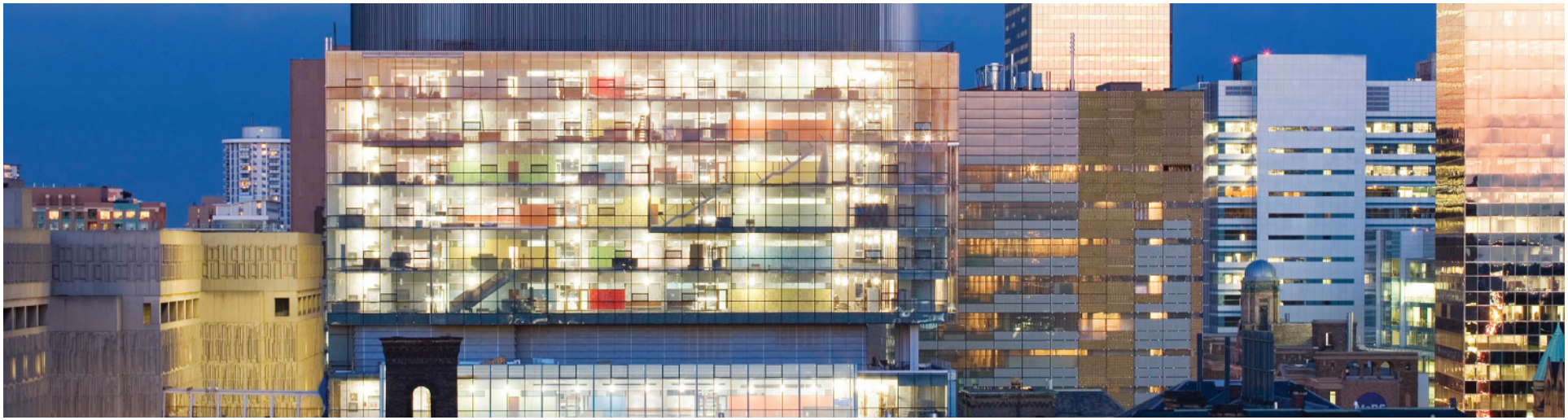
Integrative analysis of interaction networks

