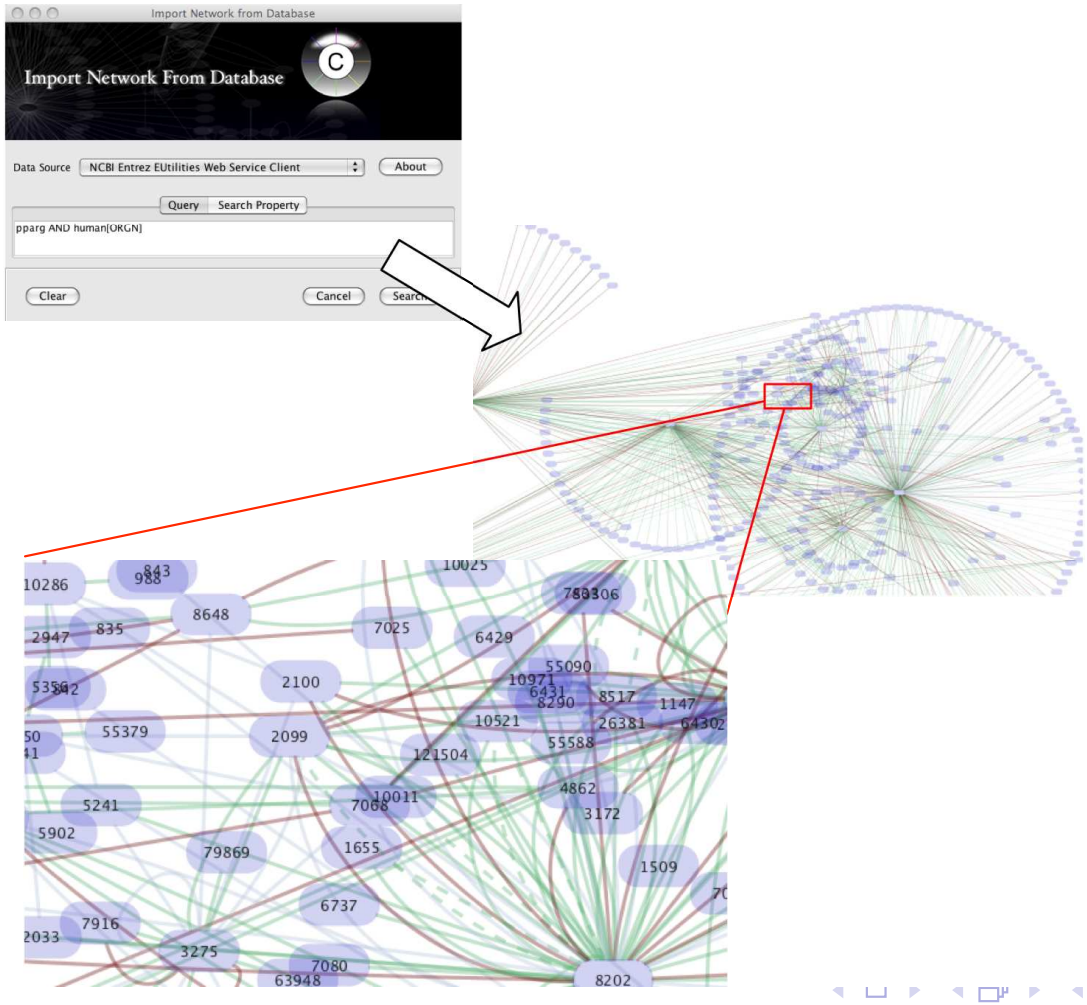
- Cytoscape extends its core functionality with *plugins*
- Developed by third parties
- Some major categories of plugins:
  - Obtain networks from online databases
  - Import node and edge attributes from online databases
  - Perform algorithmic analysis of networks



