Integration of Heterogeneous Data

Gary Bader

Terrence Donnelly Center for Cellular and Biomolecular Research (CCBR)

BBDMR, SLRI

University of Toronto

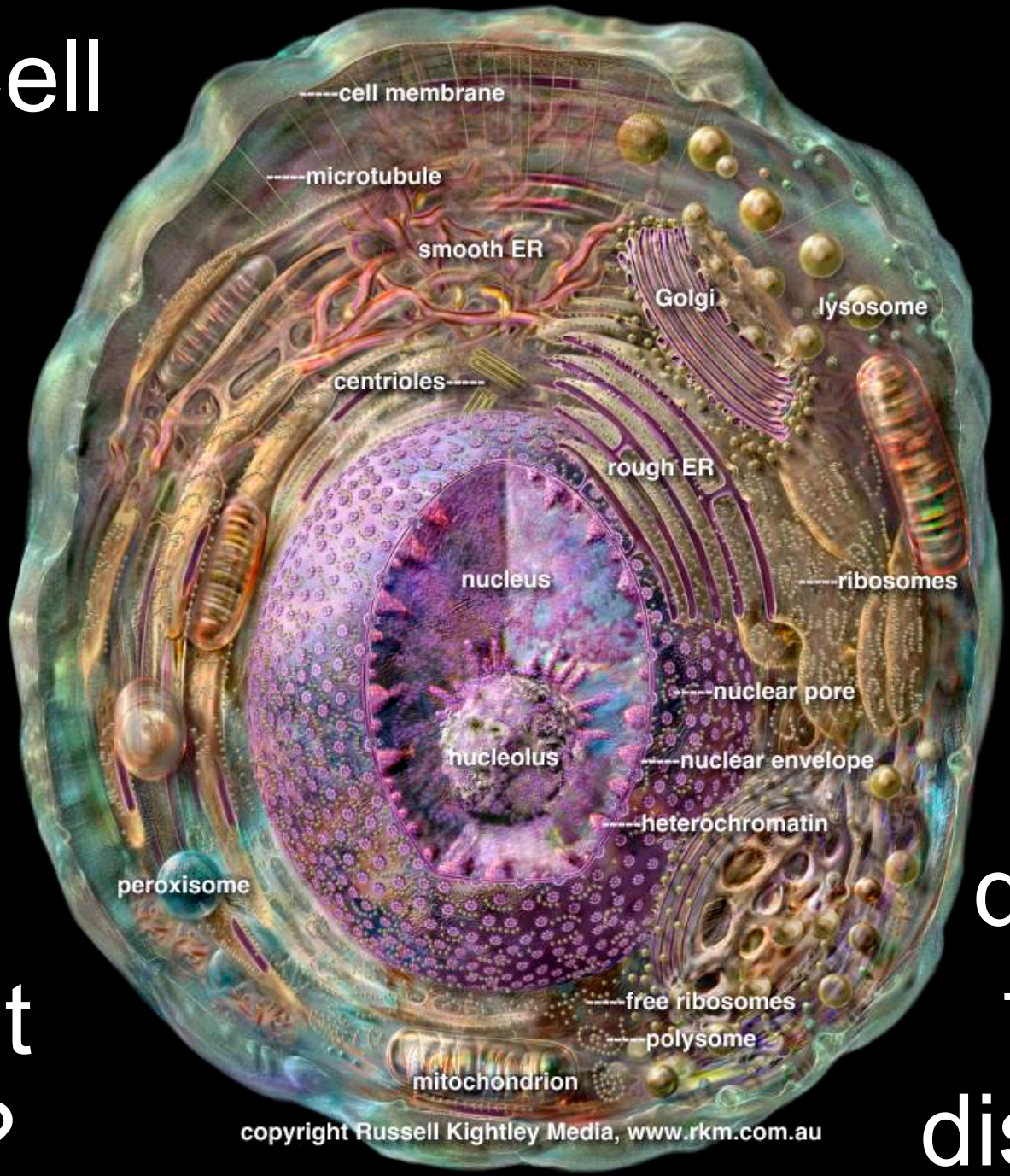
MoGen Topic Course - Mar.31.2008

http://baderlab.org

Outline

- Networks as scaffold for data integration
- Statistical integration
  - Gene function prediction
- Database and semantic integration
  - Biological pathways
- Visualizing integrated data with Cytoscape
- Example: Breast cancer classification

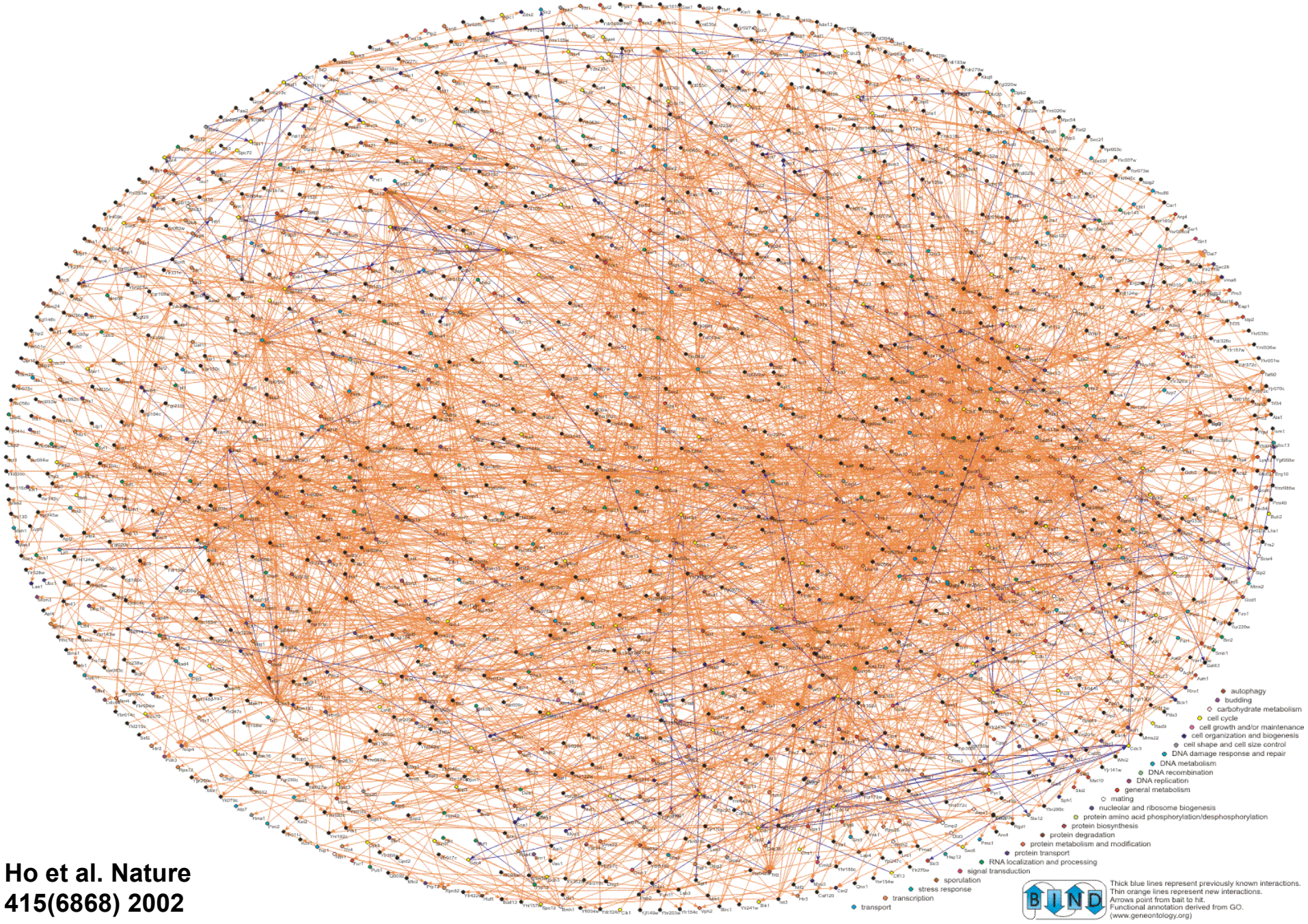
The Cell



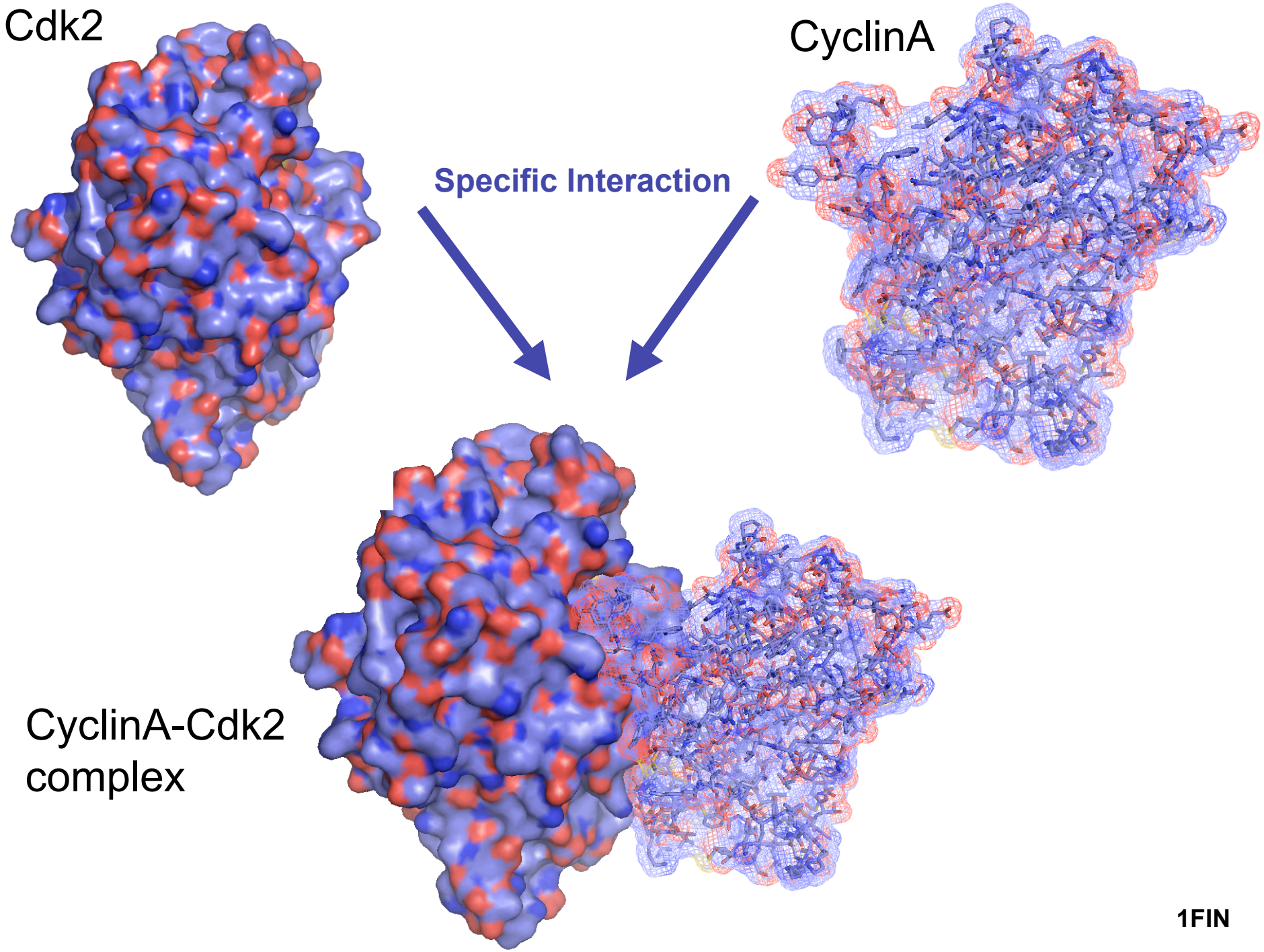
How does it work?

How does it fail in disease?

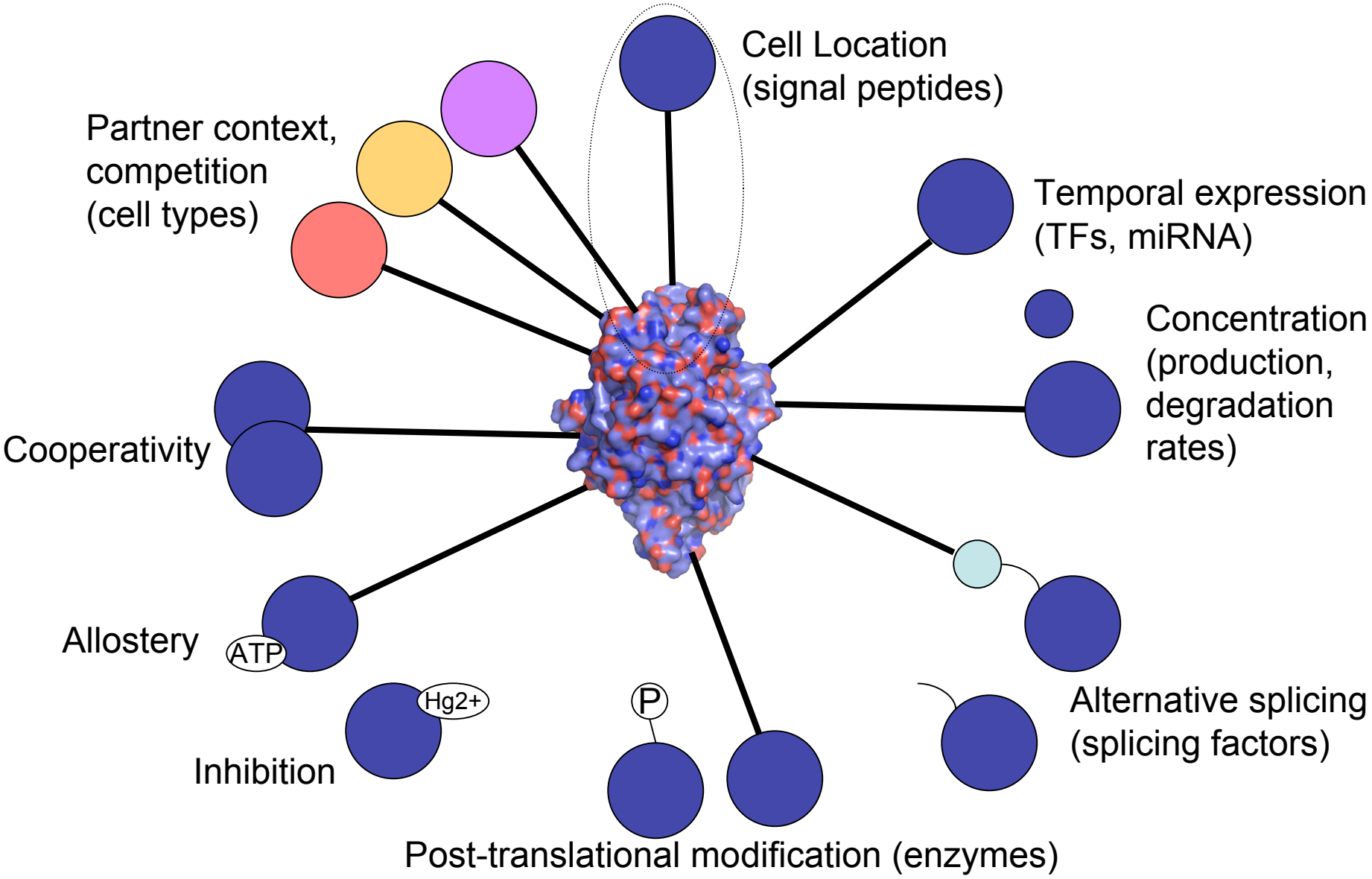
Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry

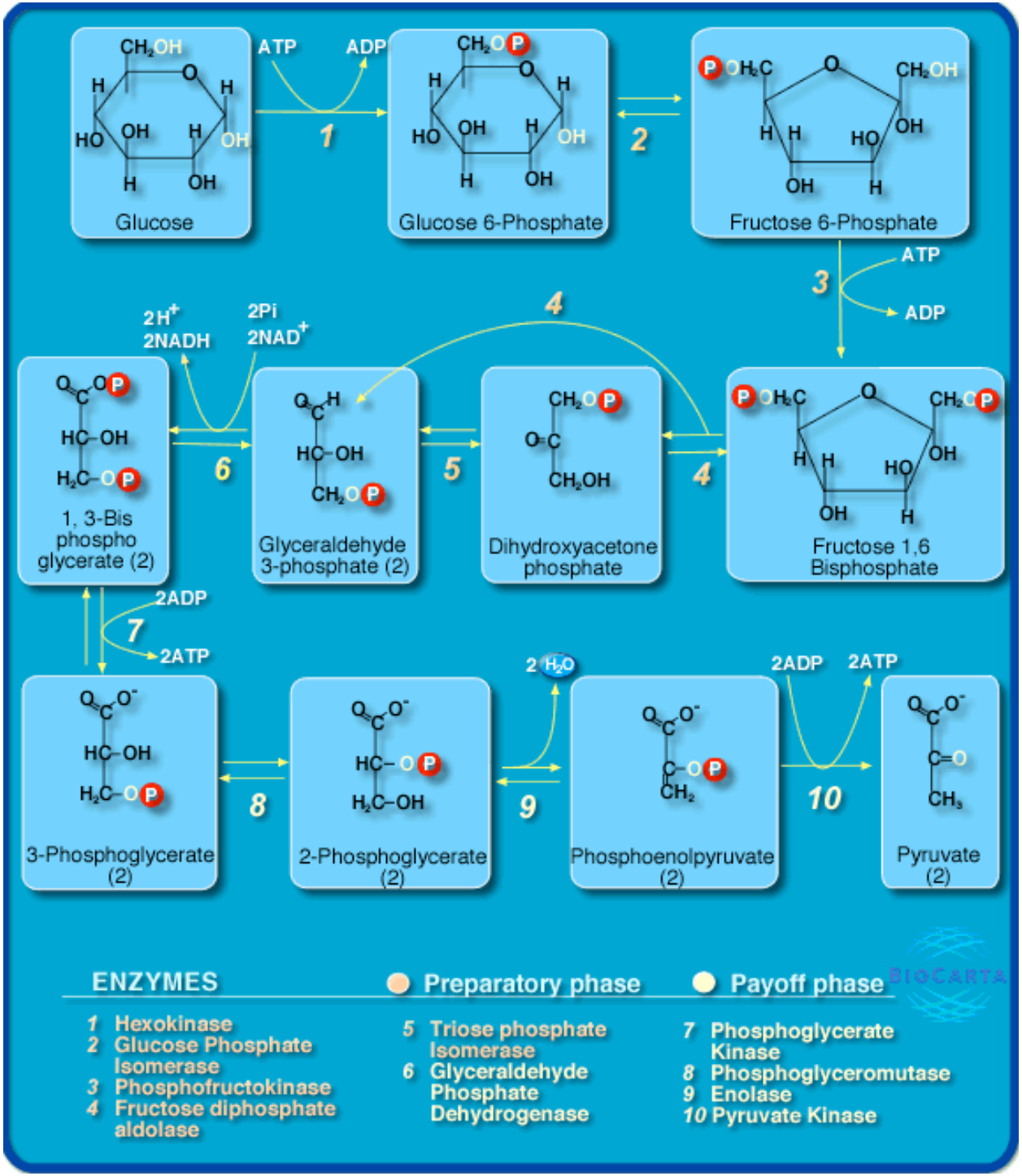


Ho et al. Nature
415(6868) 2002


Cdk2
CyclinA
Specific Interaction
CyclinA-Cdk2
complex
1FIN

Towards More Biologically Relevant Networks




Glucose
ATP
ADP
1
Glucose 6-Phosphate
2
Fructose 6-Phosphate
3
ATP
ADP
4
2H+
2NADH
2Pi
2NAD+
6
Glyceraldehyde 3-phosphate (2)
5
Dihydroxyacetone phosphate
4
Fructose 1,6 Bisphosphate
1, 3-Bis phospho glycerate (2)
2ADP
7
2ATP
3-Phosphoglycerate (2)
8
2-Phosphoglycerate (2)
2 H2O
9
Phosphoenolpyruvate (2)
2ADP
2ATP
10
Pyruvate (2)
ENZYMES
Preparatory phase
Payoff phase
1 Hexokinase
2 Glucose Phosphate Isomerase
3 Phosphofructokinase
4 Fructose diphosphate aldolase
5 Triose phosphate Isomerase
6 Glyceraldehyde Phosphate Dehydrogenase
7 Phosphoglycerate Kinase
8 Phosphoglyceromutase
9 Enolase
10 Pyruvate Kinase