Gary Bader http://www.baderlab.org
JTB2010 - Nov.23.2009

Donnelly Centre
for Cellular + Biomolecular Research

UNIVERSITY OF TORONTO

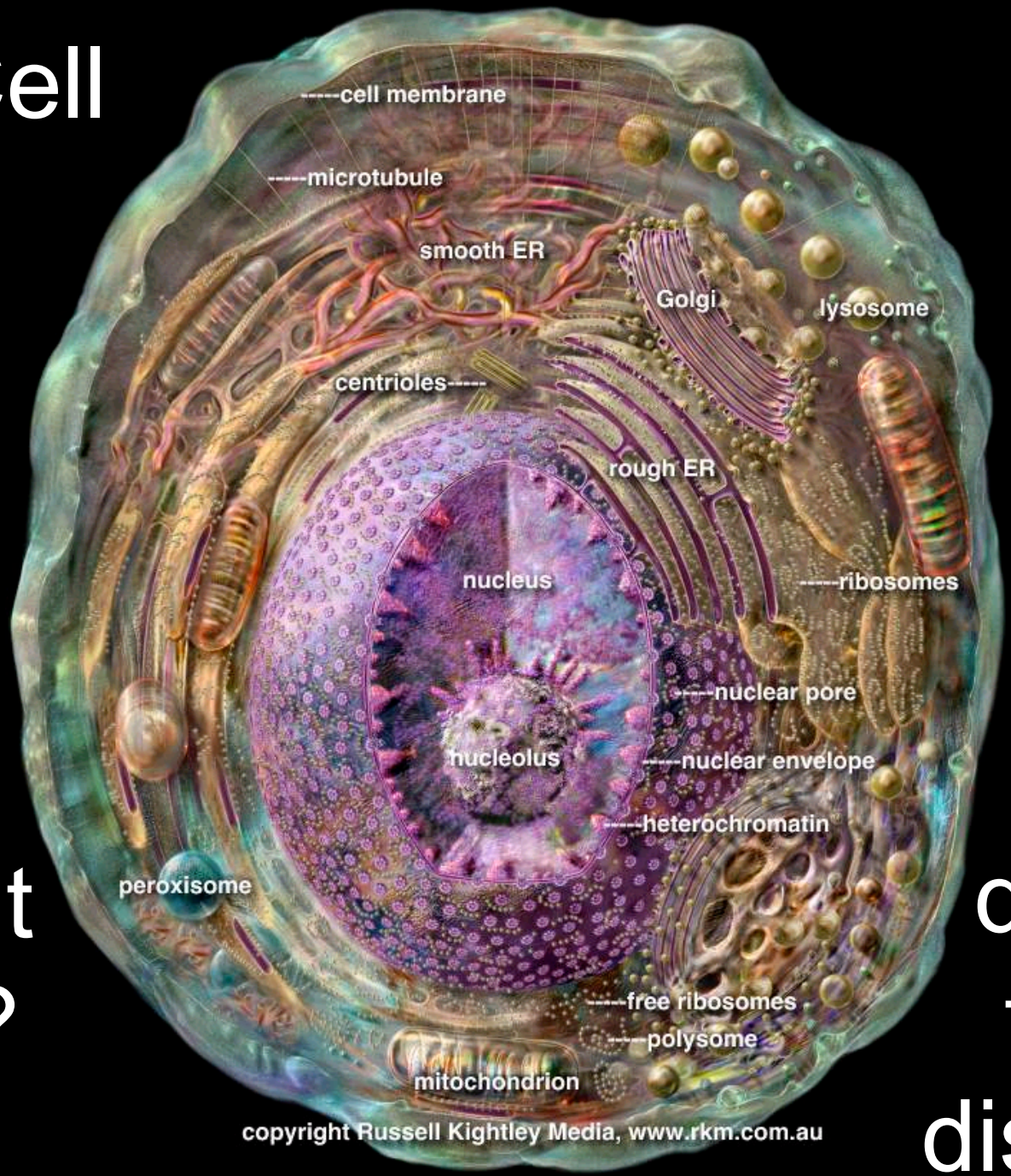


Outline

- Data integration using networks
- Network analysis
- Network data
- Network visualization and analysis
- Analyzing molecular profiles