Cytoscape's Extended Functionality: Networks from Online Databases

- We can import networks from NCBI Entrez and Pathway Commons

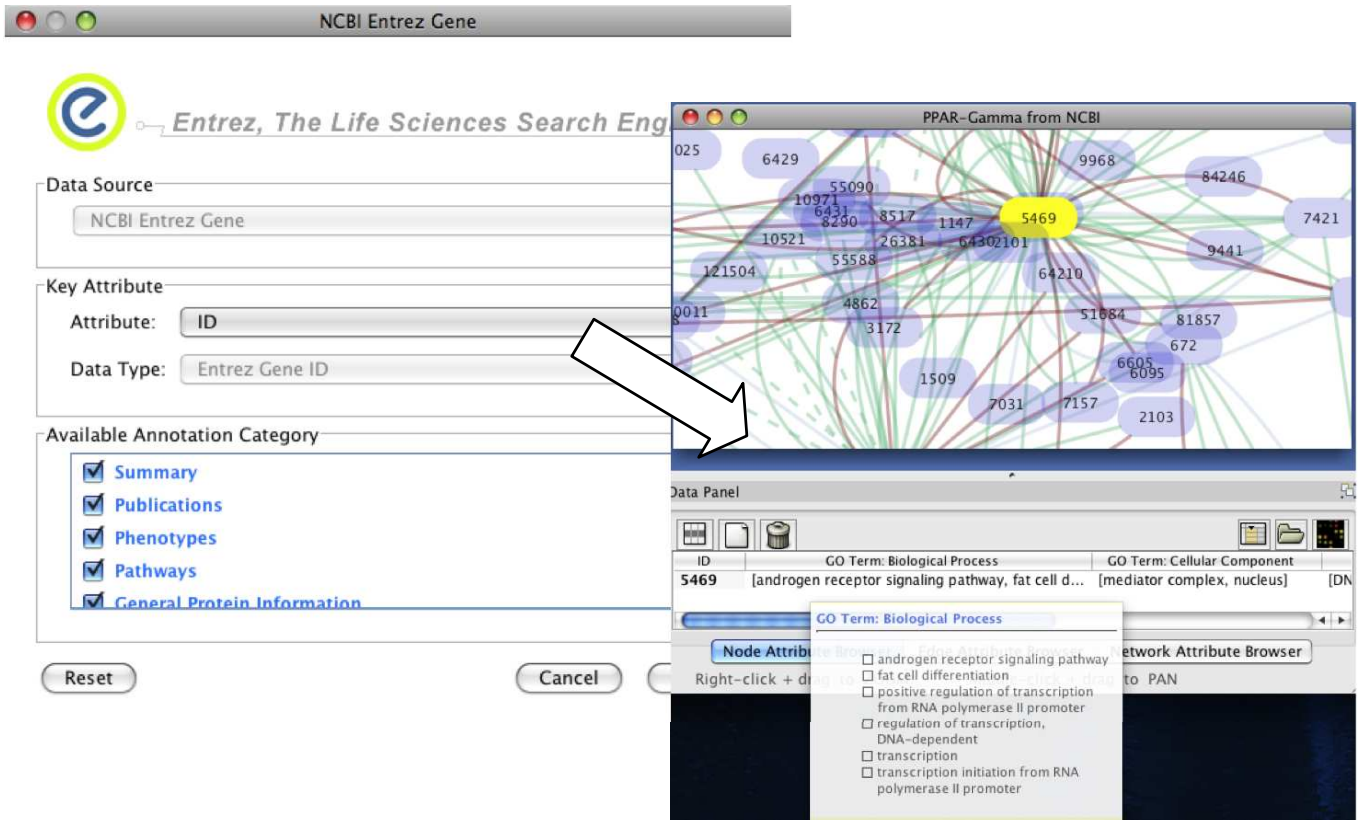
Obtaining Interactions of PPAR-Gamma in Humans from NCBI Entrez



Cytoscape's Extended Functionality: Node and Edge Attributes from Online Databases