Signaling Pathway

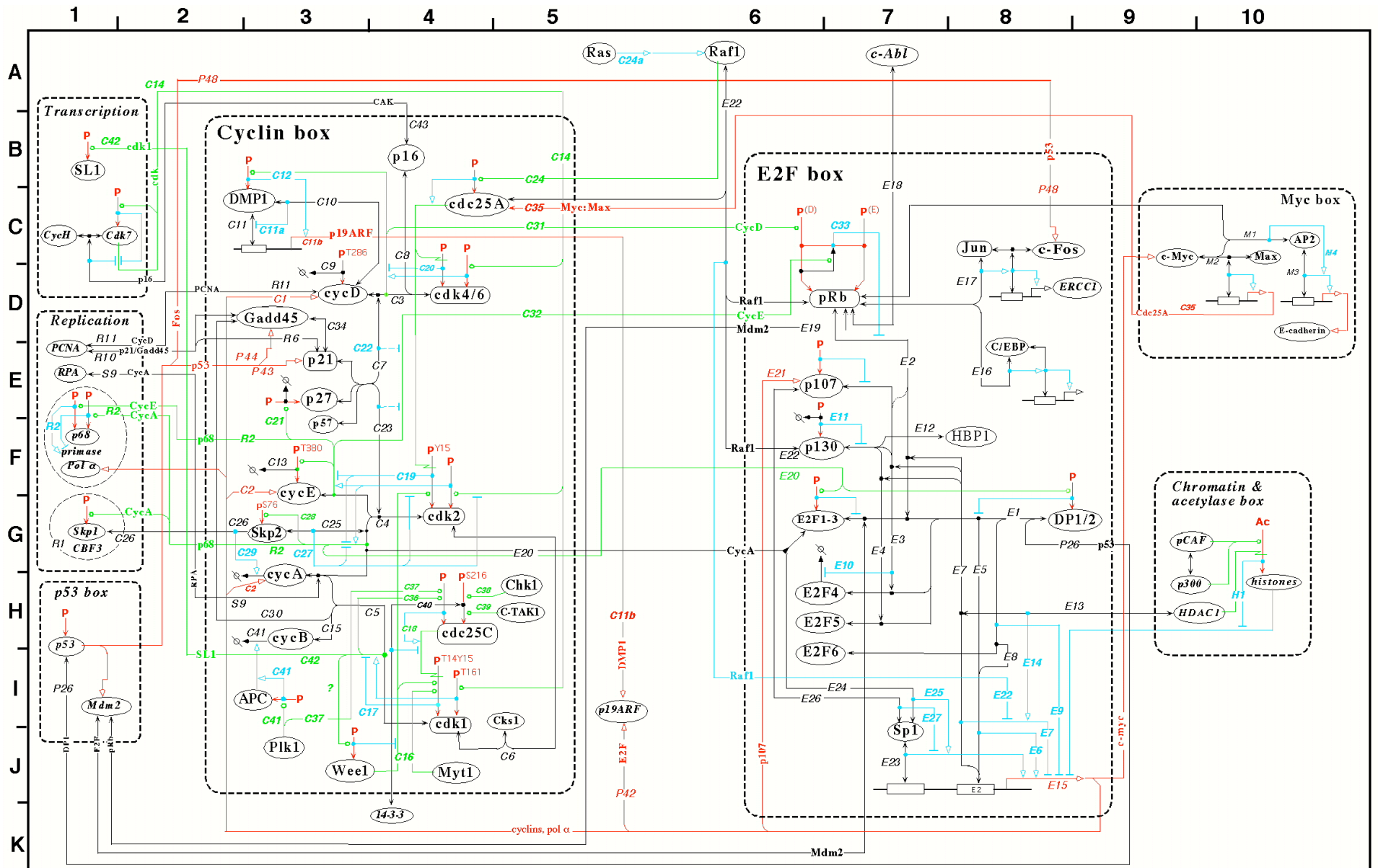
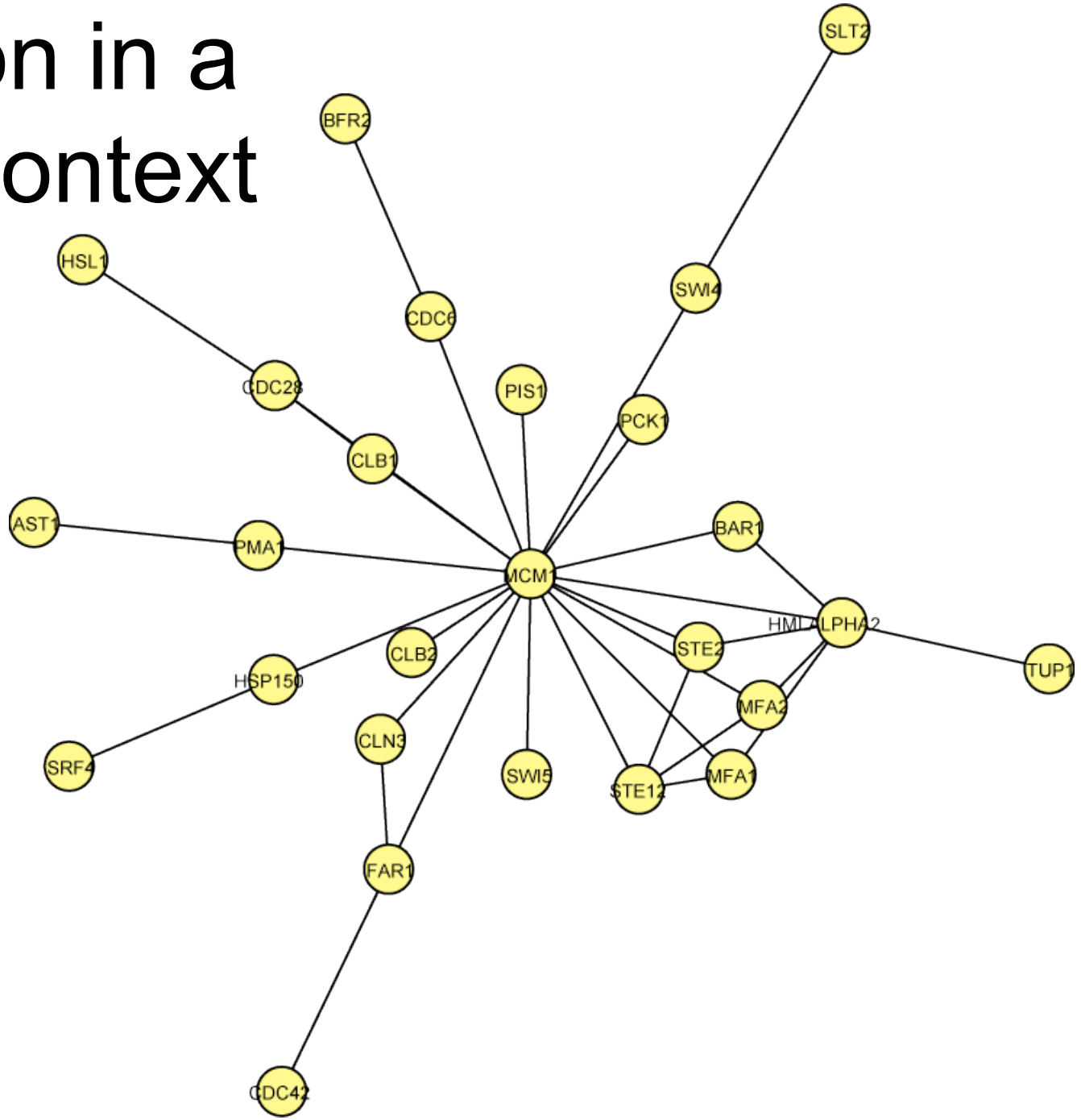


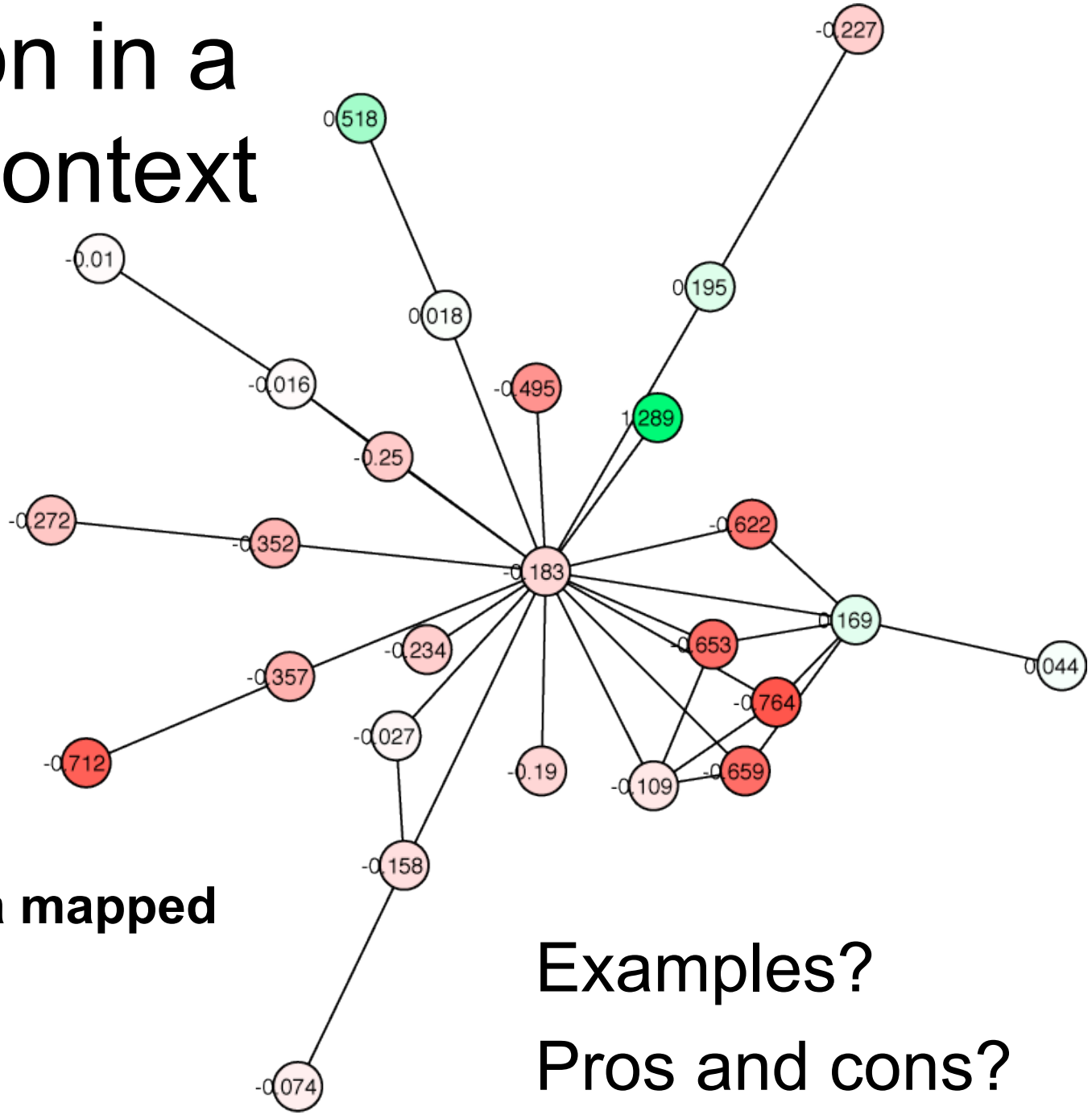
Figure 6A: The Cyclin - E2F cell cycle control system (version 3a - June 8, 1999)

http://discover.nci.nih.gov/kohnk/interaction_maps.html

Integration in a network context



Integration in a network context



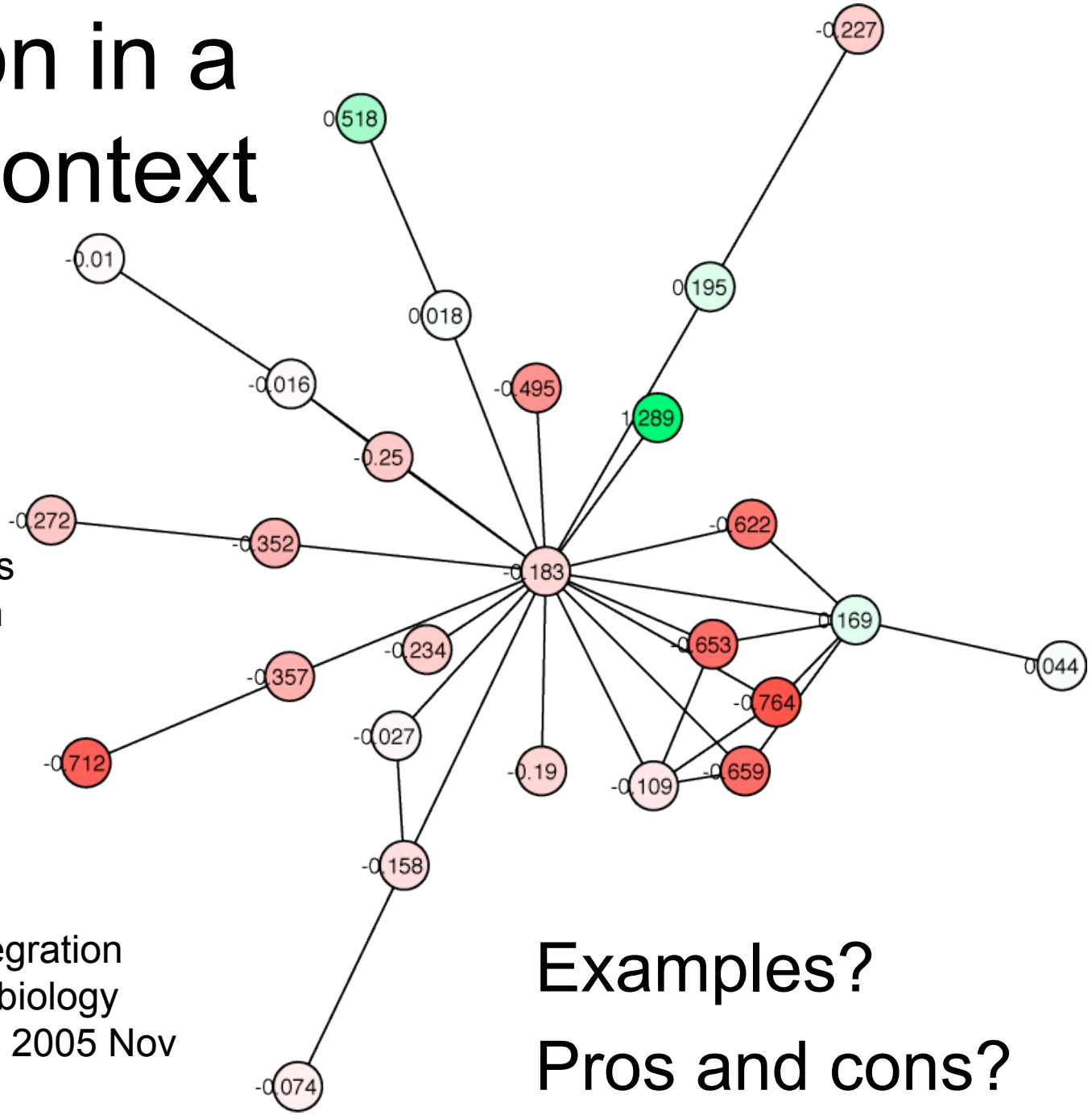
Expression data mapped to node colours

Examples?

Pros and cons?

Integration in a network context

Advantages:
-Broader coverage
-Error reduction
Challenges:
-Must carefully match data sets to avoid errors e.g. different interaction experiments
-Consider data set bias
-Consider binary vs. discrete vs. continuous



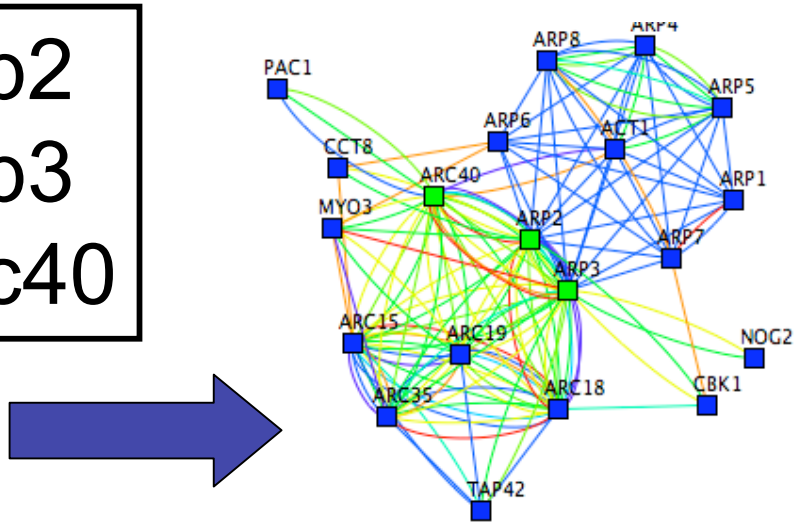
Hwang D et al. A data integration methodology for systems biology Proc Natl Acad Sci U S A. 2005 Nov 29;102(48):17296-301

Examples?

Pros and cons?

Predicting Gene Function

arp2
arp3
arc40



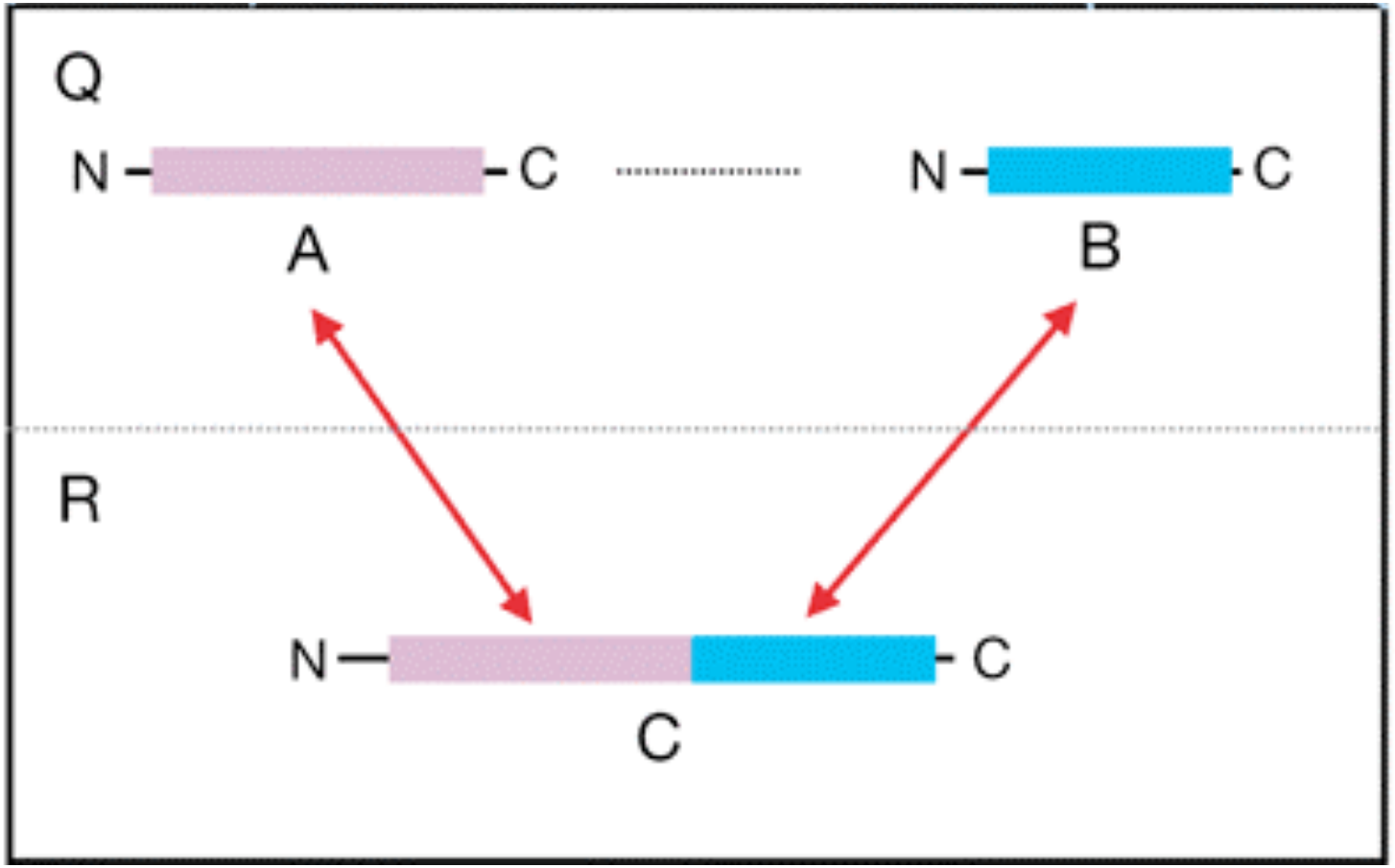
- STRING
  - http://string.embl.de/
- bioPIXIE
  - http://pixie.princeton.edu/pixie/
- GeneMania
  - http://genemania.org

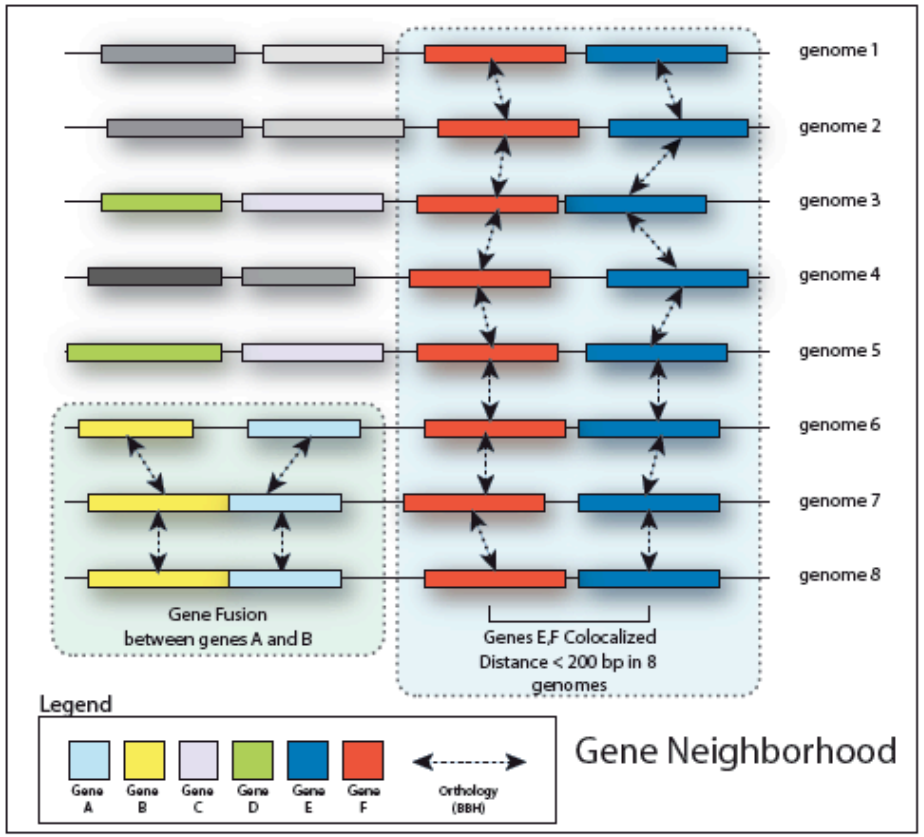
Top-Scoring Genes

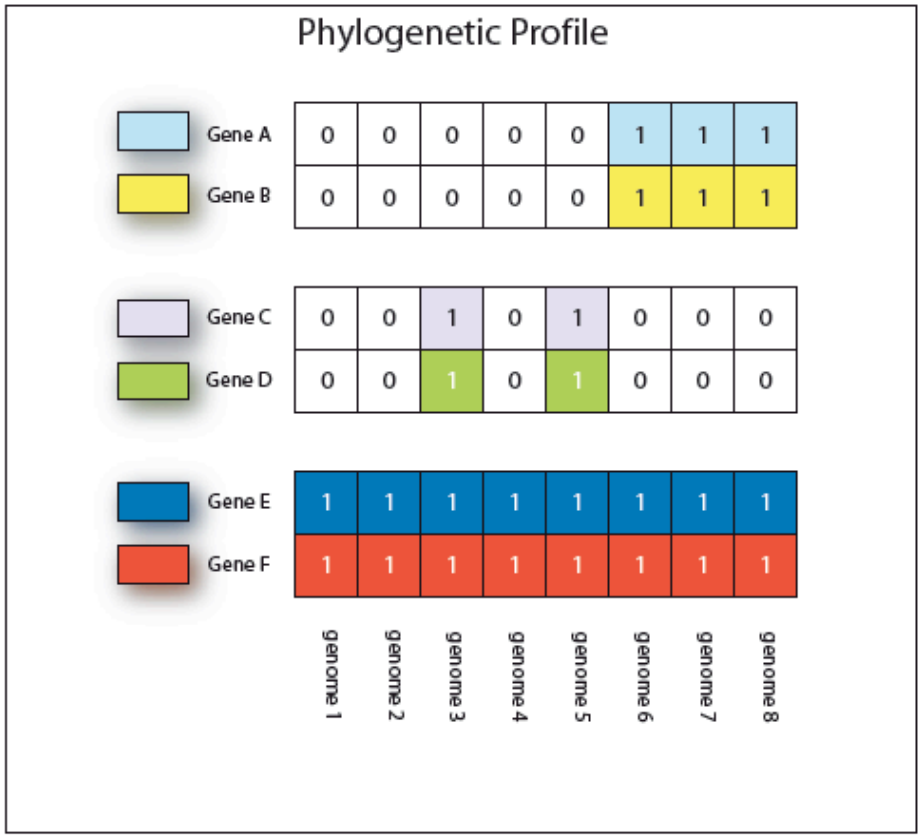
ARC15	0.09026
ARC19	0.08677
ARC35	0.08414
ARC18	0.07793
ARC40	0.03239
ARP8	0.02344
ARP5	0.02293
ARP6	0.02031
TAP42	0.02017
ACT1	0.01854
ARP4	0.01841
ARP1	0.01752
NOG2	0.01676
PAC1	0.01563
ARP7	0.01561
MYO3	0.01551

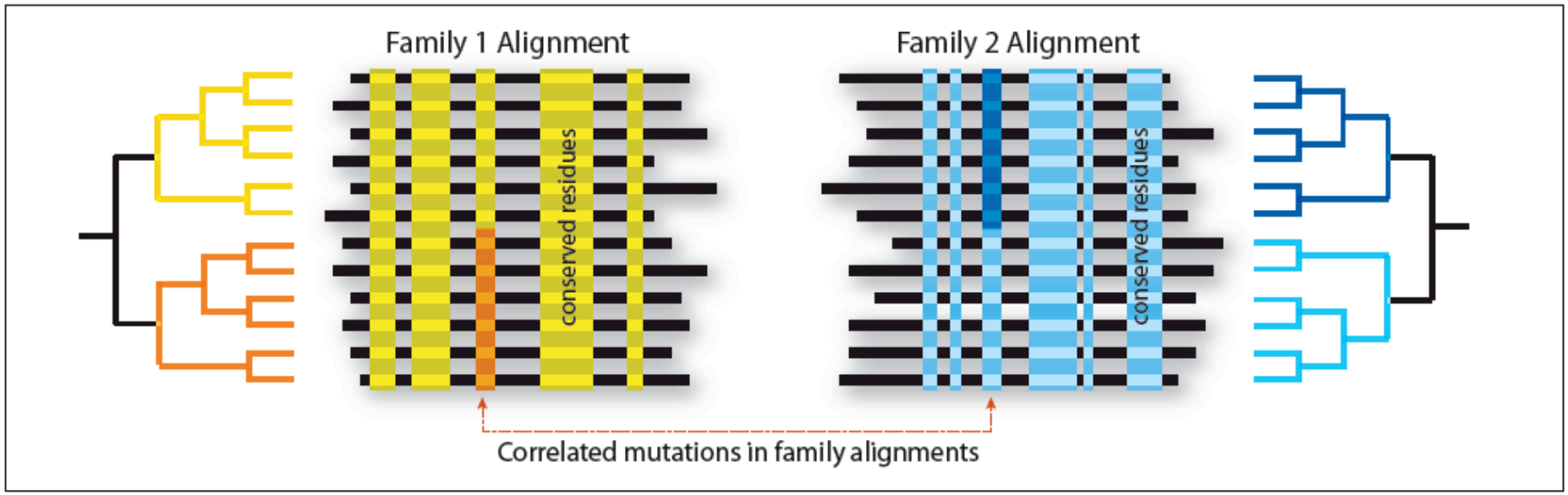
Fraser AG, Marcotte EM - A probabilistic view of gene function - Nat Genet. 2004 Jun;36(6):559-64