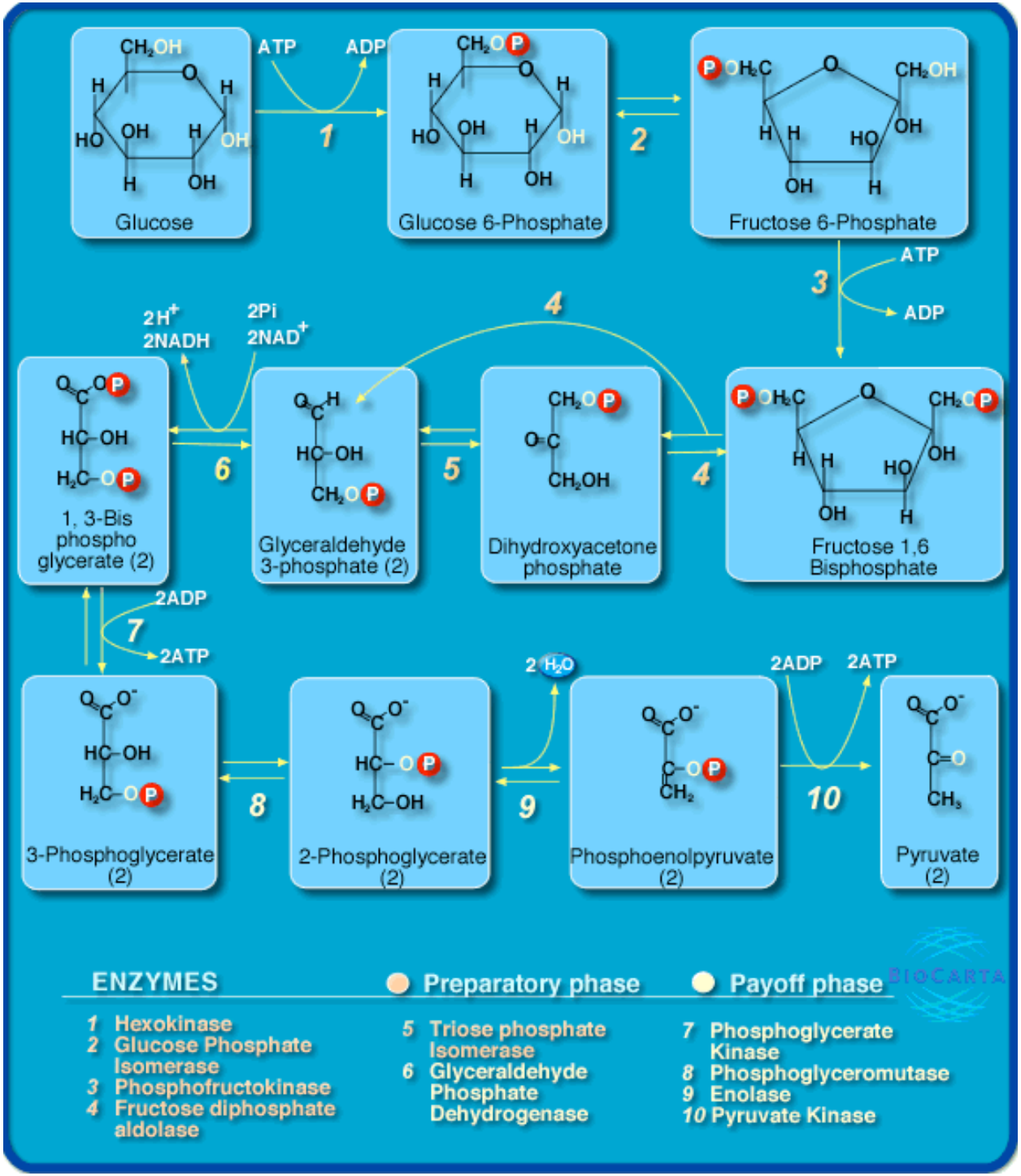
Data integration using networks

The Cell



How does it work?

How does it fail in disease?


Glucose
ATP
ADP
1
Glucose 6-Phosphate
2
Fructose 6-Phosphate
3
ATP
ADP
4
2H+
2NADH
2Pi
2NAD+
1, 3-Bis phospho glycerate (2)
6
Glyceraldehyde 3-phosphate (2)
5
Dihydroxyacetone phosphate
4
Fructose 1,6 Bisphosphate
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
BioCarta
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