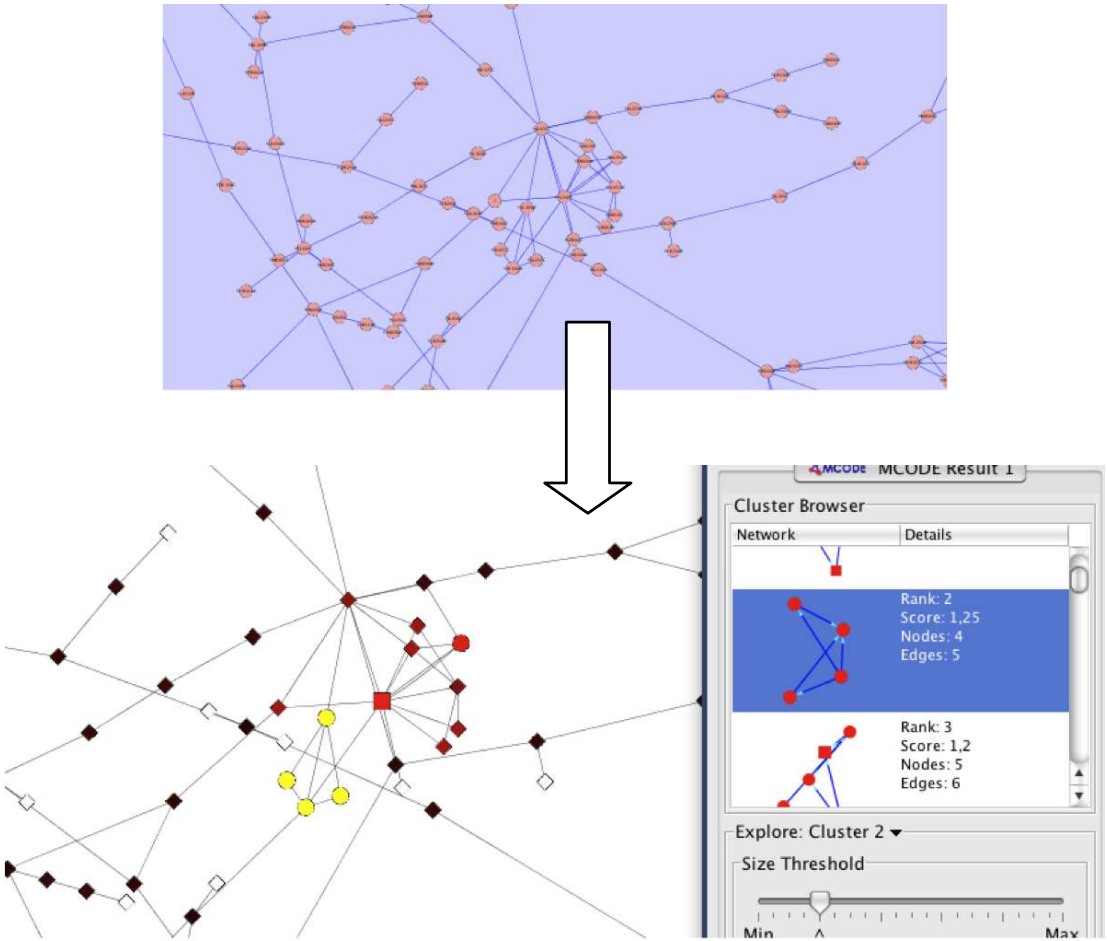
- Given a network where nodes are database IDs: overlay information about nodes onto network from a database

Obtain Node Information from NCBI Entrez



Cytoscape's Extended Functionality: Network Analysis

- Extract subnetworks and score them; the means to do this could be done by looking at:
  - Network topology
  - Values of node or edge attributes



Problem of Glioblastoma

- Study of glioblastoma, a common form of brain cancer[3]
- Biological problem: People newly diagnosed with glioblastoma have a variety of genetic aberrations.
- How can one develop treatments if there are a large number of genetic aberrations?

[3] Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblatoma genes and core pathways. *Nature* vol. 455 (23 Oct 2008).