Gene Fusion




genome 1
genome 2
genome 3
genome 4
genome 5
genome 6
genome 7
genome 8
Gene Fusion
between genes A and B
Genes E,F Colocalized
Distance < 200 bp in 8
genomes
Legend
Gene A
Gene B
Gene C
Gene D
Gene E
Gene F
Orthology
(BBH)
Gene Neighborhood


Phylogenetic Profile
Gene A
Gene B
Gene C
Gene D
Gene E
Gene F
genome 1
genome 2
genome 3
genome 4
genome 5
genome 6
genome 7
genome 8


Family 1 Alignment
Family 2 Alignment
conserved residues
conserved residues
Correlated mutations in family alignments

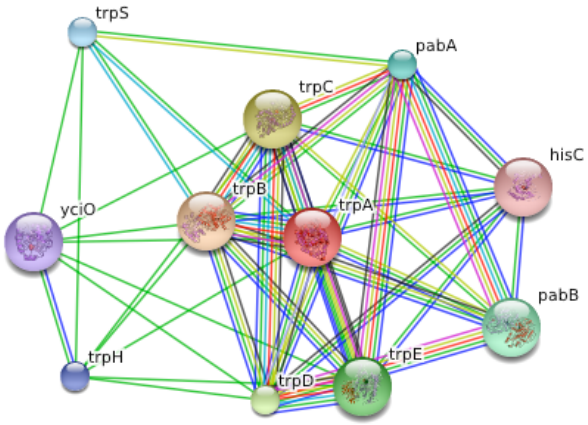
STRING

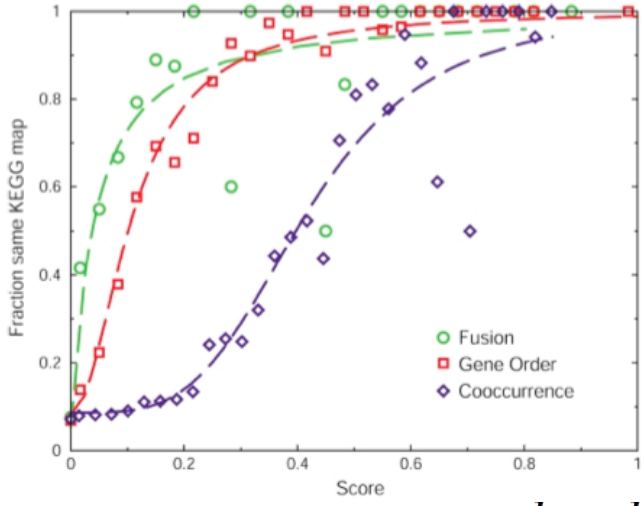
Your Input:

trpA Tryptophan synthase alpha chain (EC 4.2.1.20) (268 aa)
(Escherichia coli K12)

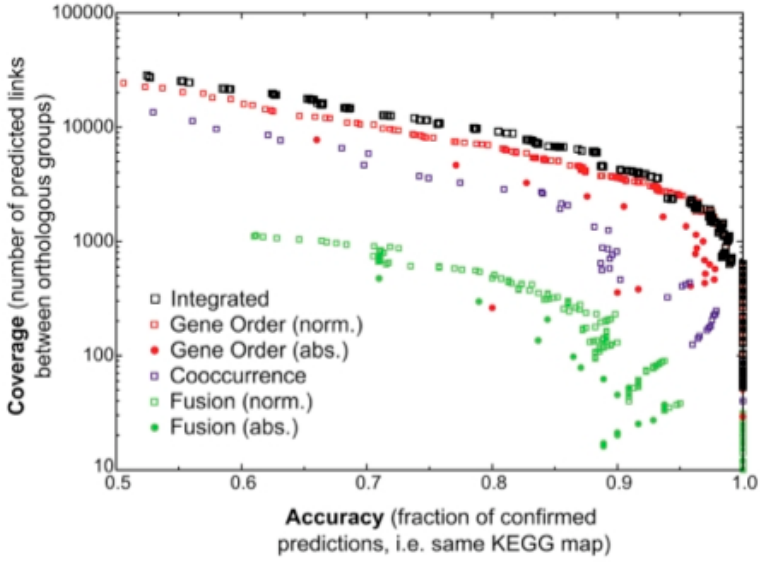
Predicted Functional Partners:

		Neighborhood	Gene Fusion	Cooccurrence	Coexpression	Experiments	Databases	Textmining	[Homology]	Score
trpB	Tryptophan synthase beta chain (EC 4.2.1.20) (397 aa)	●	●	●	●	●	●	●		0.999
trpC	Tryptophan biosynthesis protein trpCF [Includes- Indole-3-glycerol phosphate sy	●		●	●		●	●		0.999
trpD	Anthranilate synthase component II (EC 4.1.3.27) [Includes- Glutamine amidotr	●		●	●			●		0.999
trpE	Anthranilate synthase component 1 (EC 4.1.3.27) (Anthranilate synthase compo	●		●	●			●		0.999
pabB	Para-aminobenzoate synthase component 1 (EC 6.3.5.8) (Para- aminobenzoate	●						●		0.914
pabA	Para-aminobenzoate synthase glutamine amidotransferase component II (EC 6.	●						●		0.903
trpS	Tryptophanyl-tRNA synthetase (EC 6.1.1.2) (Tryptophan--tRNA ligase) (TrpRS) (	●					●			0.864
trpH	Protein trpH {UniProtKB/Swiss-Prot-P77766} (293 aa)	●								0.845
yciO	Protein yciO {UniProtKB/Swiss-Prot-P0AFR4} (206 aa)	●								0.841
hisC	Histidinol-phosphate aminotransferase (EC 2.6.1.9) (Imidazole acetol- phosphat	●		●						0.830





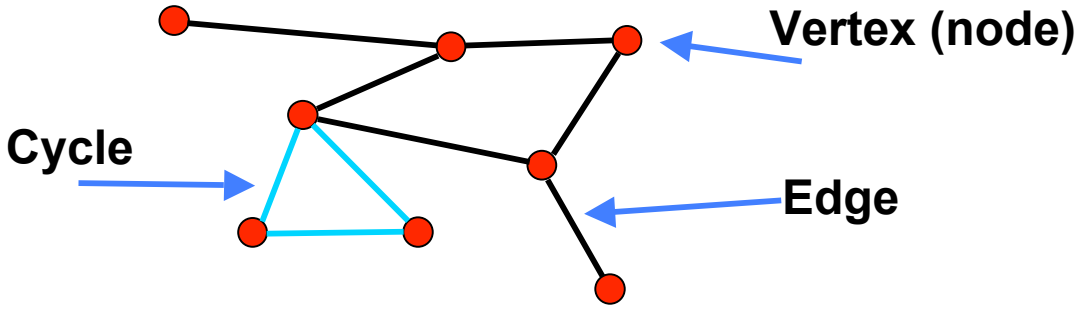
x=score
a=intercept
b=cooperativity
c=half max for x

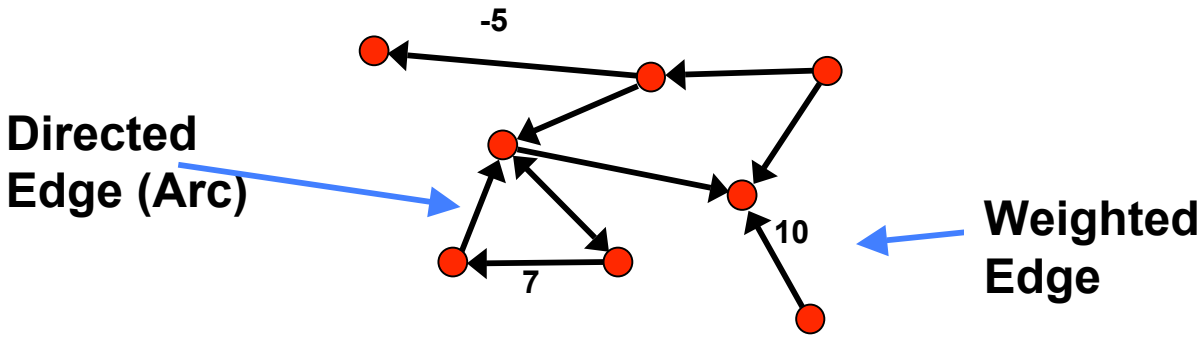


$$f(x) = a + [(1 - a)x^b/(c^b + x^b)]$$

von Mering C et al. Nucleic Acids Res. 2003 Jan 1;31(1):258-61

Graph Theory

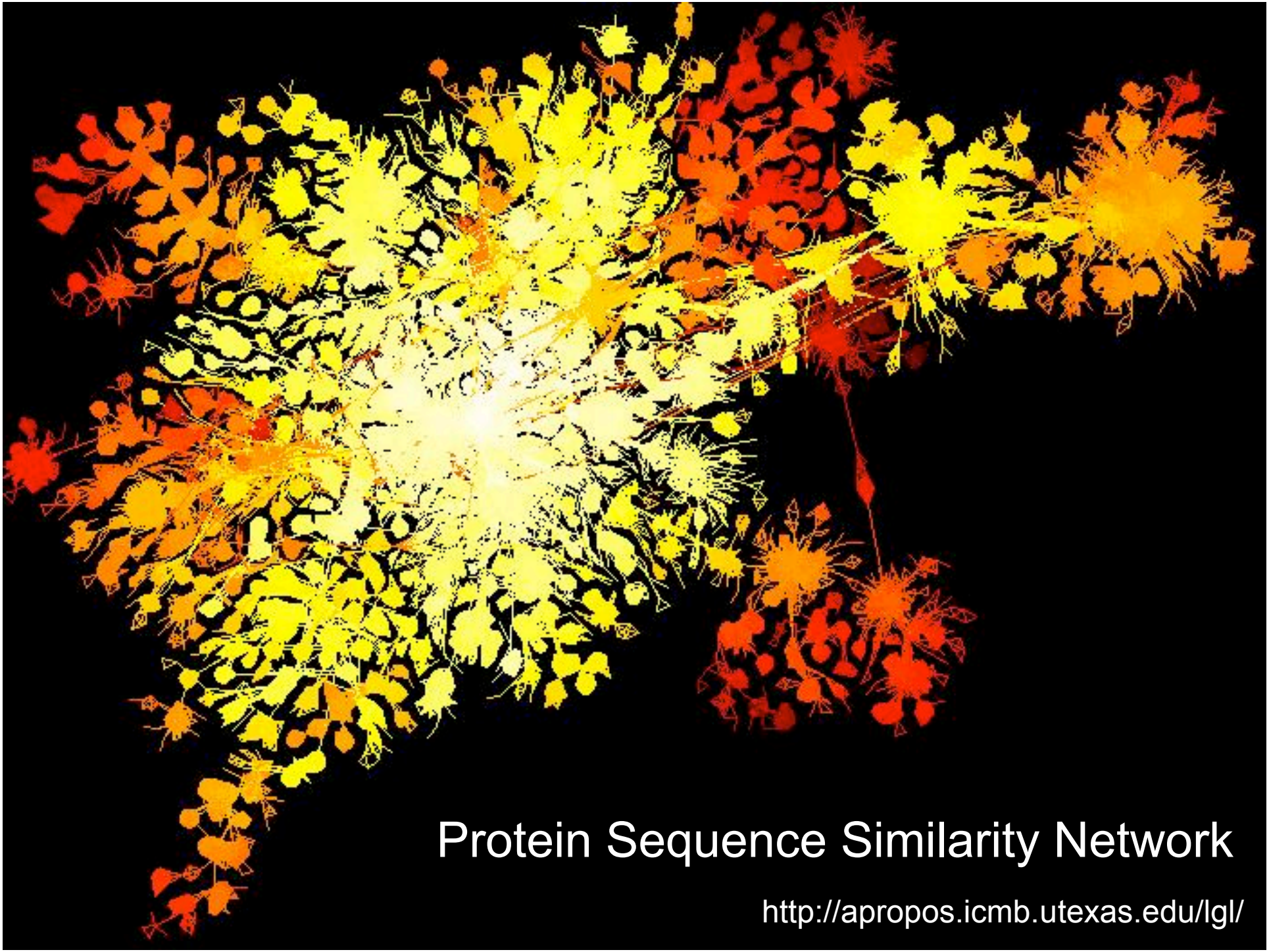




We map molecular interaction networks to graphs

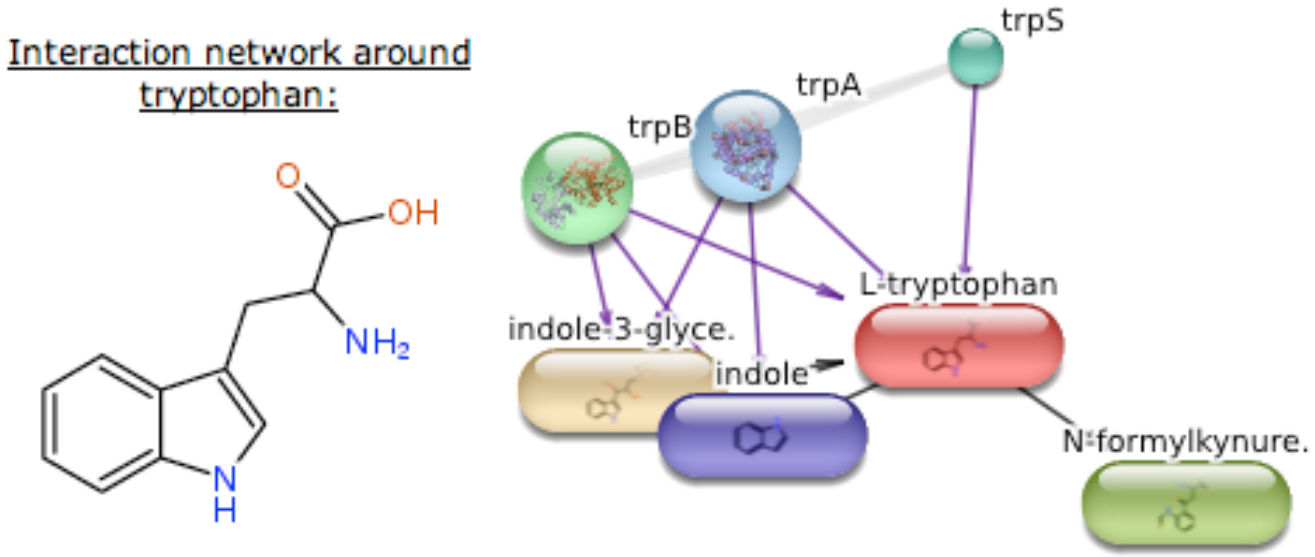
Mapping Biology to a Network

- A simple mapping
  - one compound/node, one interaction/edge
- A more realistic mapping
  - Cell localization, cell cycle, cell type, taxonomy
  - Only represent physiologically relevant interaction networks
- Edges can represent other relationships
- Critical: understand the mapping for network analysis