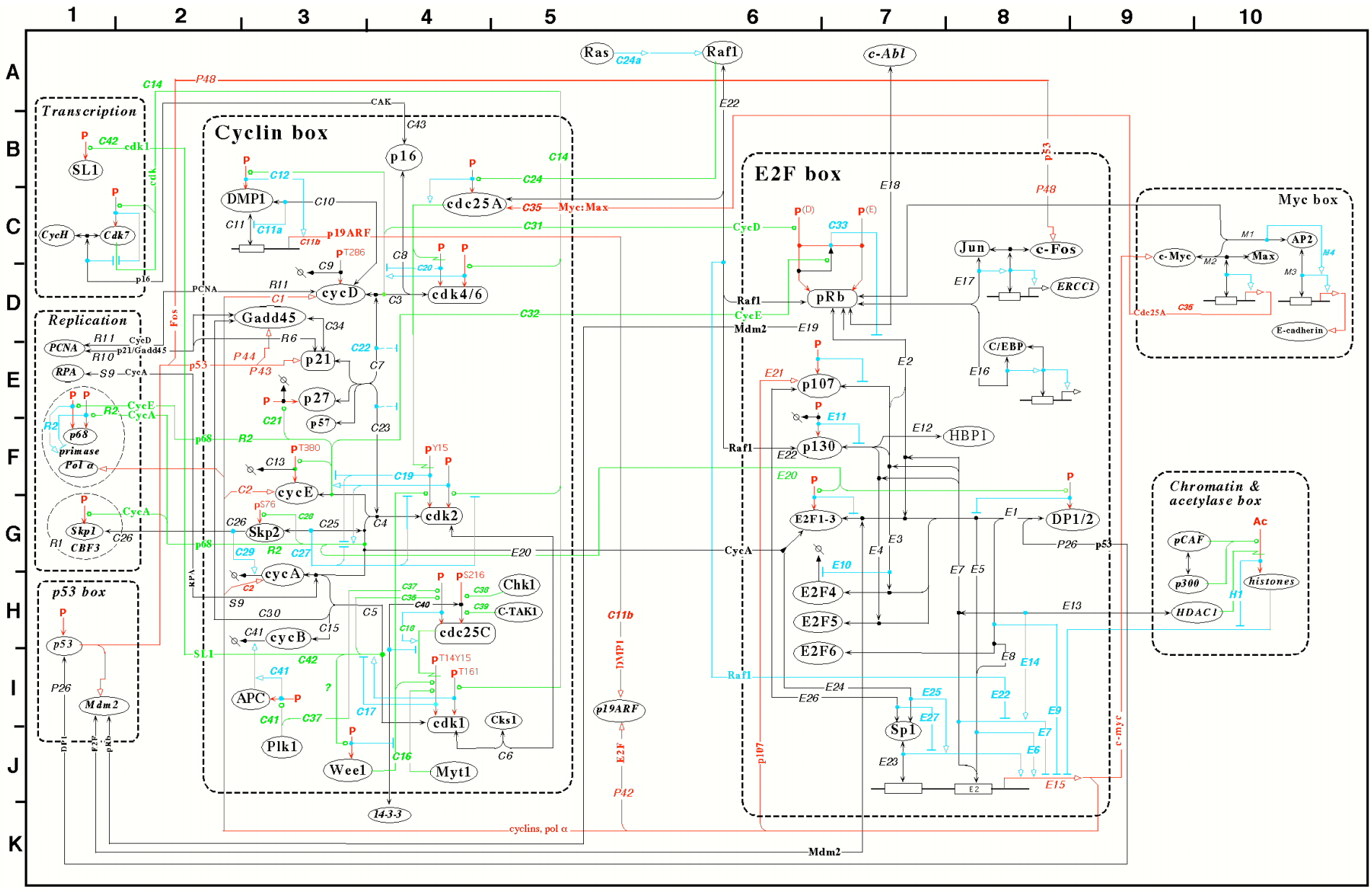
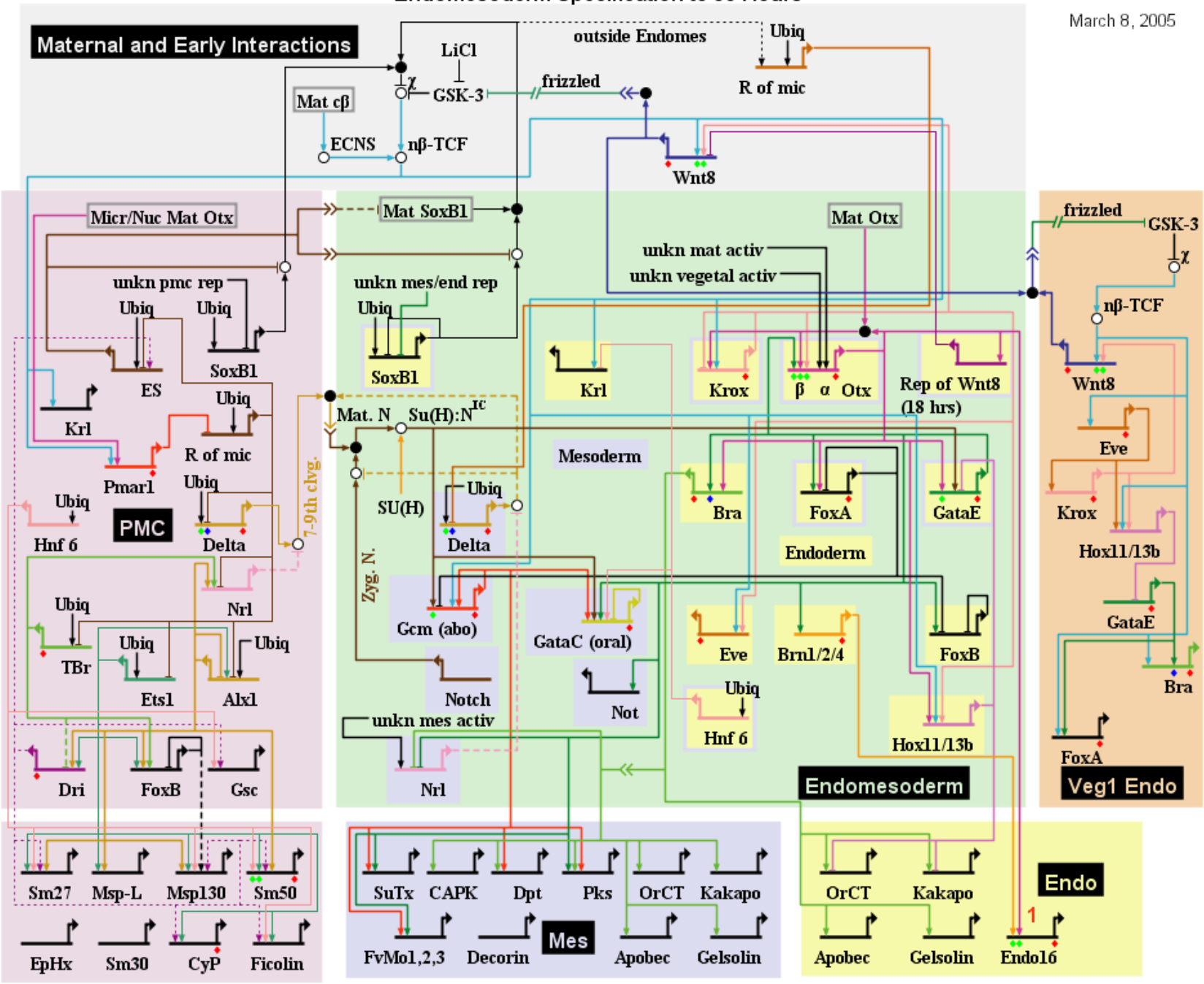