Data of Glioblastoma Patients

- Analyzed biospecimens from people in early stages of glioblastoma
- Looked at copy number alterations (CNAs) $\Rightarrow$ about $\frac{3}{4}$ of affected genes had expression levels proportional to number of copies
- Determined what genes had non–silent and silent mutations
  - p53 (regulates cell growth cycle) was mutated
  - NF1 (tumor suppressor gene) was inactivated or deleted
  - EGFR (signals the cell to grow) was activated
  - PI(3)K complex (signals a transition in cell cycle) was mutated
  - MGMT (DNA repair enzyme) was methylated, thus reducing its expression

		P53 signaling				RB signaling							RTK-Ras-PI3K-AKT signaling network										PI3K class 2
	Tumor #	CDKN2A (ARF)	MDM2	MDM4	TP53	CDKN2A (P16)	CDKN2B	CDKN2C	CDK4	CDK6	CCND2	RB1	EGFR	ERBB2	PDGFRA	MET	SPRY2	RAS	NF1	AKT/FOXO	PI3KCA	PI3KR1	PTEN	PIKC2A/B/G
untreated and primary	02-0003				2 mutations				amplified				amp+mut									mutation		
	02-0006		amplified						amplified														homo del	
	02-0009	homo del				homo del	homo del						amplified											
	02-0011				mutation																			
	02-0027	homo del				homo del	homo del																del+mut	
	02-0033				del+mut							del+mut							del+mut					
	02-0034	homo del			2 mutations	homo del	homo del																homo del	
	02-0037	homo del			mutation	homo del	homo del														mutation			
	02-0038	homo del				homo del	homo del						amplified										del+mut	
	02-0046				2 mutations				amplified				amp+mut								mutation			
	02-0047	homo del				homo del	homo del														mutation			
	02-0052		amplified									del+mut											del+mut	
	02-0054				mutation								mutation											
	02-0055				del+mut							del+mut							homo del				del+mut	
	02-0060		amplified						amplified															
	02-0064	homo del				homo del	homo del						amplified											
	02-0069													2 mutations	amplified		homo del	KRAS mut						
	02-0071	homo del	amplified	amplified		homo del	homo del						amplified										del+mut	C2B amp
	02-0074				mutation			homo del	amplified						amplified							mutation		
	02-0075				mutation														mutation				del+mut	
	02-0085	homo del	mutation			homo del	homo del											NRAS mut						
	02-0086	homo del				homo del	homo del												2 mutations					
	02-0115	homo del				homo del	homo del							3 mutations					mutation					
	06-0122	homo del				homo del	homo del						amplified											
	06-0124	homo del				homo del	homo del						amplified						2 mutations				del+mut	
	06-0125	homo del				homo del	homo del						amp+mut										del+2 mut	
	06-0126	homo del				homo del	homo del						amplified											
	06-0128				mutation				amplified															
	06-0129			mutation					amplified													mutation		
	06-0130				del+mut																			
	06-0132												amplified											
	06-0133	homo del				homo del	homo del						amplified									mutation		
	06-0137	homo del				homo del	homo del						amplified											C2G mut
	06-0138		amplified						amplified				amplified										del+mut	
	06-0139																							
	06-0141																							
	06-0143	homo del		amplified		homo del	homo del						amplified		amplified									C2B amp
	06-0145	homo del				homo del	homo del		amplified				amplified											C2G mut
	06-0147																							
	06-0148	homo del				homo del	homo del						amp+mut							AKT1 amp				
	06-0154	homo del				homo del	homo del						amp+mut											
	06-0156	homo del			del+mut	homo del	homo del						amp+2 mut		amplified								del+mut	
	06-0157			amplified					amplified		amplified		amp+mut										del+mut	C2B amp
	06-0158	homo del				homo del	homo del						amplified											
	06-0166	homo del				homo del	homo del												homo del		mutation		del+mut	
	06-0168																						del+mut	
	06-0169	homo del				homo del	homo del						amplified											
	06-0171	homo del				homo del	homo del														mutation			
	06-0173		amplified						amplified				amplified							AKT3 amp				
	06-0174		amplified						amplified				amplified		amp+mut									
	06-0176											mutation											mutation	
	06-0178																							
	06-0184				del+mut							del+2 mut												