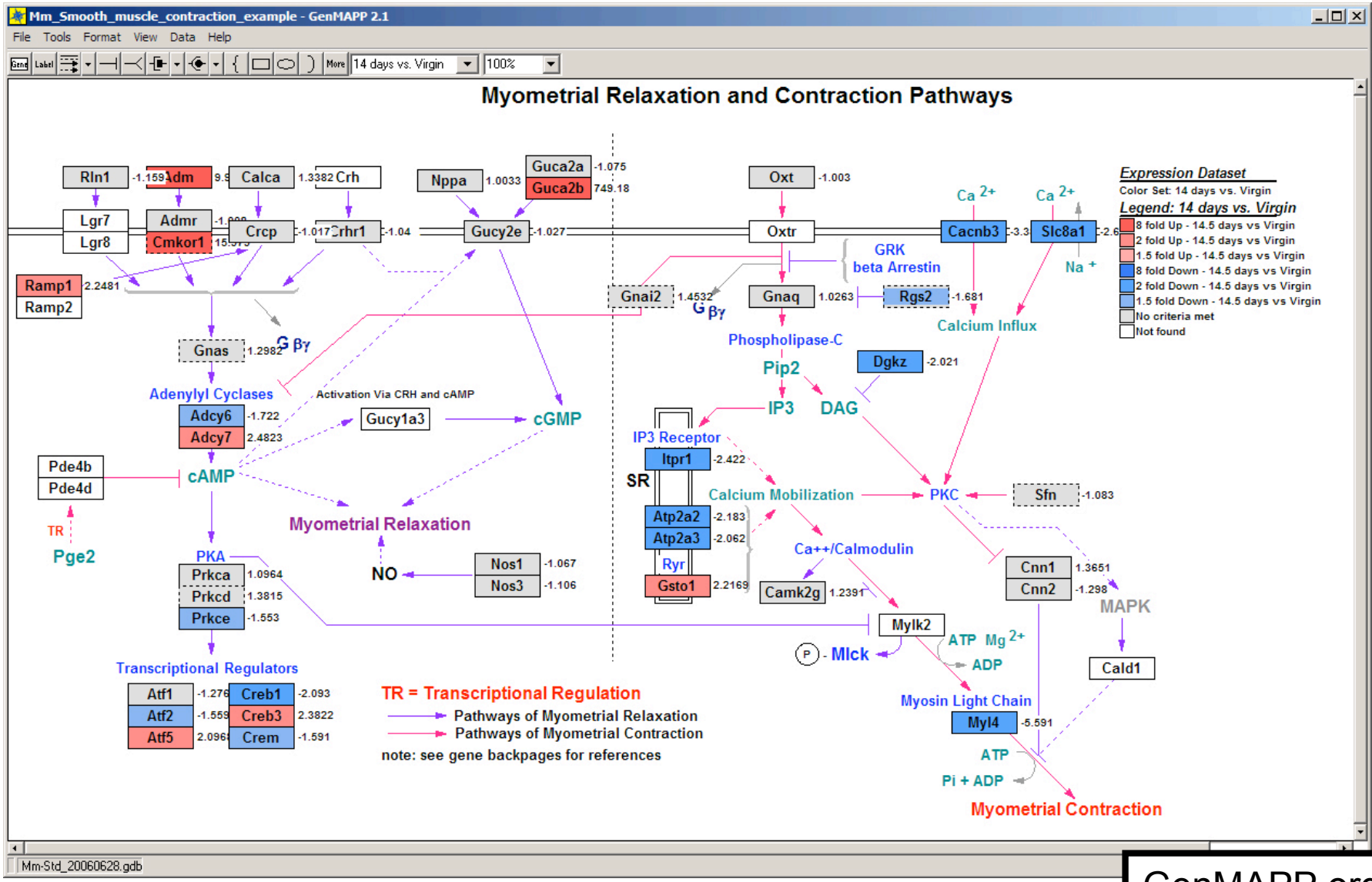

Protein Sequence Similarity Network
http://apropos.icmb.utexas.edu/lgl/

STITCH

- STITCH: Chemical-Protein Interactions
- http://stitch.embl.de/

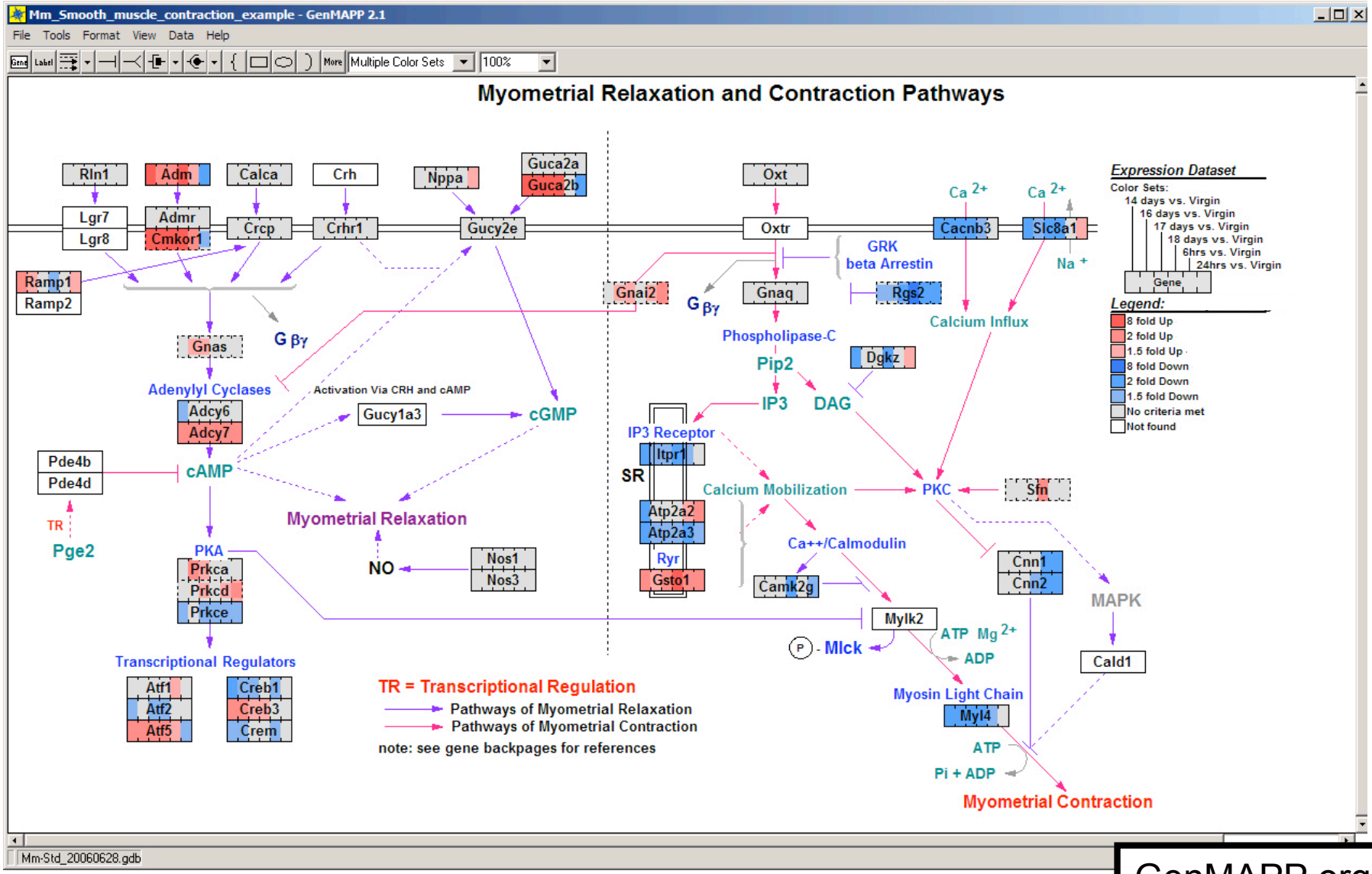


Visualizing Time Course Data on Pathways: Single Comparison View



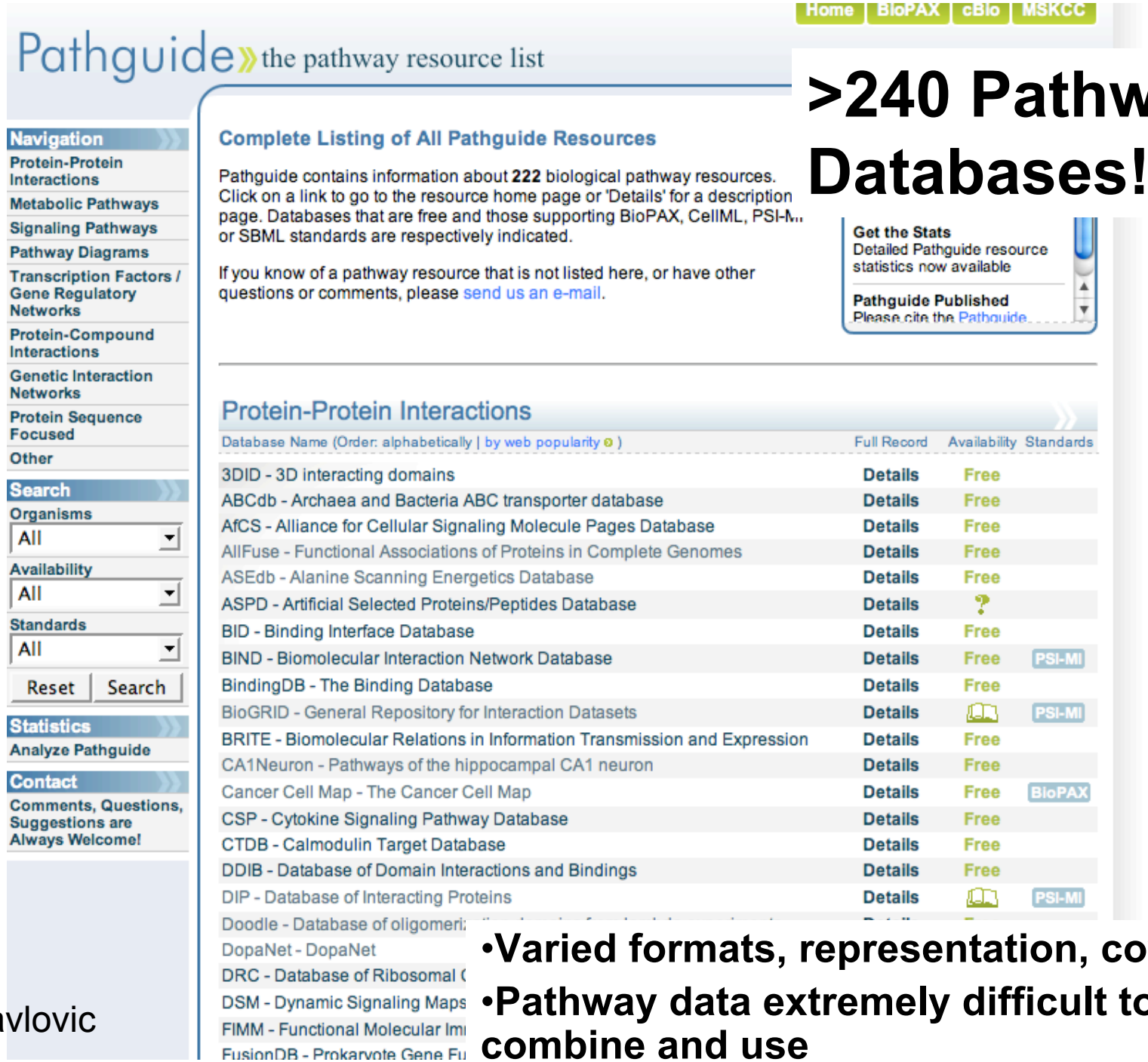
GenMAPP.org

Visualizing Time Course Data on Pathways: Multiple Comparison View



GenMAPP.org

http://pathguide.org



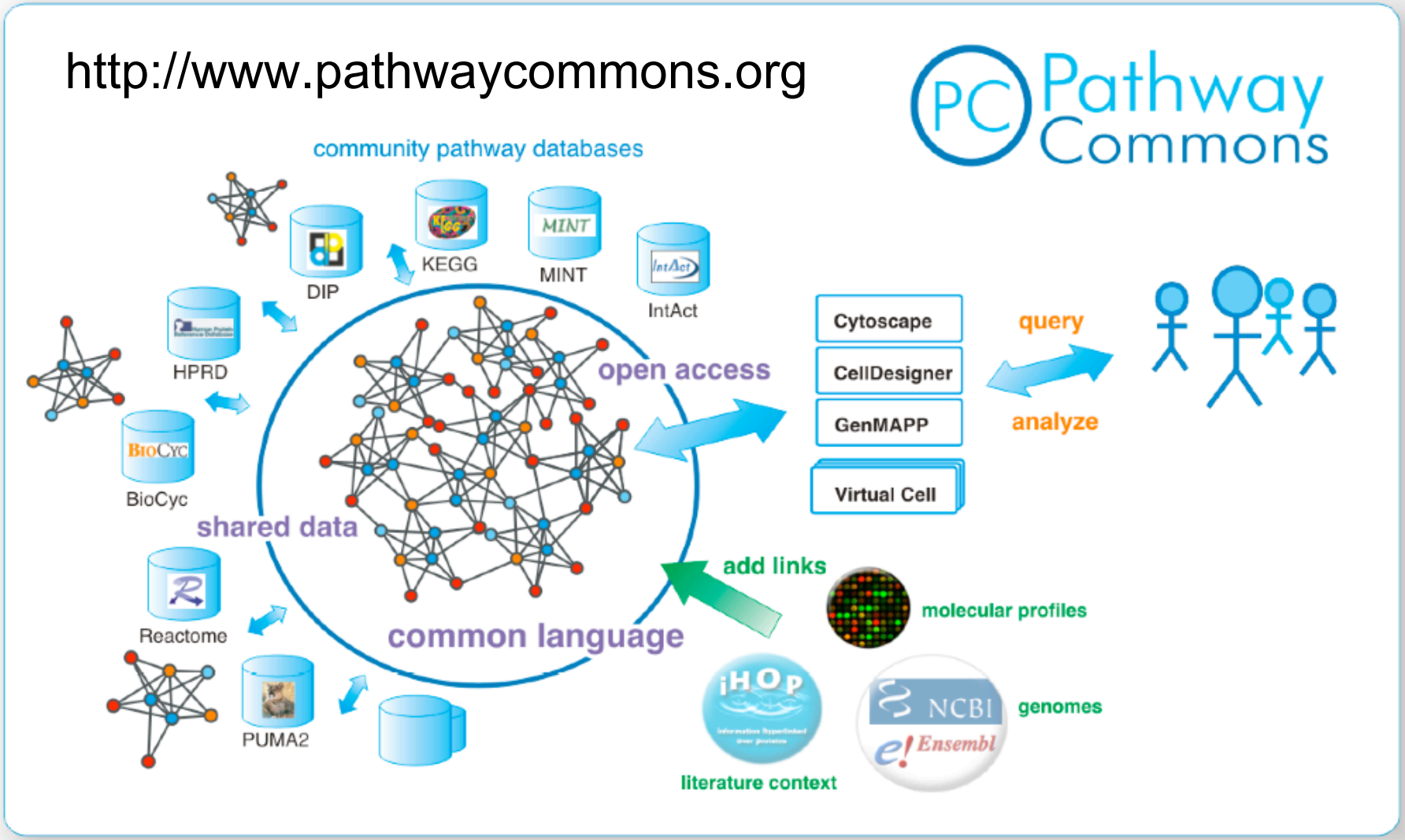
>240 Pathway Databases!

- **Varied formats, representation, coverage**
- **Pathway data extremely difficult to combine and use**

Vuk Pavlovic

Aim: Convenient Access to Pathway Information

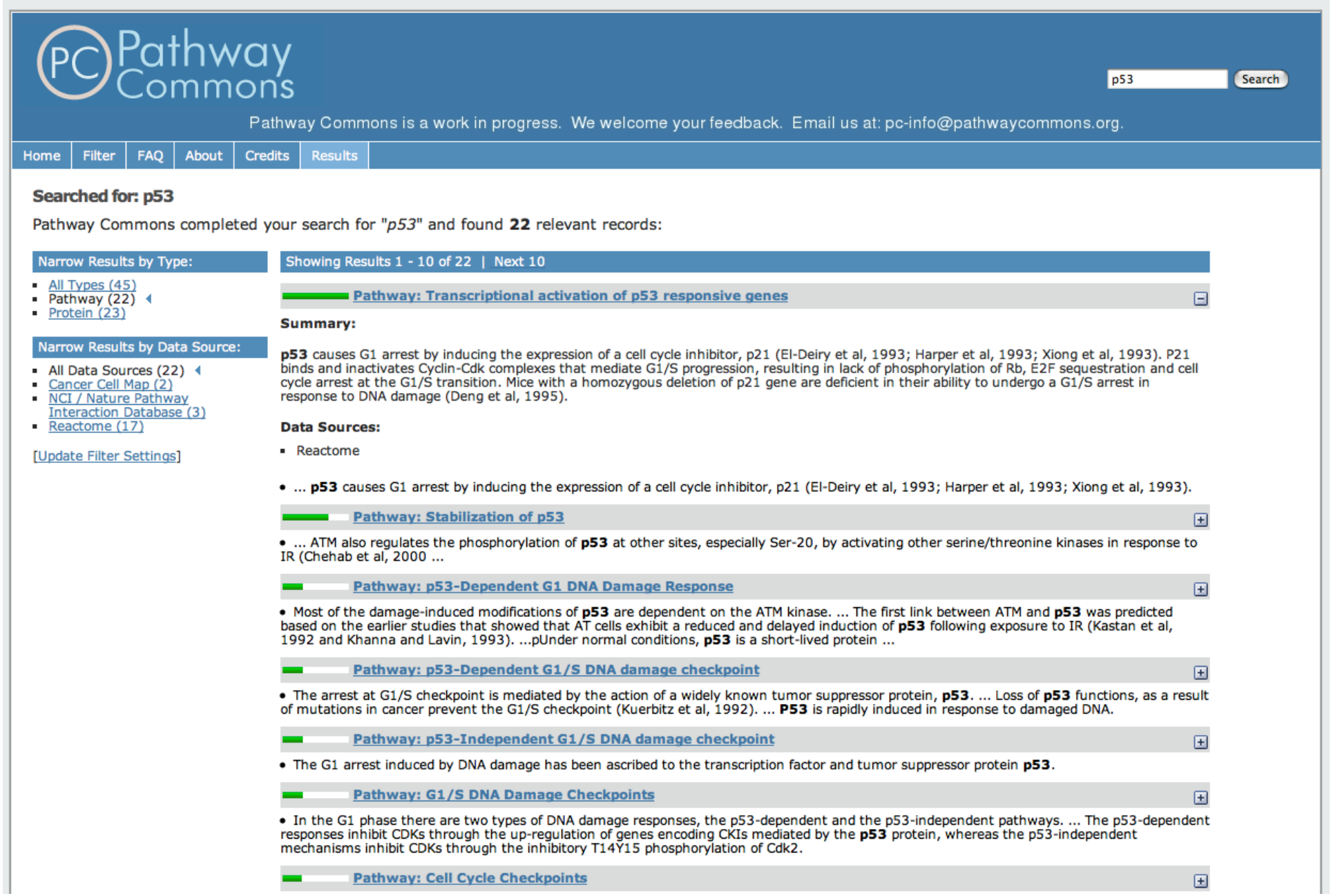
http://www.pathwaycommons.org



Facilitate creation and communication of pathway data
Aggregate pathway data in the public domain
Provide easy access for pathway analysis

Long term: Converge to integrated cell map

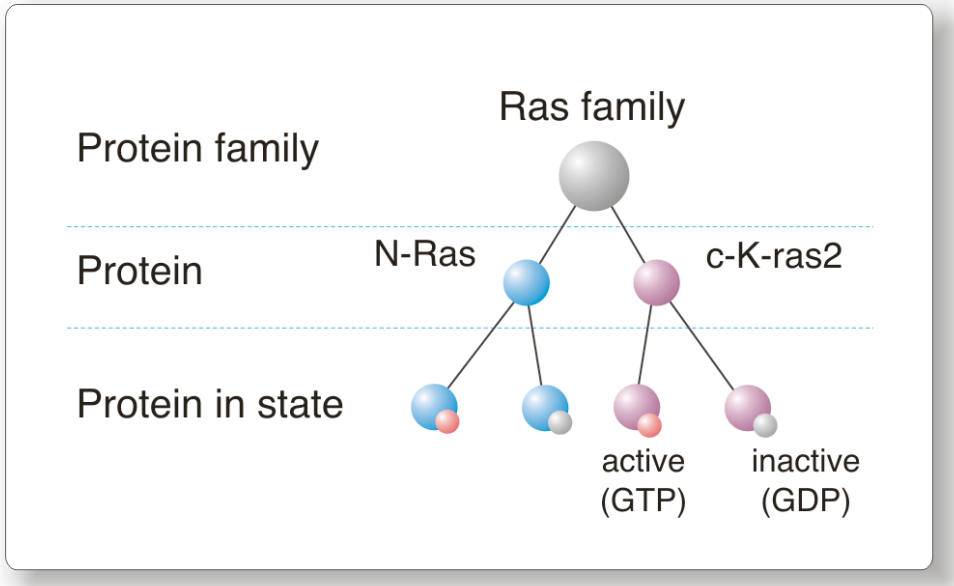
http://pathwaycommons.org



http://pathwaycommons.org

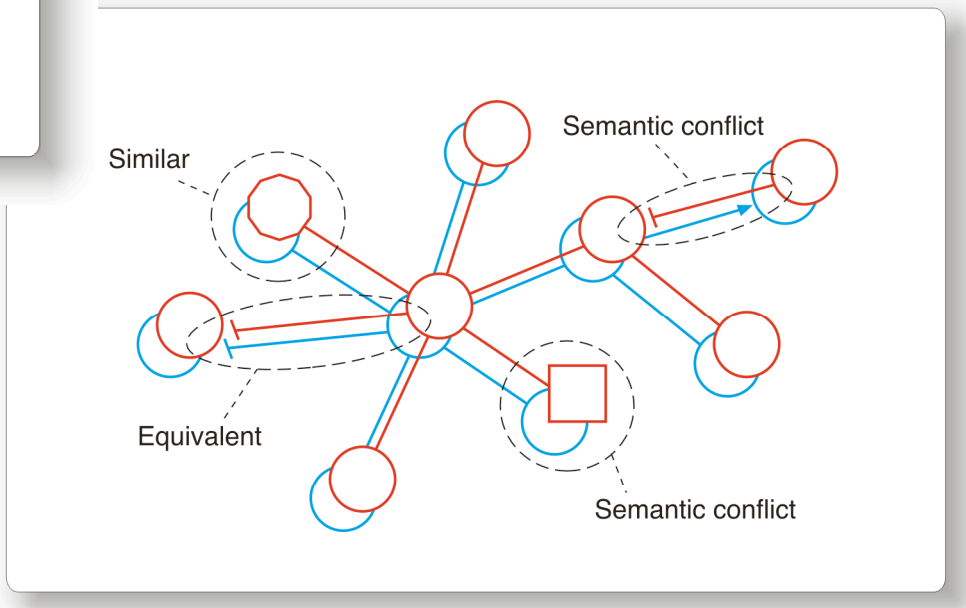
Towards an Integrated Cell Map

- Semantic pathway integration is difficult



Physical entities

Relationships

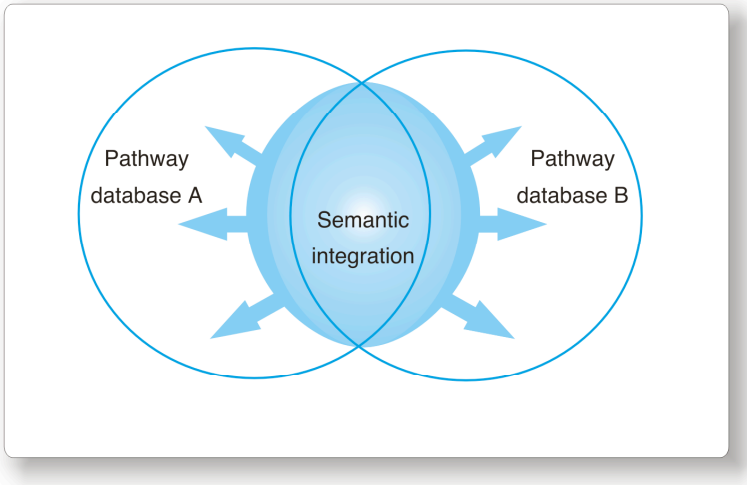


Determining equivalent entities is critical

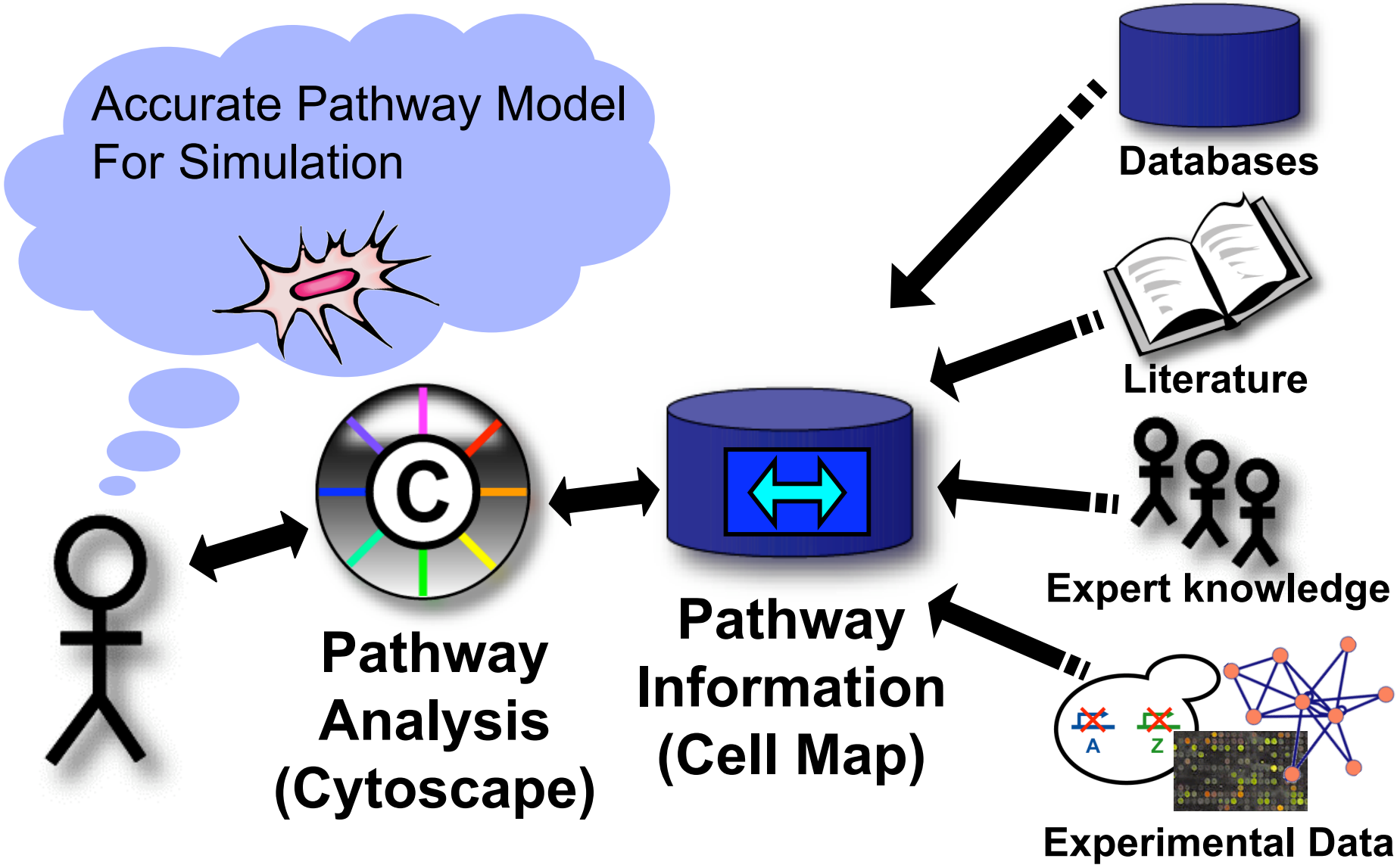
Practical Semantic Integration

- Minimize errors
  - Integrate only where possible with high accuracy
  - Detect and flag conflicts, errors for users, no revision
  - Promote best-practices to minimize future errors
  - Interaction confidence algorithms
  - Validation software
  - Allow users to filter and select trusted sources
- Converge to standard representation
  - Community process

Doable: hundreds of curators globally in >200 databases (GDP) - make it more efficient

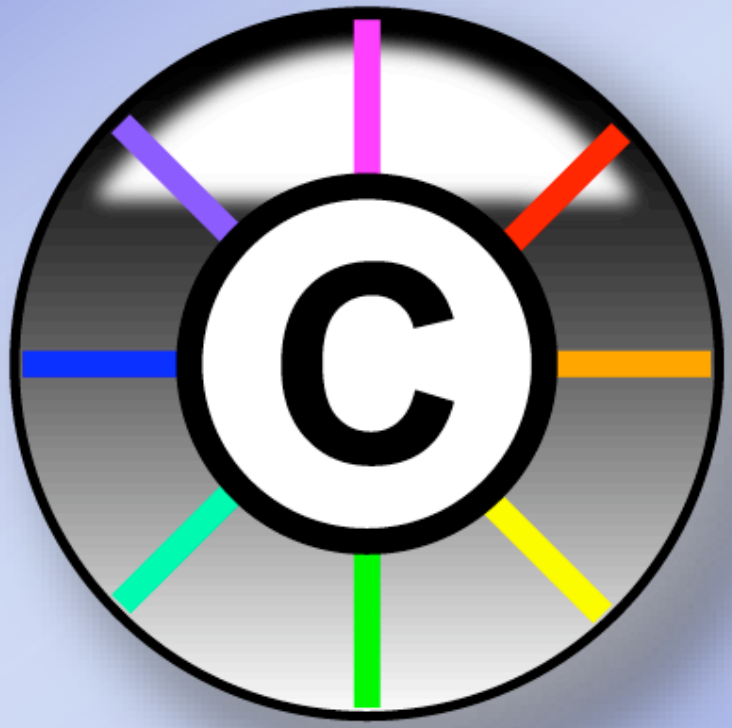


Using Pathway Information



Cytoscape

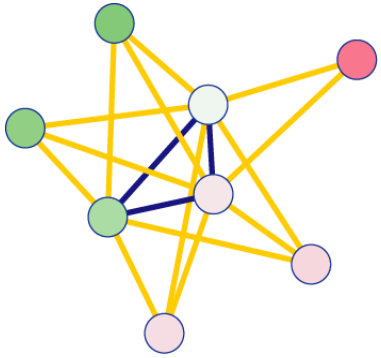
http://cytoscape.org



MEMORIAL SLOAN-KETTERING CANCER CENTER
ESTABLISHED 1884

Agilent Technologies

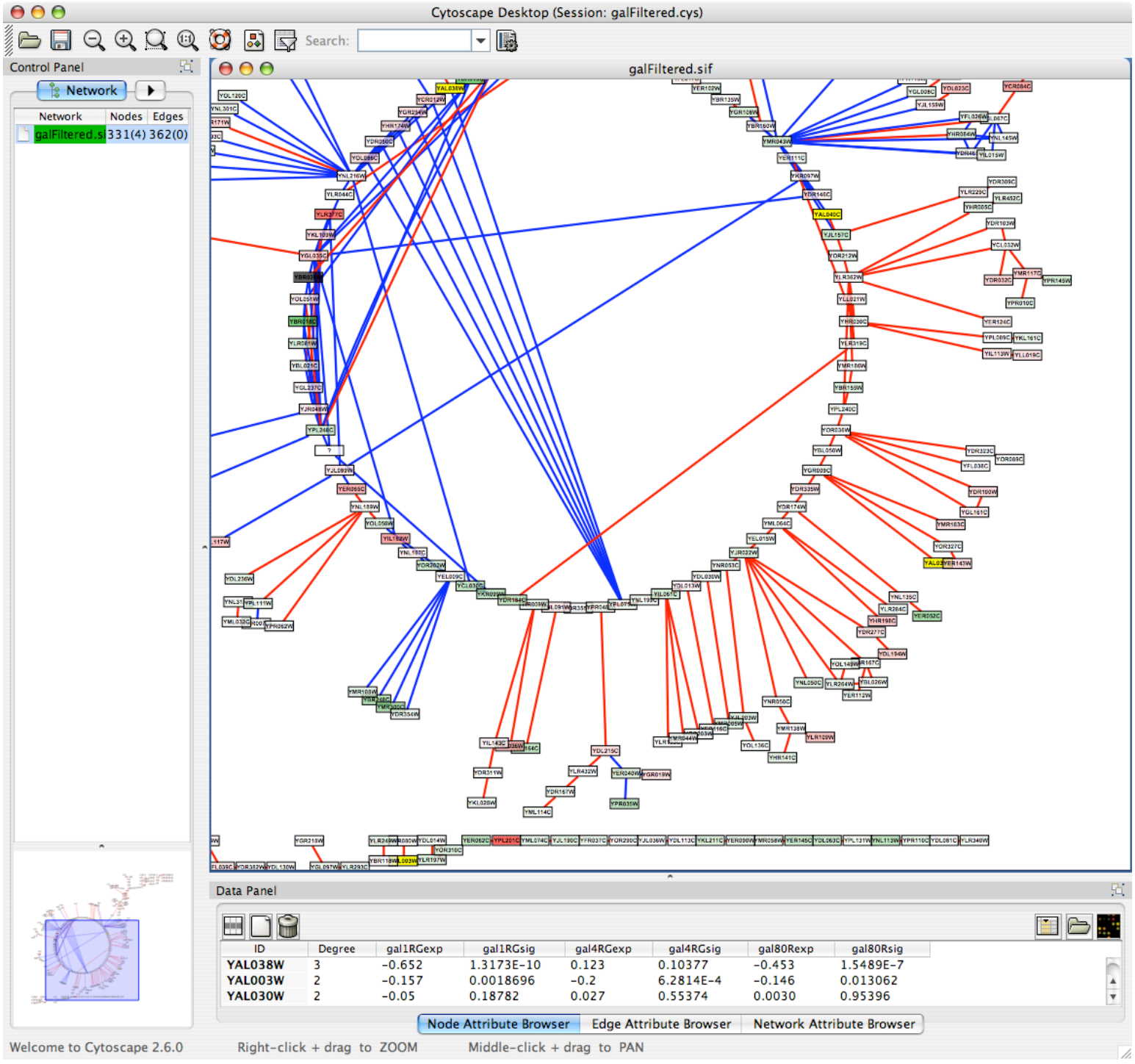
Cytoscape - Network Visualization and Analysis

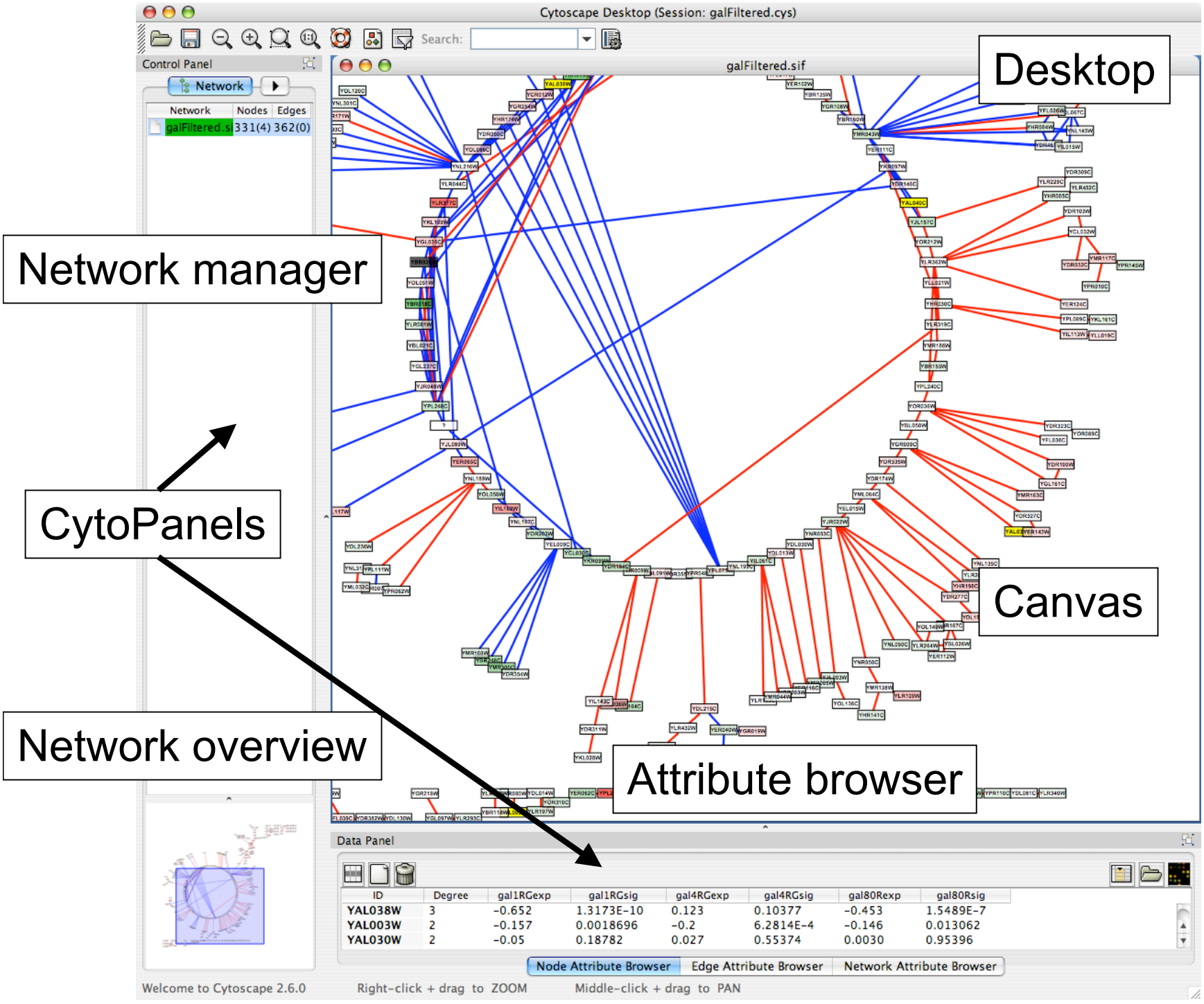


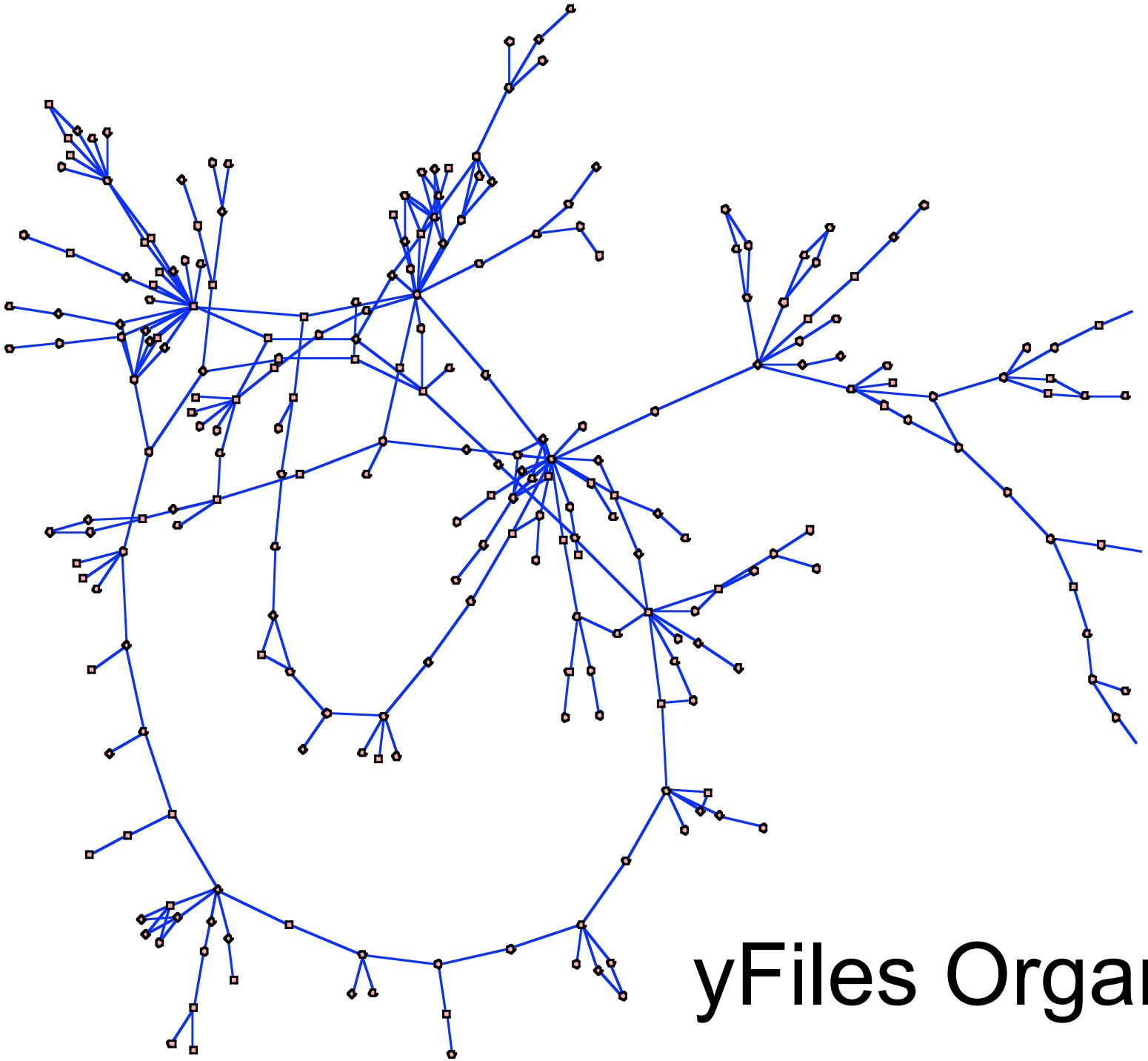
http://cytoscape.org

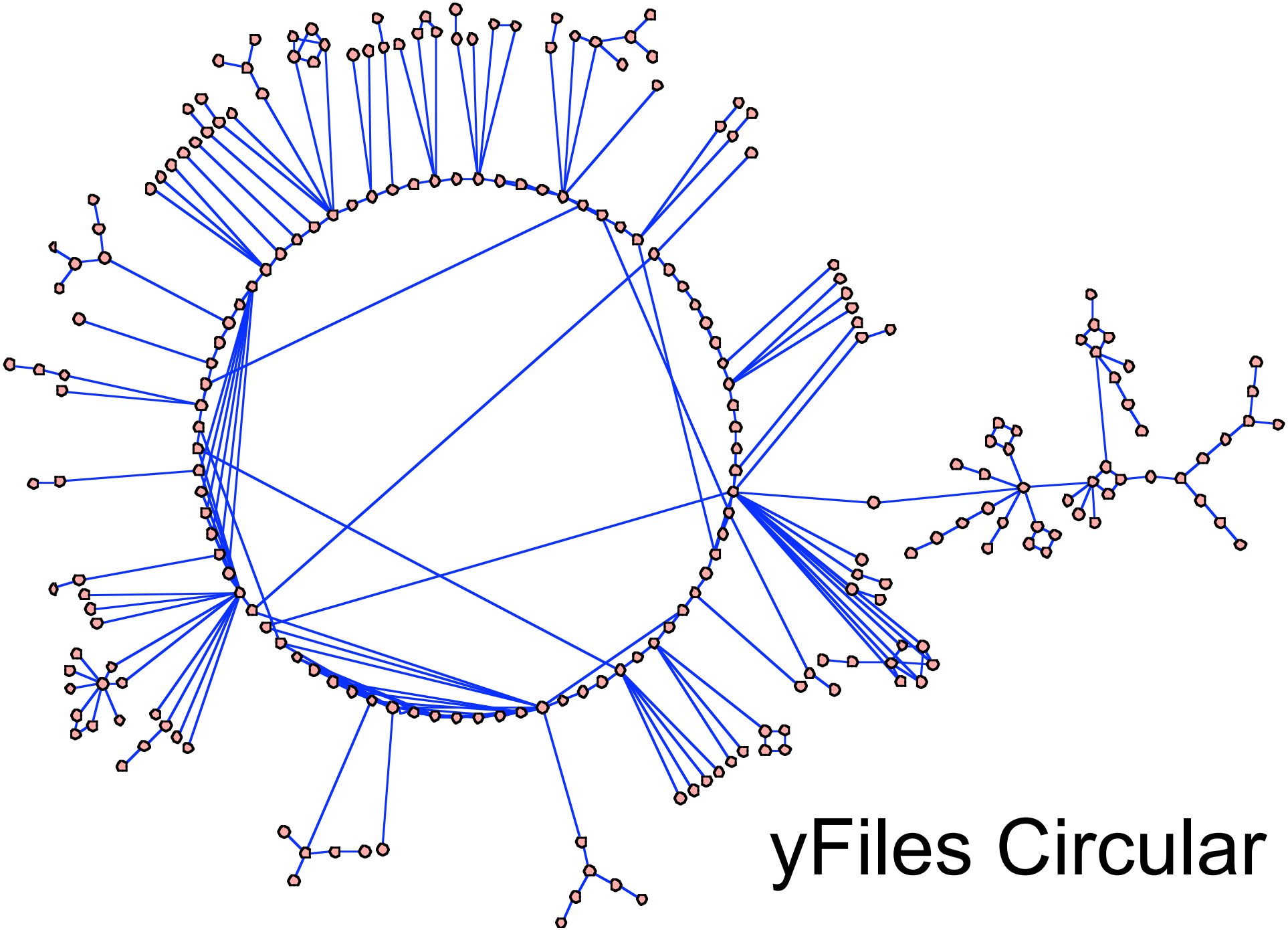
- Freely-available (open-source, java) software
- Visualizing biological networks (e.g. molecular interaction networks)
- Analyzing networks with gene expression profiles and other cell state data

Other software: Osprey, BioLayout, VisANT, Navigator, PIMWalker, ProViz


Cytoscape Desktop (Session: galFiltered.cys)
Search:
Control Panel
Network
Network
Nodes
Edges
galFiltered.si
331(4)
362(0)
galFiltered.sif
Data Panel
ID
Degree
gal1RGexp
gal1RGsig
gal4RGexp
gal4RGsig
gal80Rexp
gal80Rsig
YAL038W
3
-0.652
1.3173E-10
0.123
0.10377
-0.453
1.5489E-7
YAL003W
2
-0.157
0.0018696
-0.2
6.2814E-4
-0.146
0.013062
YAL030W
2
-0.05
0.18782
0.027
0.55374
0.0030
0.95396
Node Attribute Browser
Edge Attribute Browser
Network Attribute Browser
Welcome to Cytoscape 2.6.0
Right-click + drag to ZOOM
Middle-click + drag to PAN

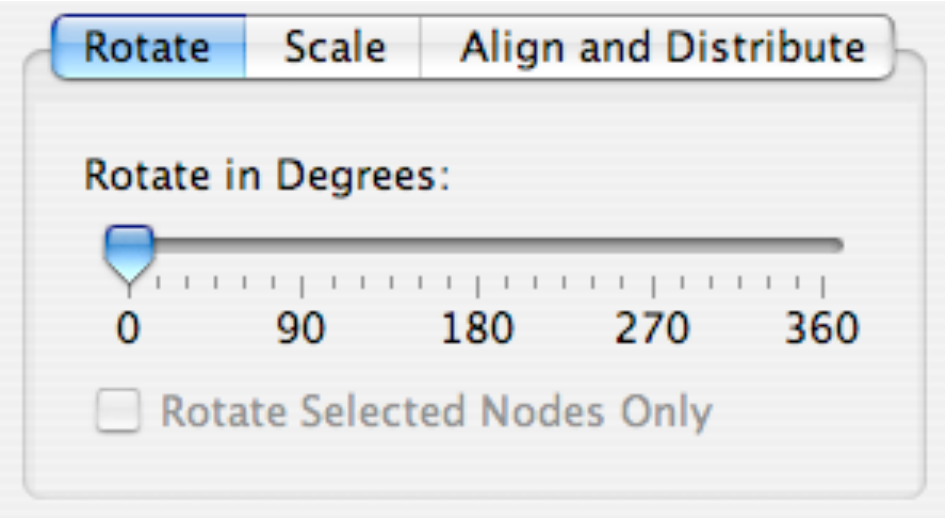

Cytoscape Desktop (Session: galFiltered.cys)
Search:
Control Panel
Network
Network
Nodes
Edges
galFiltered.si
331(4)
362(0)
galFiltered.sif
Desktop
Network manager
CytoPanels
Network overview
Canvas
Attribute browser
Data Panel
ID
Degree
gal1RGexp
gal1RGsig
gal4RGexp
gal4RGsig
gal80Rexp
gal80Rsig
YAL038W
3
-0.652
1.3173E-10
0.123
0.10377
-0.453
1.5489E-7
YAL003W
2
-0.157
0.0018696
-0.2
6.2814E-4
-0.146
0.013062
YAL030W
2
-0.05
0.18782
0.027
0.55374
0.0030
0.95396
Node Attribute Browser
Edge Attribute Browser
Network Attribute Browser
Welcome to Cytoscape 2.6.0
Right-click + drag to ZOOM
Middle-click + drag to PAN

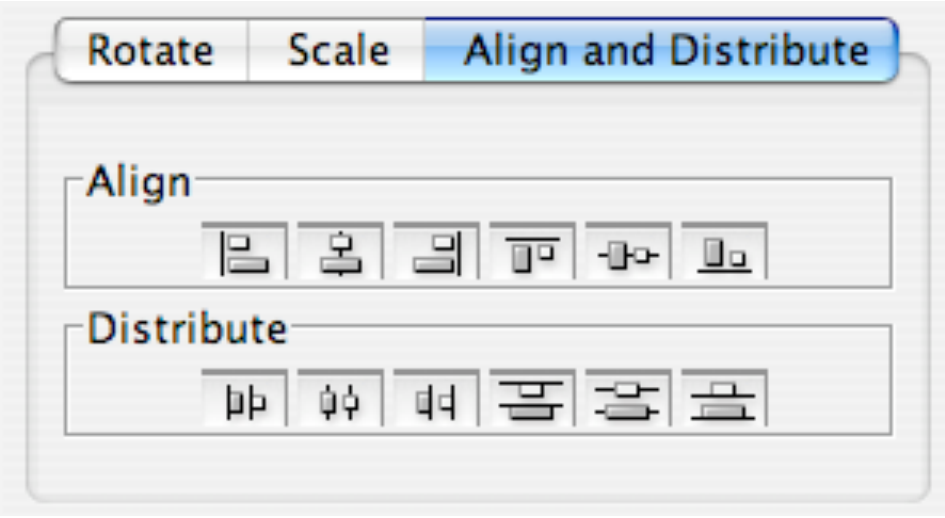

yFiles Organic


yFiles Circular

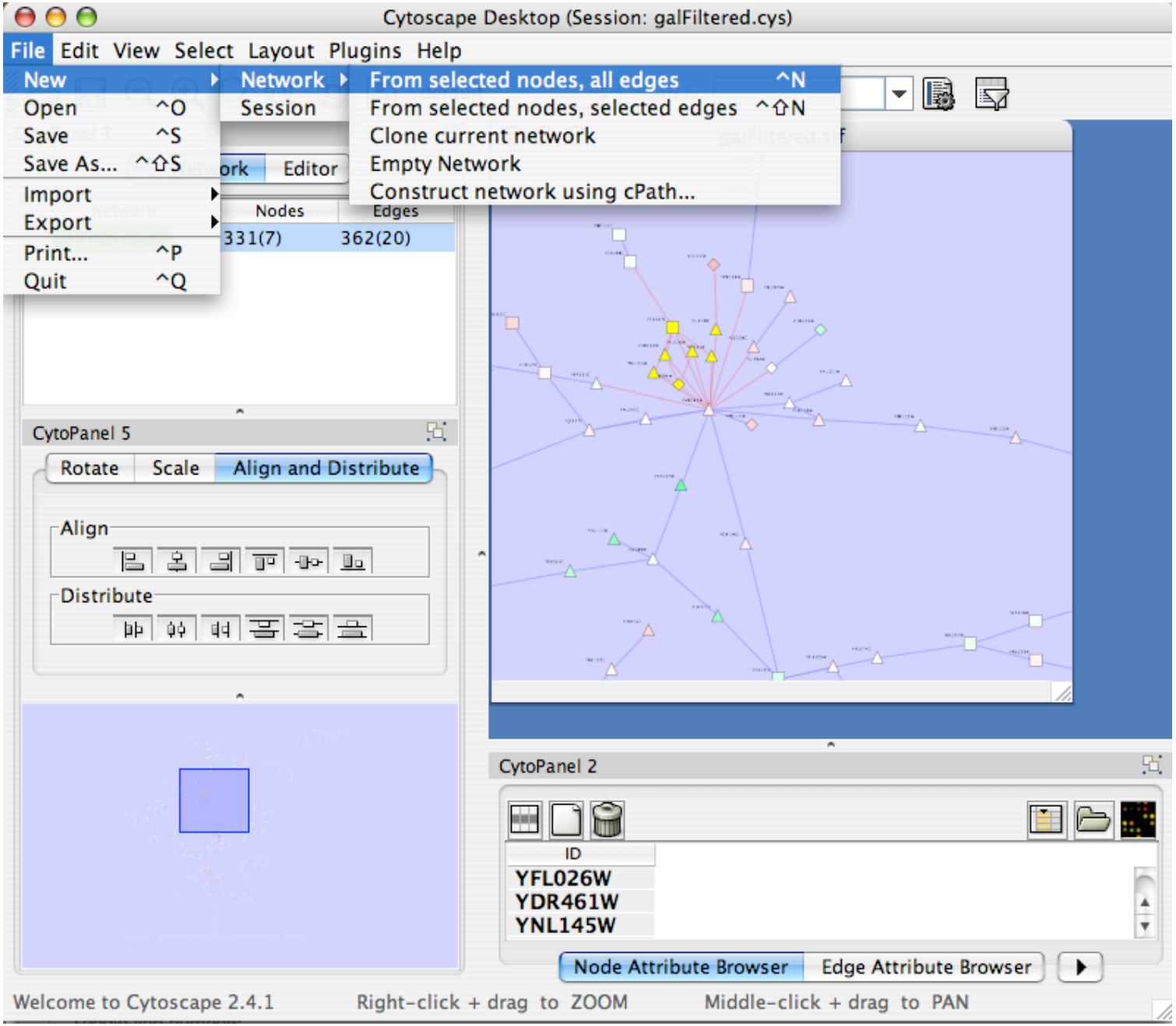
Network Layout

- 15 algorithms available through plugins
- Demo: Move, zoom/pan, rotate, scale, align