	06-0185	homo del				homo del	homo del						amp+mut			amplified			del+mut				del+mut	
	06-0187	homo del	amplified			homo del	homo del						amplified										del+mut	
	06-0188	homo del			del+2 mut	homo del	homo del	homo del															del+mut	
	06-0189																							
	06-0190				del+mut														mutation				del+mut	
	06-0195				2 mutations							del+mut											del+mut	
	06-0197				del+mut																		del+2 mut	
	06-0201	homo del				homo del	homo del												2 mutations					C2G mut
	06-0206				del+mut							del+mut							homo del				del+mut	
	06-0208	homo del				homo del	homo del						amplified											
	06-0209		amplified						amplified				amplified											C2A mut
	06-0210	homo del		amplified		homo del	homo del								amplified									C2B amp
	06-0211	homo del				homo del	homo del						amplified		amplified									
	06-0213				homo del							del+mut											del+mut	C2G mut
	06-0214	homo del	amplified			homo del	homo del						amplified		amplified				mutation				homo del	
	06-0219	homo del				homo del	homo del						amplified										homo del	
	06-0221				2 mutations				amplified	amplified	amplified				amplified									
	06-0237				mutation								amp+mut											
	06-0241	homo del			del+mut	homo del	homo del								amplified									
	02-0001	homo del			mutation	homo del	homo del					homo del					homo del		mutation				mutation	
	02-0007			amplified								del+mut	mutation										del+mut	C2B amp
	02-0010				2 mutations	del+mut							mutation			mutation			mutation					
	02-0014	homo del			mutation	homo del	homo del								amplified								del+mut	
	02-0021	homo del				homo del	homo del						amp+mut									mutation	del+mut	
	02-0024	homo del			del+mut	homo del	homo del								amplified	amplified								
	02-0028																		3 mutations			mutation		
	02-0043	homo del				homo del	homo del						amplified										homo del	C2B mut
	02-0057	homo del				homo del	homo del																	
	02-0058				mutation																			
	02-0080				mutation									mutation										
	02-0083	del+2 mut			mutation	del+mut							amplified							FOXO3+4 mut				
	02-0089	homo del			mutation	homo del							amp+mut	2 mutations								mutation		
	02-0099		amplified						amplified														del+mut	
	02-0102		amplified						amplified				amplified											
	02-0107													mutation					2 mutations					
	02-0113													2 mutations									del+mut	
	02-0114				2 mutations	del+mut										amplified						mutation		
	02-0116	homo del				homo del	homo del						amp+mut	mutation							mutation			
	gene altered all	49.5%	14.3%	6.6%	35.2%	51.6%	47.3%	2.2%	17.6%	1.1%	2.2%	11.0%	45.1%	7.7%	13.2%	4.4%	2.2%	2.2%	17.6%	3.3%	6.6%	8.8%	36.3%	
	gene altered not treated	50.0%	15.3%	6.9%	31.9%	50.0%	50.0%	2.8%	19.4%	1.4%	2.8%	11.1%	45.8%	2.8%	13.9%	1.4%	1.4%	2.8%	16.7%	2.8%	6.9%	5.6%	36.1%	
	gene altered treated	47.4%	10.5%	5.3%	47.4%	57.9%	36.8%	0.0%	10.5%	0.0%	0.0%	10.5%	42.1%	26.3%	10.5%	15.8%	5.3%	0.0%	21.1%	5.3%	5.3%	21.1%	36.8%	
	pathway altered all	86.8%				78.0%							87.9%											12.1%
	pathway altered not treated	87.5%				79.2%							87.5%											12.5%
	pathway altered treated	84.2%				73.7%							89.5%											10.5%
	RTK altered												59.3%											

What Can We Do with the Data?

- Problem: data from a variety of sources: CNAs, mutations of many genes
- How can one draw conclusions from disparate data sources?
- Solution they used: project this data onto an established glioblastoma pathway network[4]

[4] Furnari, F. B. *at al.* Malignant astrocytic glioma: genetics, biology, and paths to treatment. *Genes Dev.* 21, 2683-2710 (2007).

What Can We Learn from the Data?

- Almost all patients had mutations in any of these four pathways
- These pathways are *highly* interconnected
- While a variety of these genes were affected by different types of mutations, *these genes are strongly related to each other*

Concluding Remarks

We can use Cytoscape for bringing data together under a network paradigm; we could potentially improve our understanding of a network by:

- obtaining metadata from a variety of online databases,
- visualizing it,
- and performing algorithmic analysis of it.