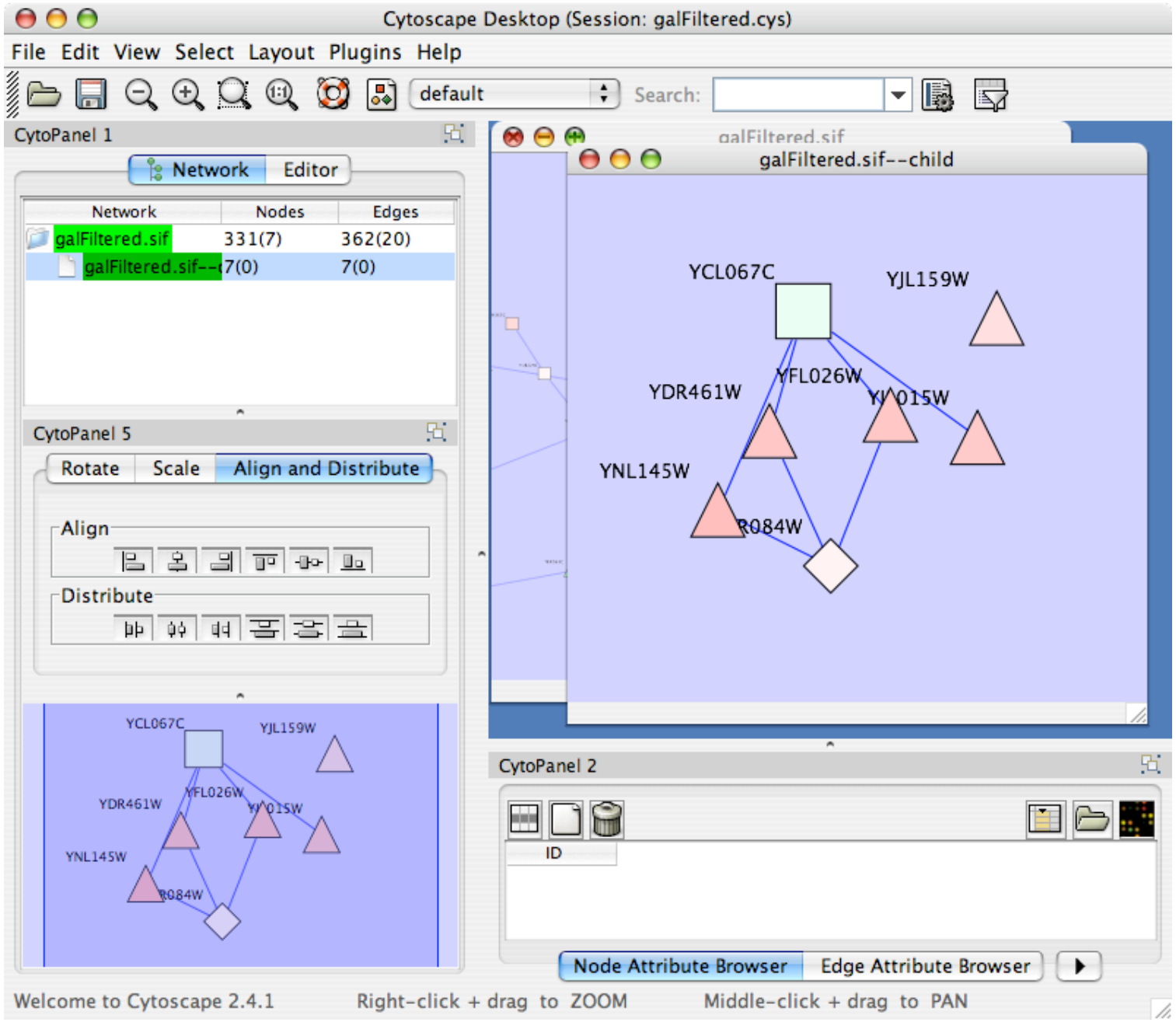




Create subnetwork



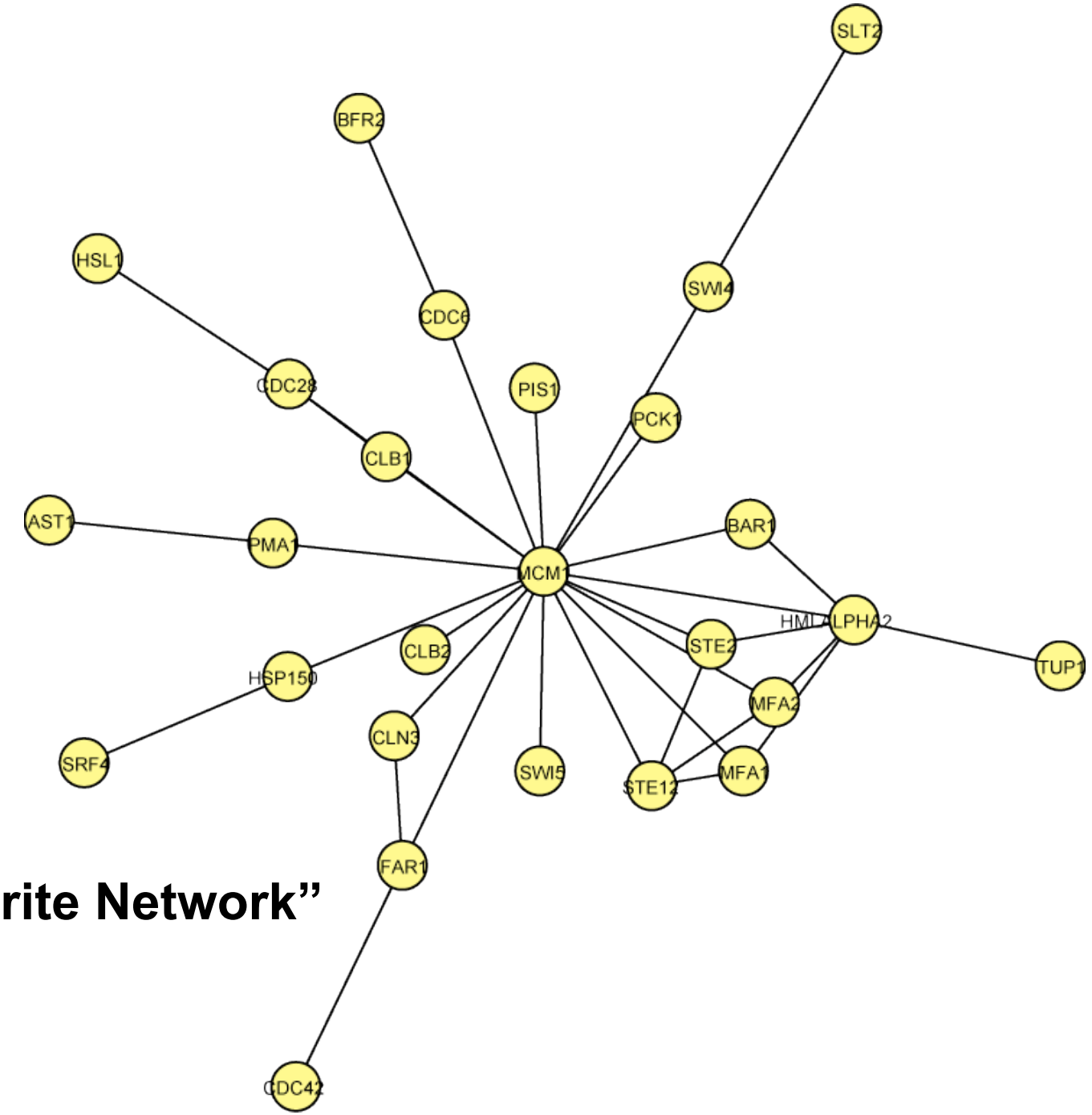
Create subnetwork



Visual Style

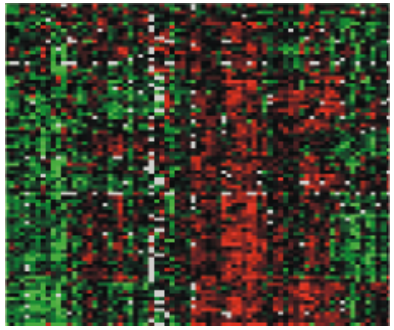
- Customized views of experimental data in a network context
- Network has node and edge attributes
  - E.g. expression data, GO function, interaction type
- Mapped to visual attributes
  - E.g. node/edge size, shape, colour…
- E.g. Visualize gene expression data as node colour gradient on the network

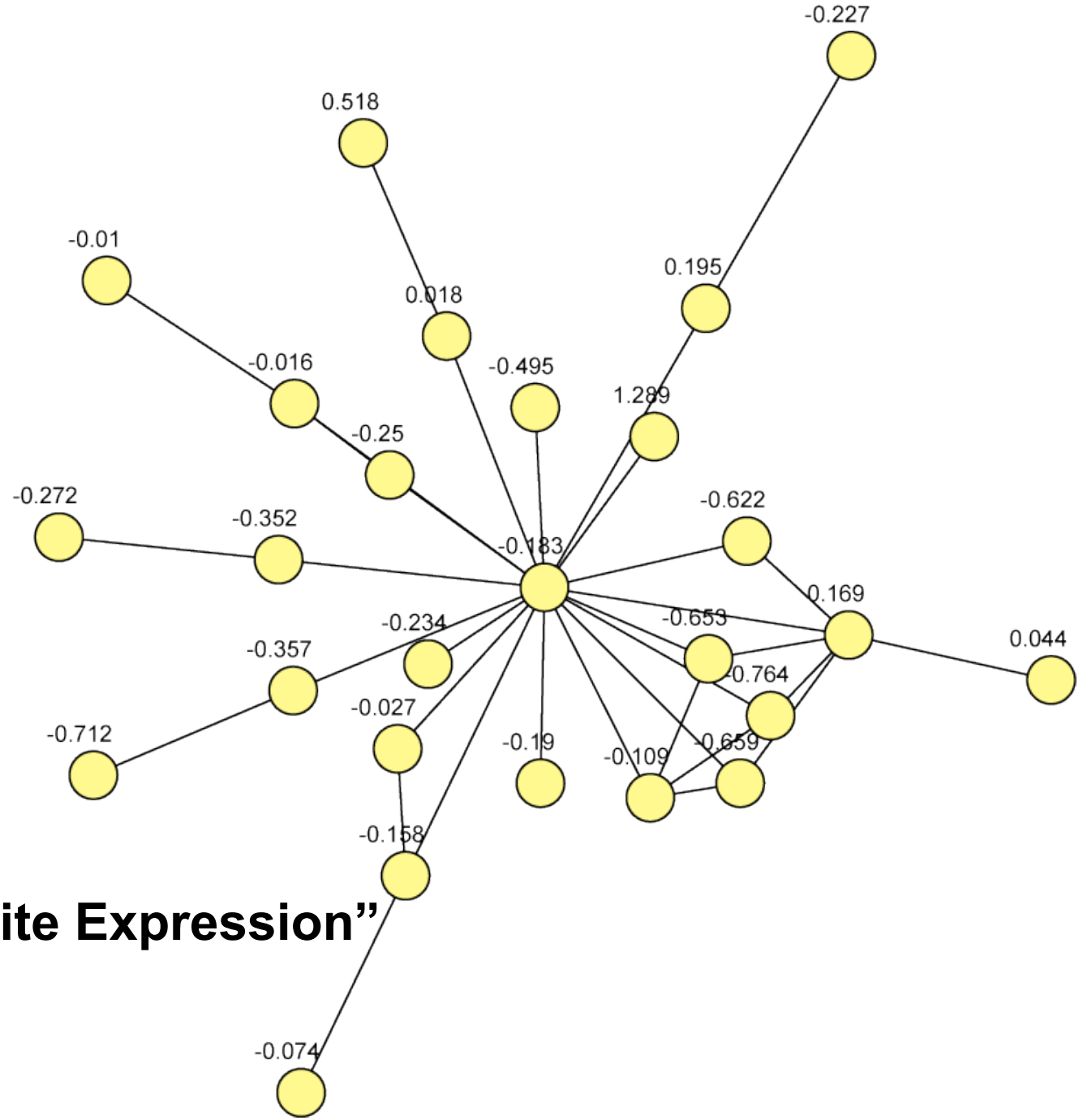
Visual Style



Load “Your Favorite Network”

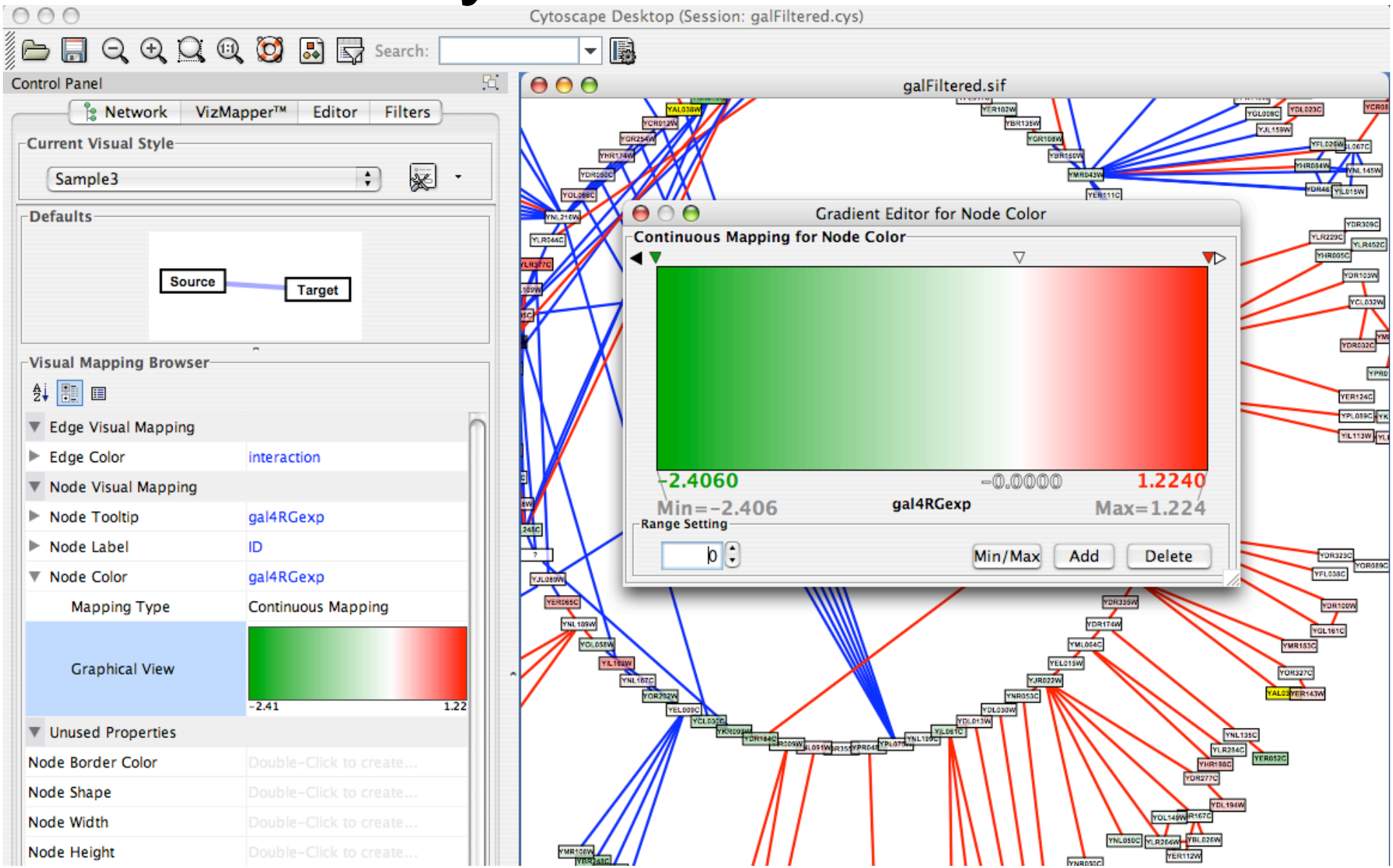
Visual Style





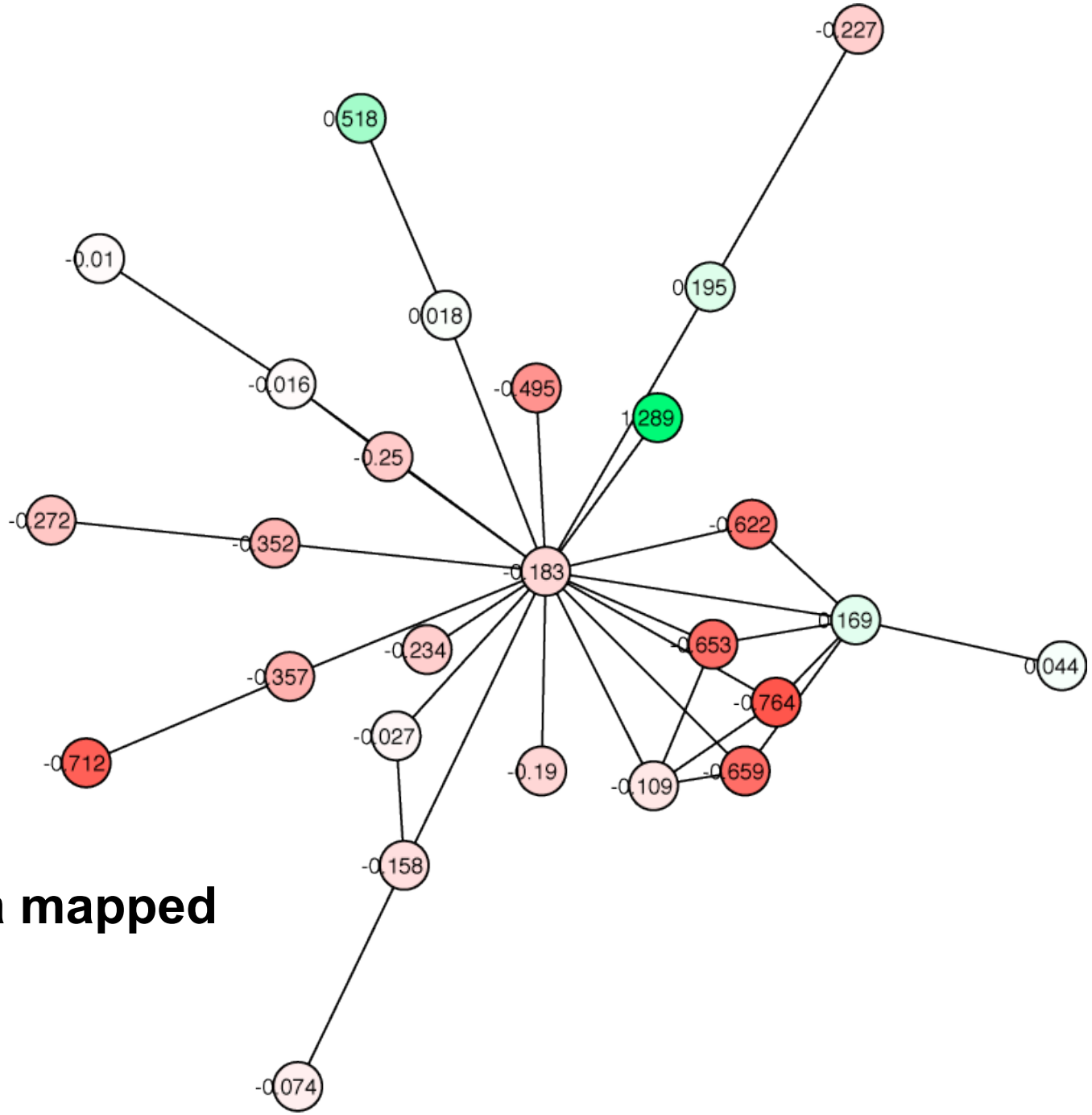
Load “Your Favorite Expression” Dataset

Visual Style



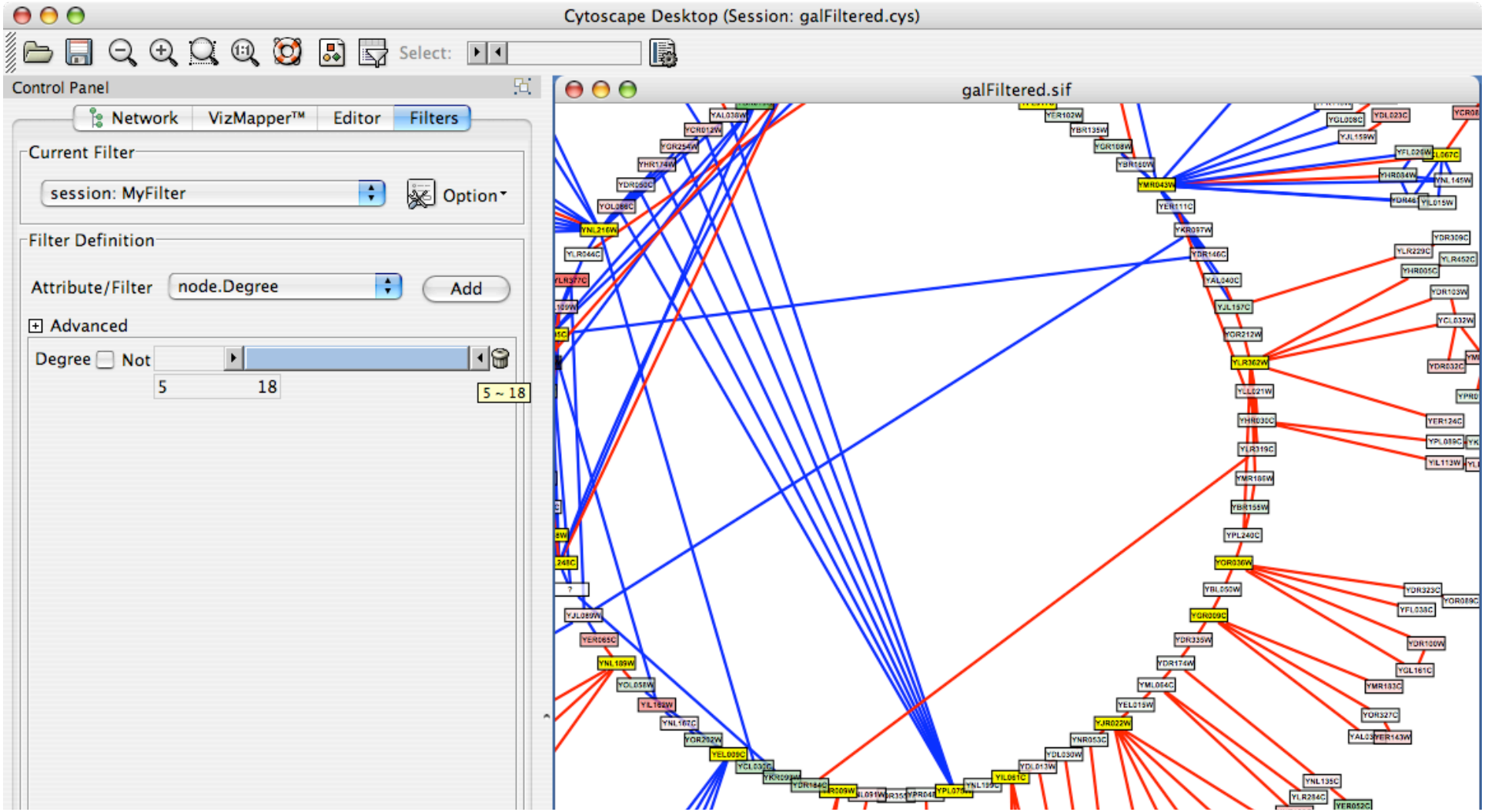
Map expression values to node colours using a continuous mapper

Visual Style



Expression data mapped to node colours

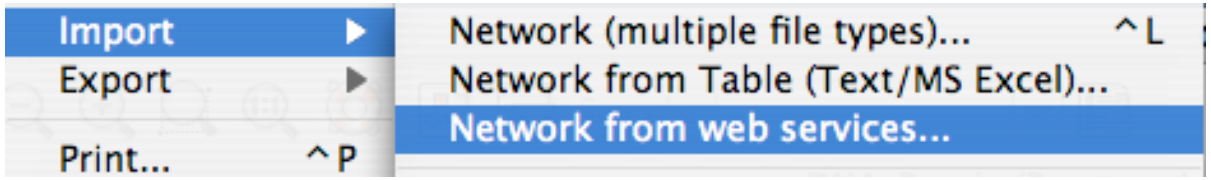
Network Filtering

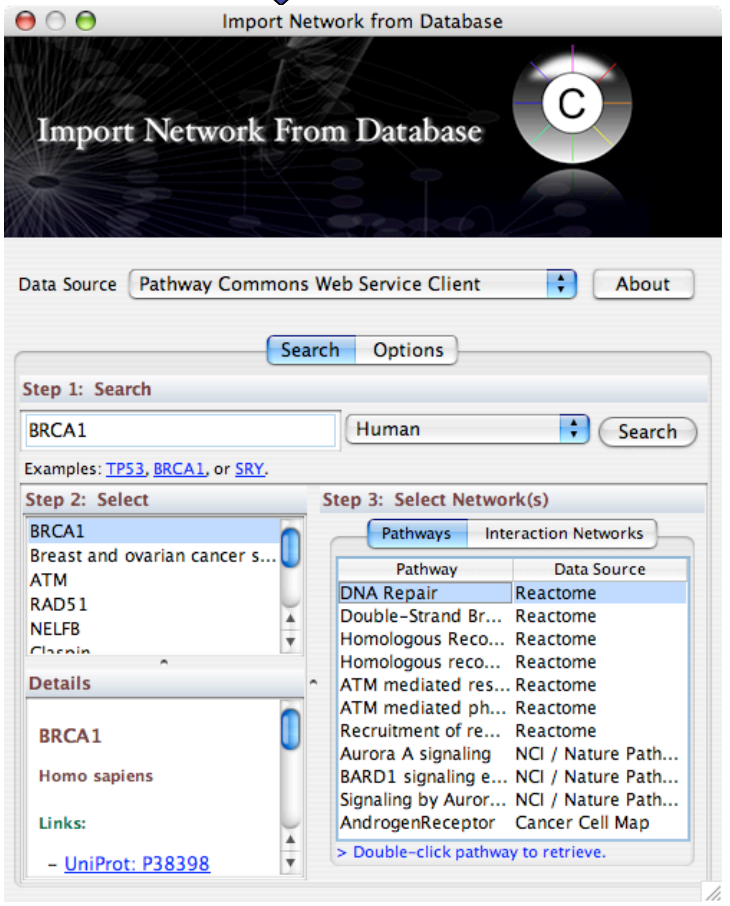


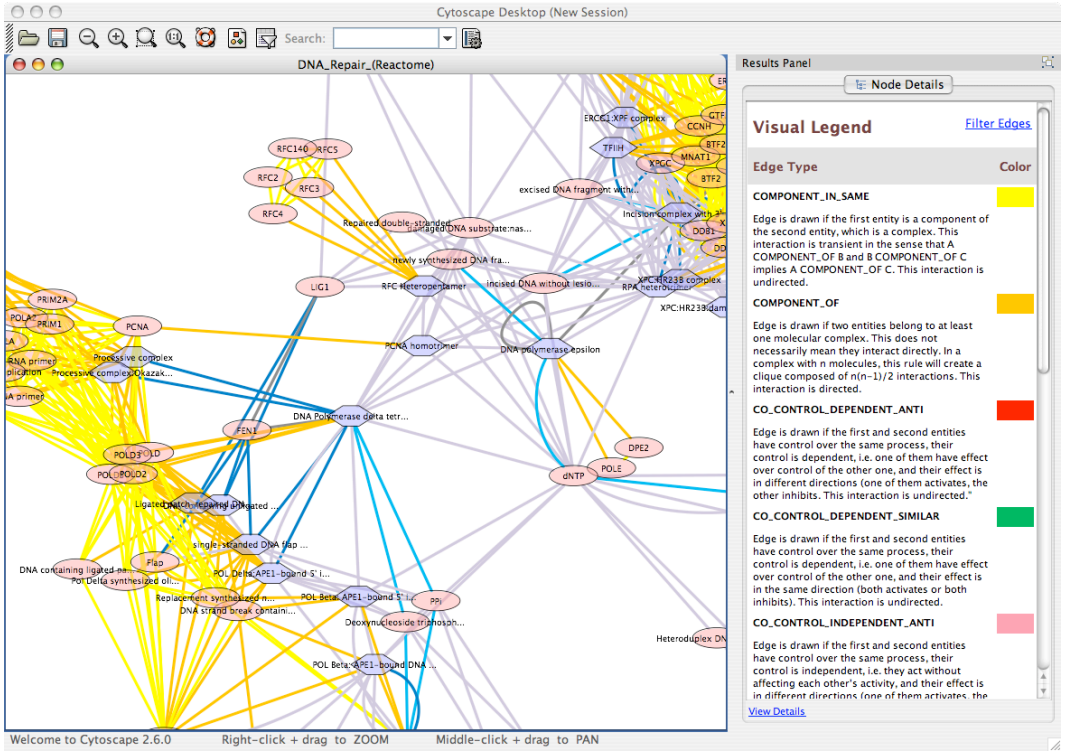
Analyzing gene expression data in a network context

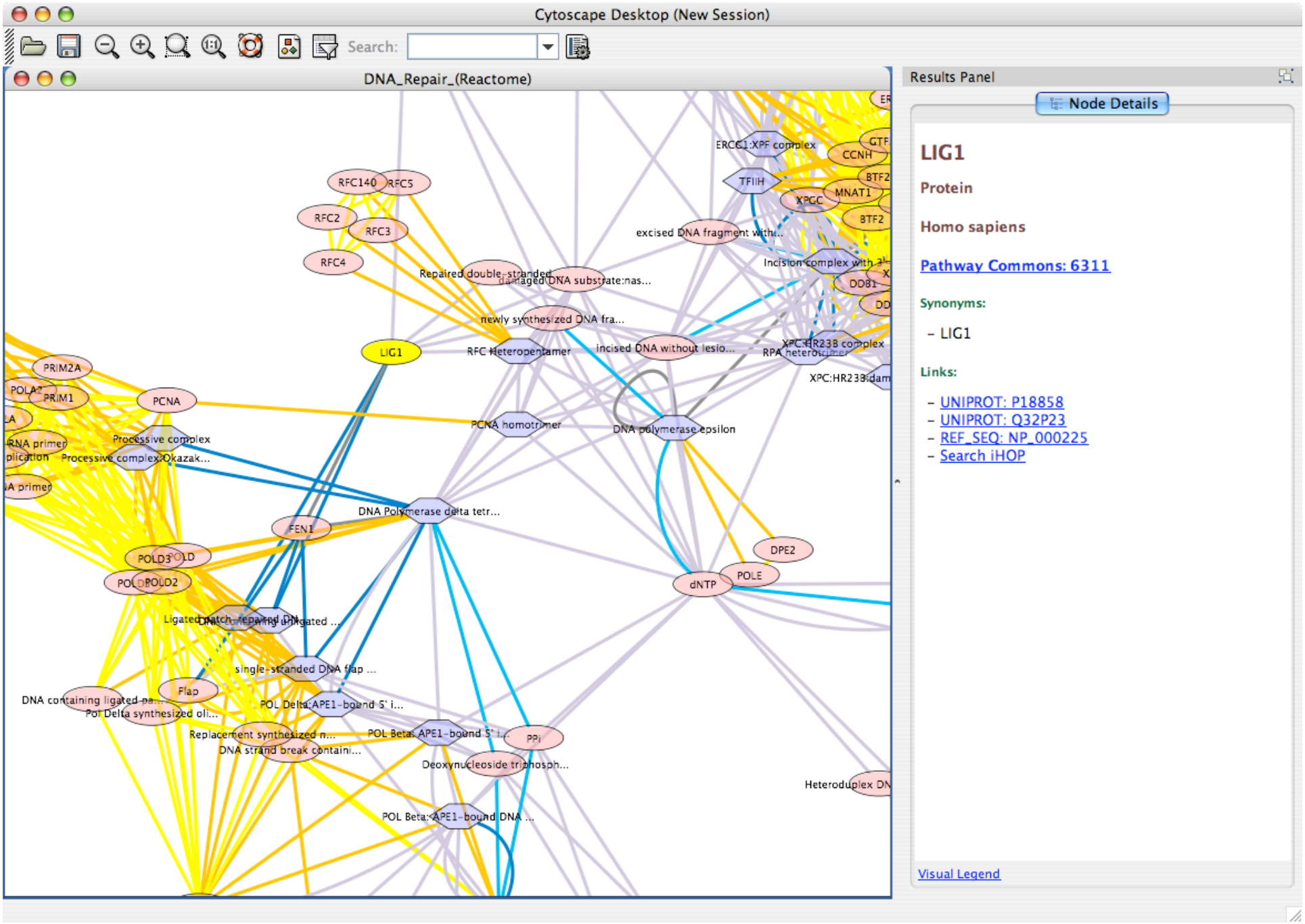
- Input
  - Gene expression data
  - Network data
- Output
  - Visual diagram of expression data on network
  - Active network regions
- Outline
  - Where to find network data?
    - Interaction database (cPath)
    - Literature associations via text mining
  - Load expression data
  - Identify active pathways

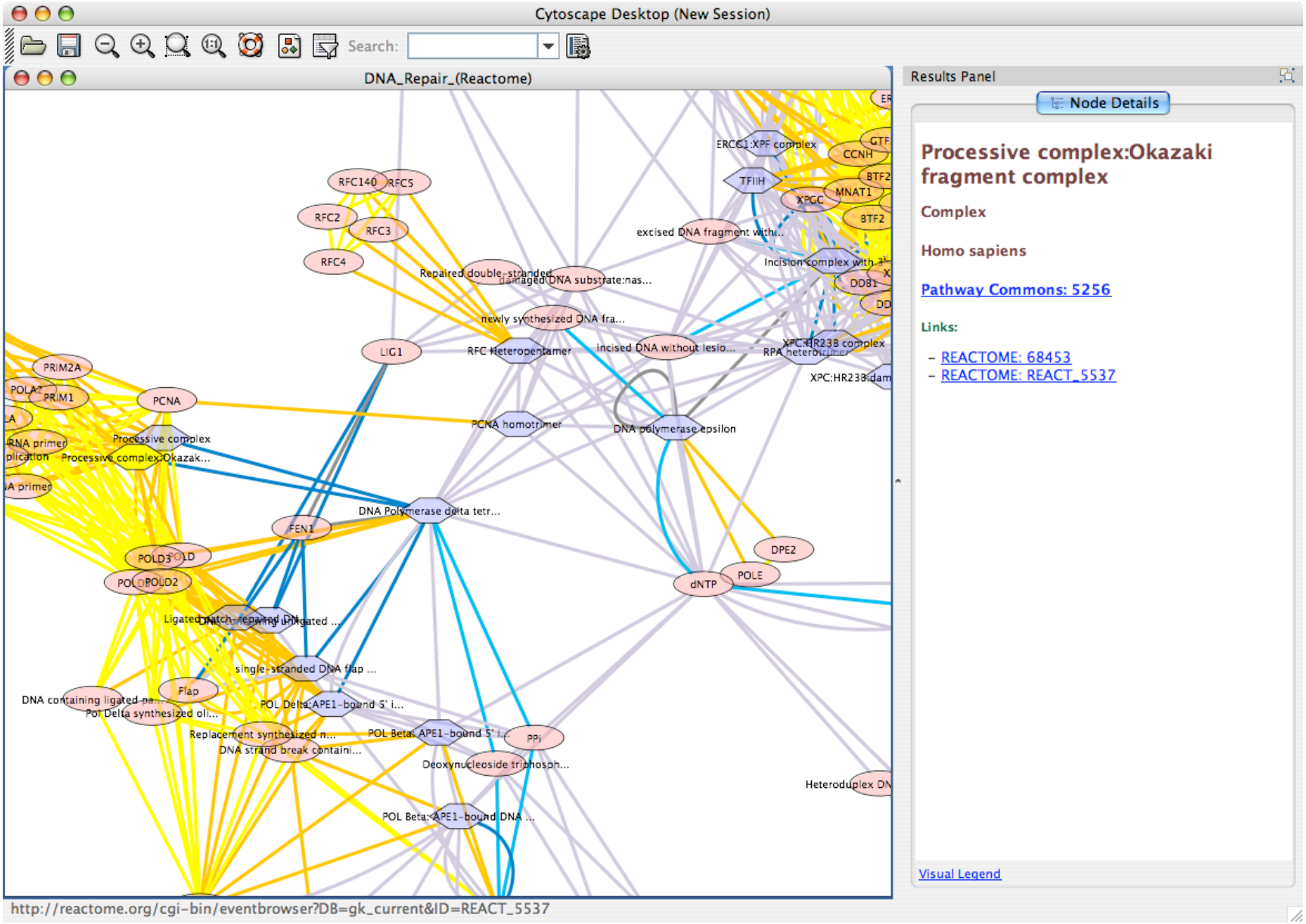
Interaction Database Search





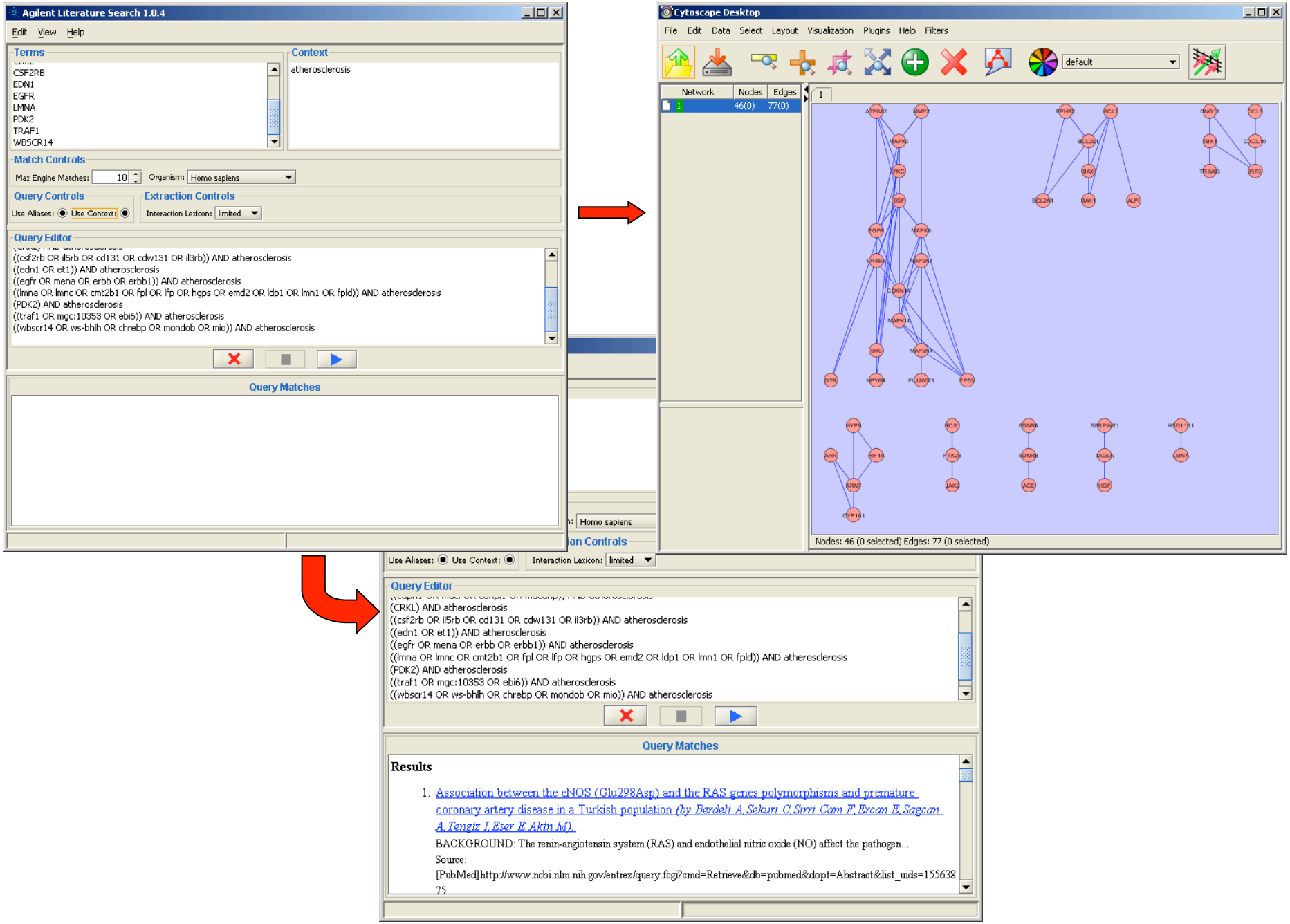



Cytoscape Desktop (New Session)
Search:
DNA_Repair_(Reactome)
Results Panel
Node Details
LIG1
Protein
Homo sapiens
Pathway Commons: 6311
Synonyms:
- LIG1
Links:
- UNIPROT: P18858
- UNIPROT: Q32P23
- REF_SEQ: NP_000225
- Search iHOP
Visual Legend
RFC140
RFC5
RFC2
RFC3
RFC4
LIG1
RFC Heteropentamer
Repaired double-stranded
damaged DNA substrate:nas...
newly synthesized DNA fra...
incised DNA without lesio...
ERCC1:XPF complex
TFIIH
XPCC
MNAT1
CCNH
BTF2
excised DNA fragment with...
Incision complex with 3'
DDB1
XPC:HR23B complex
RPA heterotrimer
XPC:HR23B:dam
PRIM2A
PRIM1
PCNA
PCNA homotrimer
DNA polymerase epsilon
Processive complex
Processive complex:Okazak...
DNA Polymerase delta tetr...
FEN1
POLD3
POLD2
DPE2
POLE
dNTP
Ligated DNA containing unligated ...
single-stranded DNA flap ...
Flap
DNA containing ligated na...
Pol Delta synthesized oli...
POL Delta:APE1-bound 5' i...
Replacement synthesized n...
DNA strand break containi...
POL Beta:APE1-bound 5' i...
PPi
Deoxynucleoside triphosph...
Heteroduplex DN
POL Beta:APE1-bound DNA ...

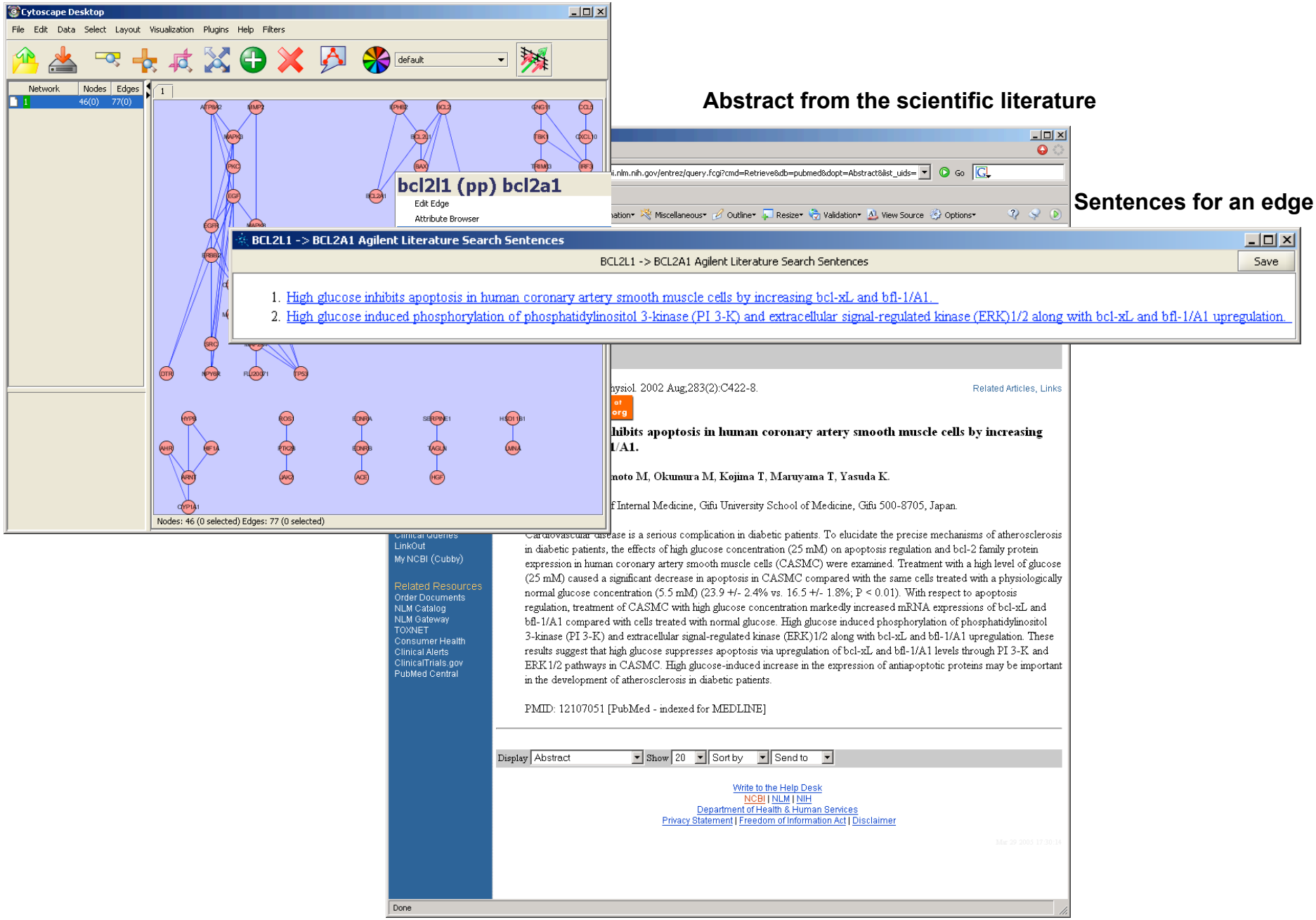

Cytoscape Desktop (New Session)
Search:
DNA_Repair_(Reactome)
Results Panel
Node Details
Processive complex:Okazaki fragment complex
Complex
Homo sapiens
Pathway Commons: 5256
Links:
REACTOME: 68453
REACTOME: REACT_5537
Visual Legend
http://reactome.org/cgi-bin/eventbrowser?DB=gk_current&ID=REACT_5537
RFC140
RFC5
RFC2
RFC3
RFC4
LIG1
PRIM2A
PRIM1
PCNA
Processive complex
Processive complex:Okazak...
FEN1
POLD3
POLD2
Flap
DNA containing ligated na...
Pol Delta synthesized oli...
Replacement synthesized n...
DNA strand break containi...
Repaired double-stranded...
damaged DNA substrate:nas...
newly synthesized DNA fra...
RFC Heteropentamer
PCNA homotrimer
DNA Polymerase delta tetr...
single-stranded DNA flap ...
POL Delta:APE1-bound 5' i...
POL Beta:APE1-bound 5' i...
PPi
Deoxynucleoside triphosph...
POL Beta:APE1-bound DNA ...
ERCC1:XPF complex
TFIIH
GTF
CCNH
BTF2
XPCC
MNAT1
excised DNA fragment with...
Incision complex with 3'
DDB1
incised DNA without lesio...
RPA heterotrimer
XPC:HR23B complex
DNA polymerase epsilon
DPE2
POLE
dNTP
Heteroduplex DN

Text Mining

- Computationally extract gene relationships from text, usually PubMed abstracts
- Literature search tool, lots of network data
- BUT not perfect
  - Problems recognizing gene names
  - Natural language processing not perfect
- Agilent Literature Search Cytoscape plugin
- Others: E.g. iHOP
  - www.ihop-net.org/UniPub/iHOP/


Agilent Literature Search 1.0.4
Edit View Help
Terms
CSF2RB
EDN1
EGFR
LMNA
PDK2
TRAF1
WBSCR14
Context
atherosclerosis
Match Controls
Max Engine Matches: 10
Organism: Homo sapiens
Query Controls
Use Aliases: Use Context:
Extraction Controls
Interaction Lexicon: limited
Query Editor
((csf2rb OR il5rb OR cd131 OR cdw131 OR il3rb)) AND atherosclerosis
((edn1 OR et1)) AND atherosclerosis
((egfr OR mena OR erbb OR erbb1)) AND atherosclerosis
((lmna OR lmnc OR cmt2b1 OR fpl OR lfp OR hgps OR emd2 OR ldp1 OR lmn1 OR fpld)) AND atherosclerosis
(PDK2) AND atherosclerosis
((traf1 OR mgc:10353 OR ebi6)) AND atherosclerosis
((wbscr14 OR ws-bhlh OR chrebp OR mondob OR mio)) AND atherosclerosis
Query Matches
Cytoscape Desktop
File Edit Data Select Layout Visualization Plugins Help Filters
default
Network Nodes Edges
1 46(0) 77(0)
Nodes: 46 (0 selected) Edges: 77 (0 selected)
Homo sapiens
Use Aliases: Use Context:
Interaction Lexicon: limited
Query Editor
(CRKL) AND atherosclerosis
((csf2rb OR il5rb OR cd131 OR cdw131 OR il3rb)) AND atherosclerosis
((edn1 OR et1)) AND atherosclerosis
((egfr OR mena OR erbb OR erbb1)) AND atherosclerosis
((lmna OR lmnc OR cmt2b1 OR fpl OR lfp OR hgps OR emd2 OR ldp1 OR lmn1 OR fpld)) AND atherosclerosis
(PDK2) AND atherosclerosis
((traf1 OR mgc:10353 OR ebi6)) AND atherosclerosis
((wbscr14 OR ws-bhlh OR chrebp OR mondob OR mio)) AND atherosclerosis
Query Matches
Results
1. Association between the eNOS (Glu298Asp) and the RAS genes polymorphisms and premature coronary artery disease in a Turkish population (by Berdeli A, Sekuri C, Sirri Can F, Ercan E, Sagcan A, Tengiz I, Eser E, Akin M)
BACKGROUND: The renin-angiotensin system (RAS) and endothelial nitric oxide (NO) affect the pathogen...
Source:
[PubMed]http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=pubmed&dopt=Abstract&list_uids=15563875

Cytoscape Network produced by Literature Search.



Gene Expression/Network Integration

- Identifier (ID) mapping
  - Translation from network IDs to gene expression IDs e.g. Affymetrix probe IDs
  - Also: Unification, link out, query
  - Entrez gene IDs (genes), UniProt (proteins)
- Synergizer
  - llama.med.harvard.edu/cgi/synergizer/translate
- More ID mapping services available
  - http://baderlab.org/IdentifierMapping

Gene Expression/Network Integration

THE SYNERGIZER

The Synergizer database is a growing repository of gene and protein identifier synonym relationships. This tool facilitates the conversion of identifiers from one naming scheme (a.k.a "namespace") to another.

load sample inputs

Select species: Saccharomyces cerevisiae

Select authority: ensembl

Select "FROM" namespace: __ANY__

Select "TO" namespace: entrezgene [854192]

(NB: The strings in [brackets] are representative IDs in the corresponding namespaces.)

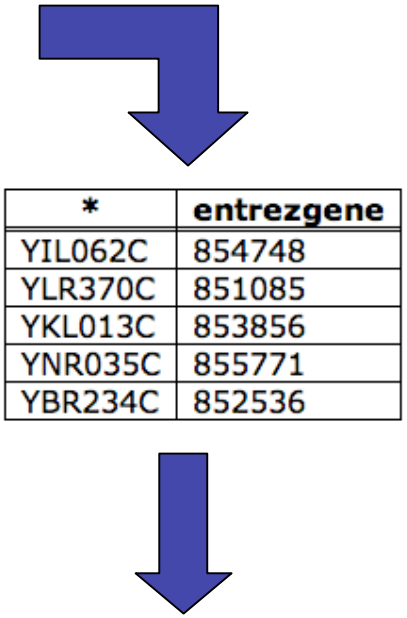
File containing IDs to translate: Browse...

and/or

IDs to translate:
YIL062C
YLR370C
YKL013C
YNR035C
YBR234C

Output as spreadsheet:

Submit

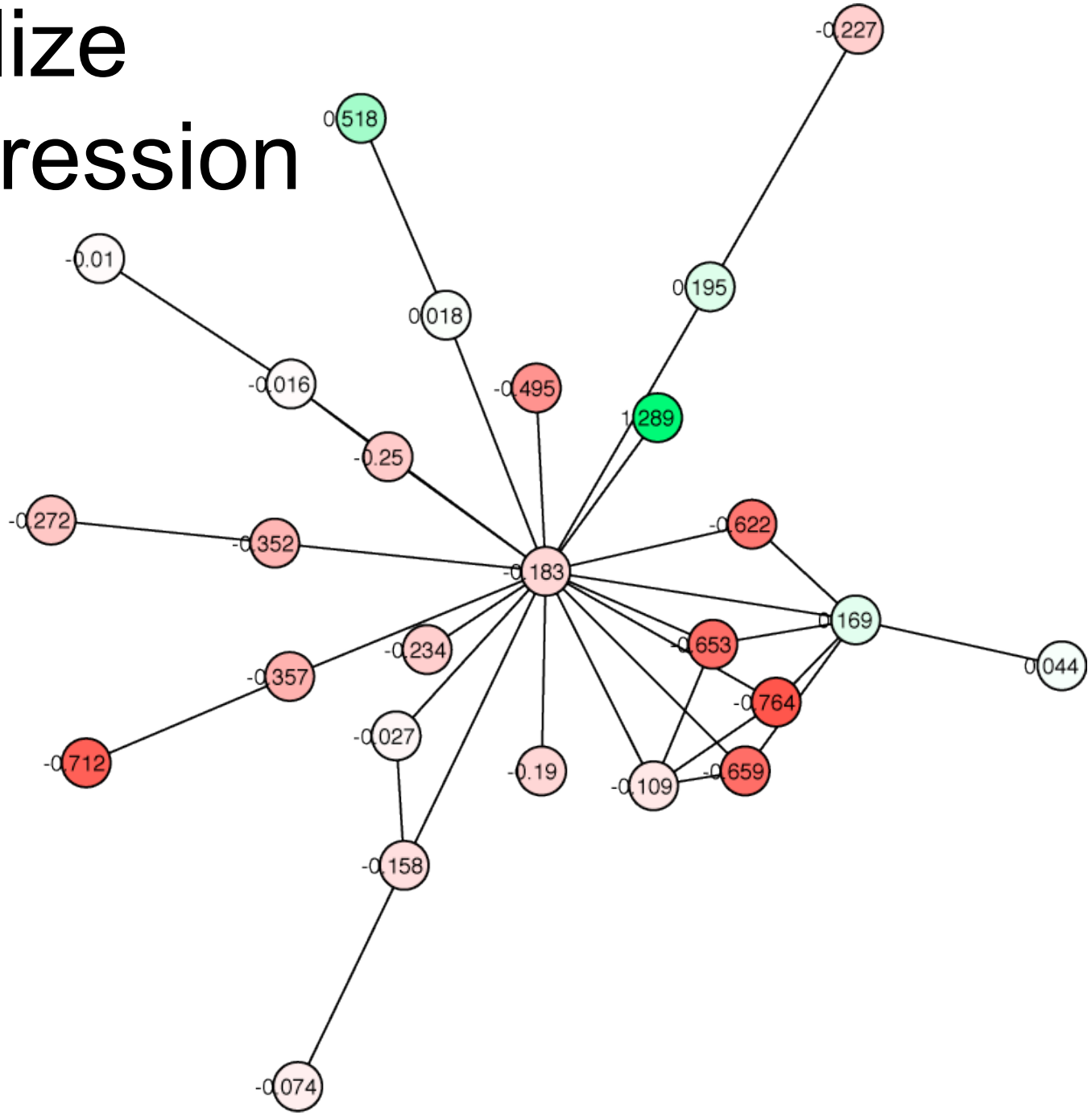


*	entrezgene
YIL062C	854748
YLR370C	851085
YKL013C	853856
YNR035C	855771
YBR234C	852536

1. Load as attributes in Cytoscape

2. Assign expression values to nodes using this attribute set

Visualize Gene Expression



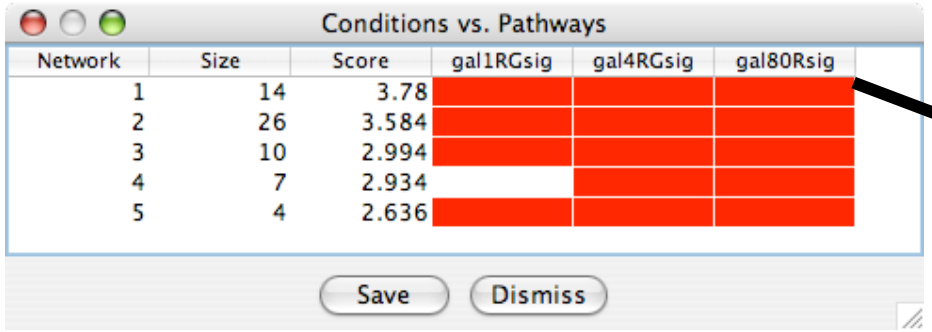
Find Active Subnetworks

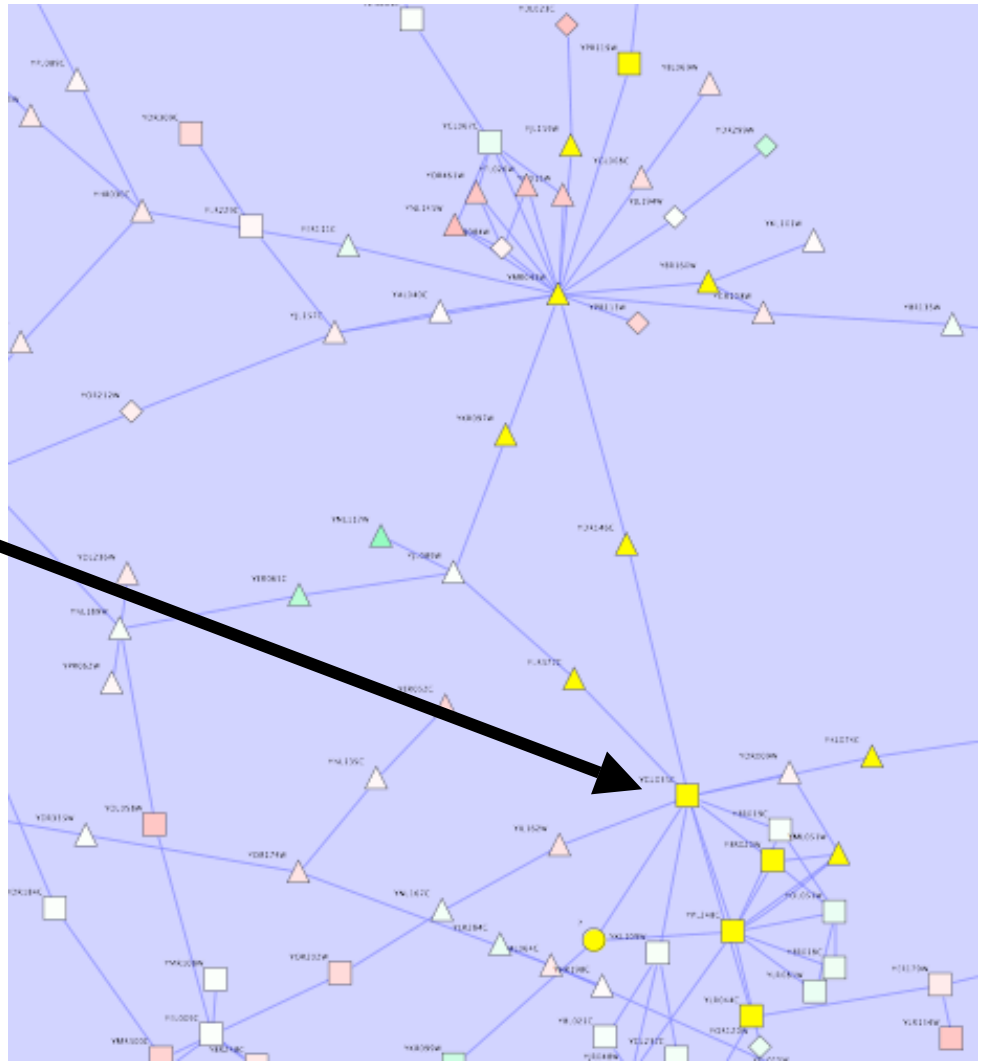
- Active modules
  - Input: network + p-values for gene expression values e.g. from GCRMA
  - Output: significantly differentially expressed subgraphs
- Method
  - Calculate z-score/node, Z_A score/subgraph, correct vs. random expression data sampling
  - Score over multiple experimental conditions
  - Simulated annealing used to find high scoring networks

Ideker T, Ozier O, Schwikowski B, Siegel AF Bioinformatics. 2002;18 Suppl 1:S233-40

Active Module Results

Network: yeast protein-protein and protein-dna network
Expression data: 3 gene knock out conditions (enzyme, TF activator, TF repressor)





Note: non-deterministic, multiple runs required for confidence of result robustness

Ideker T et al. Science. 2001 May 4;292(5518):929-34.

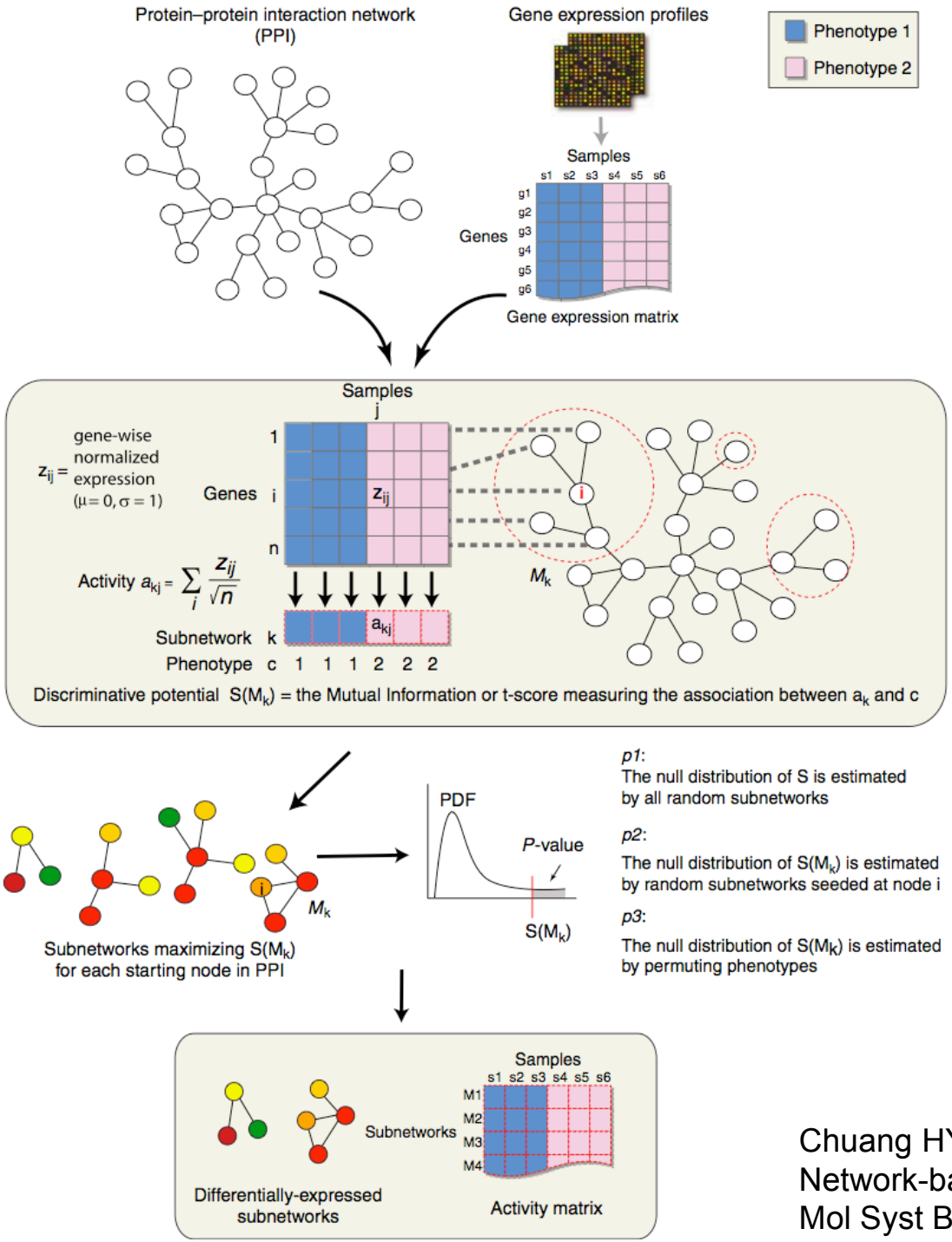
Network classification of disease

- Traditional: Gene association
- Limitations: Too many genes reduces statistical power
- New: Active cell map based approaches combining network and molecular profiles
- http://www.connotea.org/user/ACMT

Chuang HY, Lee E, Liu YT, Lee D, Ideker T
Network-based classification of breast cancer metastasis
Mol Syst Biol. 2007;3:140. Epub 2007 Oct 16

Liu M, Liberzon A, Kong SW, Lai WR, Park PJ, Kohane IS, Kasif S
Network-based analysis of affected biological processes in type 2 diabetes models
PLoS Genet. 2007 Jun;3(6):e96

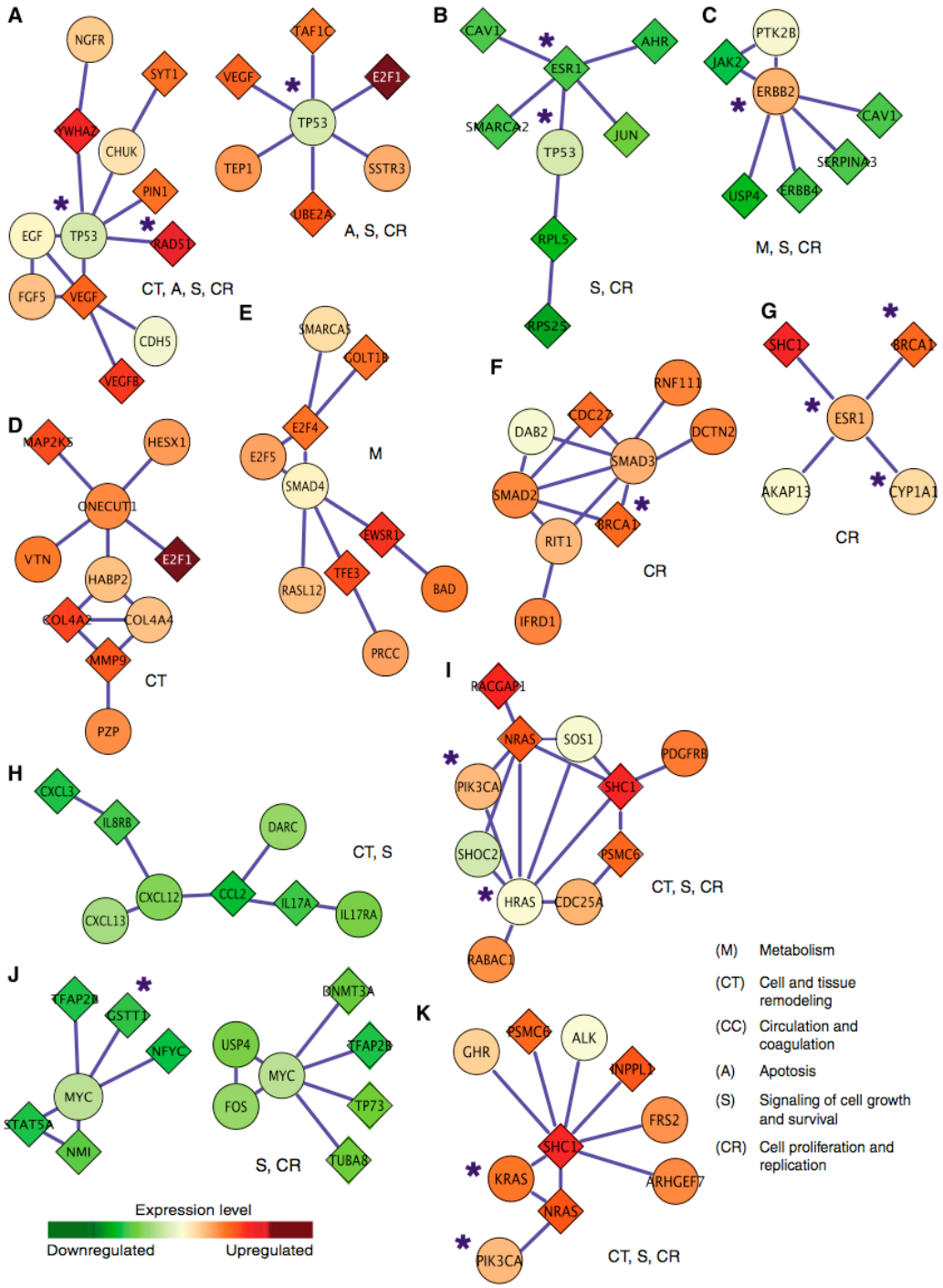
Efroni S, Schaefer CF, Buetow KH
Identification of key processes underlying cancer phenotypes using biologic pathway analysis
PLoS ONE. 2007 May 9;2(5):e425



Network-based breast cancer classification

- 57k intx from Y2H, orthology, co-citation, HPRD, BIND, Reactome
- 2 breast cancer cohorts, different expression platforms
- Metastasis: 78 vs. 217 (van de Vijver) & 106 vs. 180 (Wang)

Chuang HY, Lee E, Liu YT, Lee D, Ideker T
Network-based classification of breast cancer metastasis
Mol Syst Biol. 2007;3:140. Epub 2007 Oct 16



- Similar network markers across 2 data sets (better than original overlap)
- Increased classification accuracy
- Better coverage of known cancer risk genes (*)

Acknowledgements

Pathway Commons
Chris Sander
Ethan Cerami
Ben Gross
Emek Demir
Robert Hoffmann

Cytoscape
Trey Ideker (UCSD)
Ryan Kelley, Kei Ono, Mike Smoot, Peng Liang Wang (Nerius Landys, Chris Workman, Mark Anderson, Nada Amin, Owen Ozier, Jonathan Wang)

Lee Hood (ISB)
Sarah Killcoyne (Iliana Avila-Campillo, Rowan Christmas, Andrew Markiel, Larissa Kamenkovich, Paul Shannon)

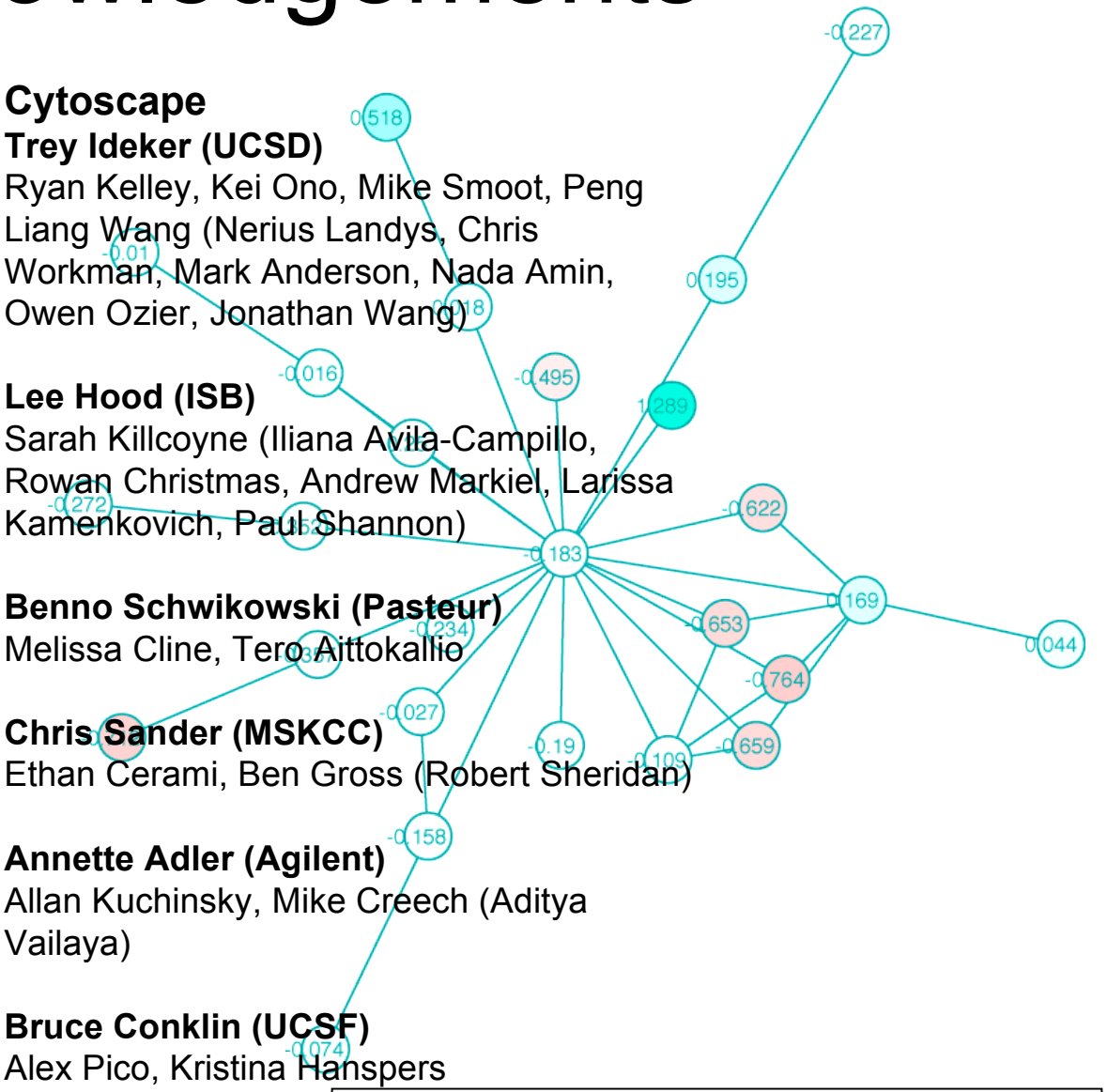
Benno Schwikowski (Pasteur)
Melissa Cline, Tero Aittokallio

Chris Sander (MSKCC)
Ethan Cerami, Ben Gross (Robert Sheridan)

Annette Adler (Agilent)
Allan Kuchinsky, Mike Creech (Aditya Vailaya)

Bruce Conklin (UCSF)
Alex Pico, Kristina Hanspers

Funding
CIHR
Genome Canada
Canada Foundation for Innovation



http://baderlab.org