Figure 6A: The Cyclin - E2F cell cycle control system (version 3a - June 8, 1999)

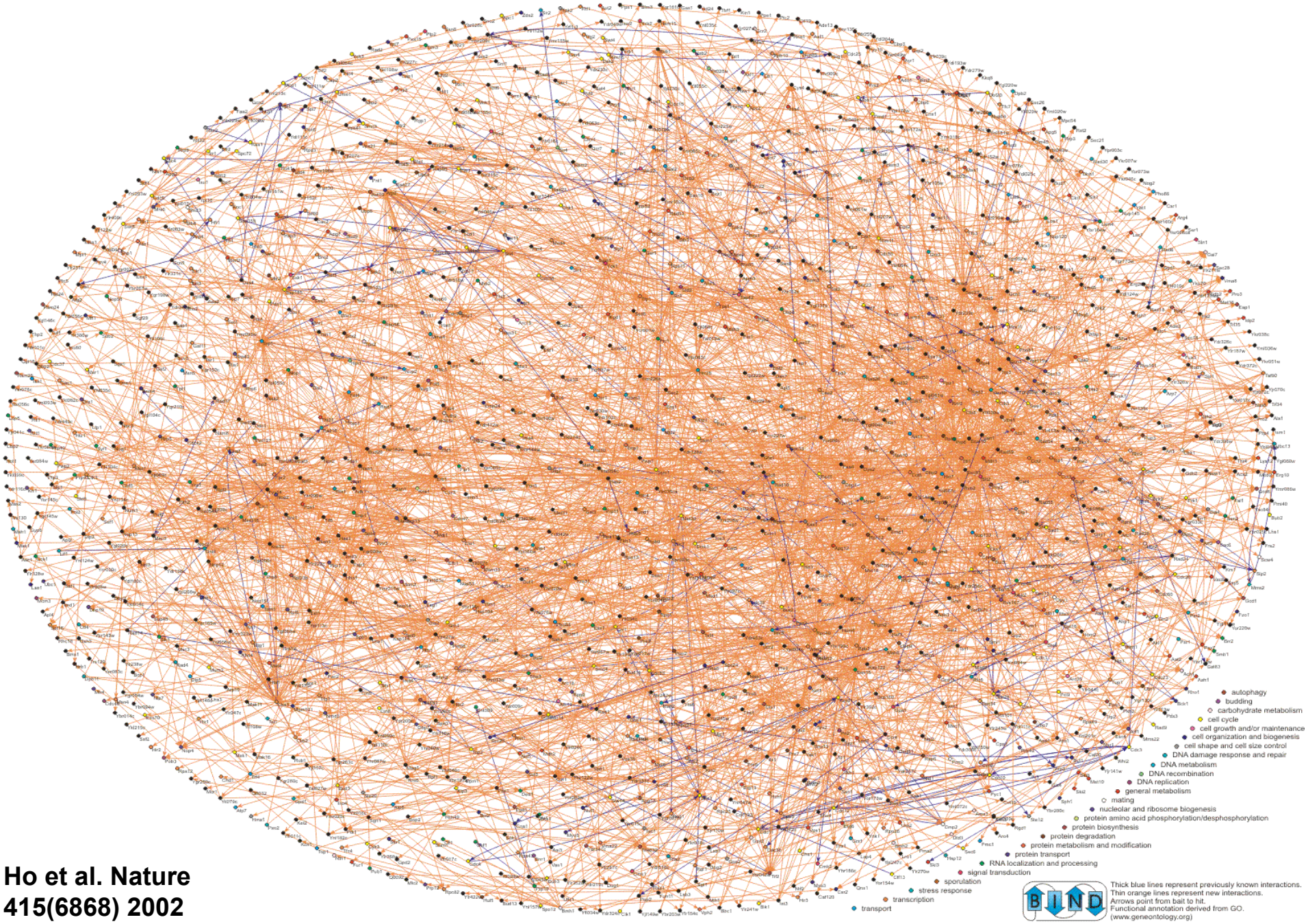
http://discover.nci.nih.gov/kohnk/interaction_maps.html

Endomesoderm Specification to 30 Hours

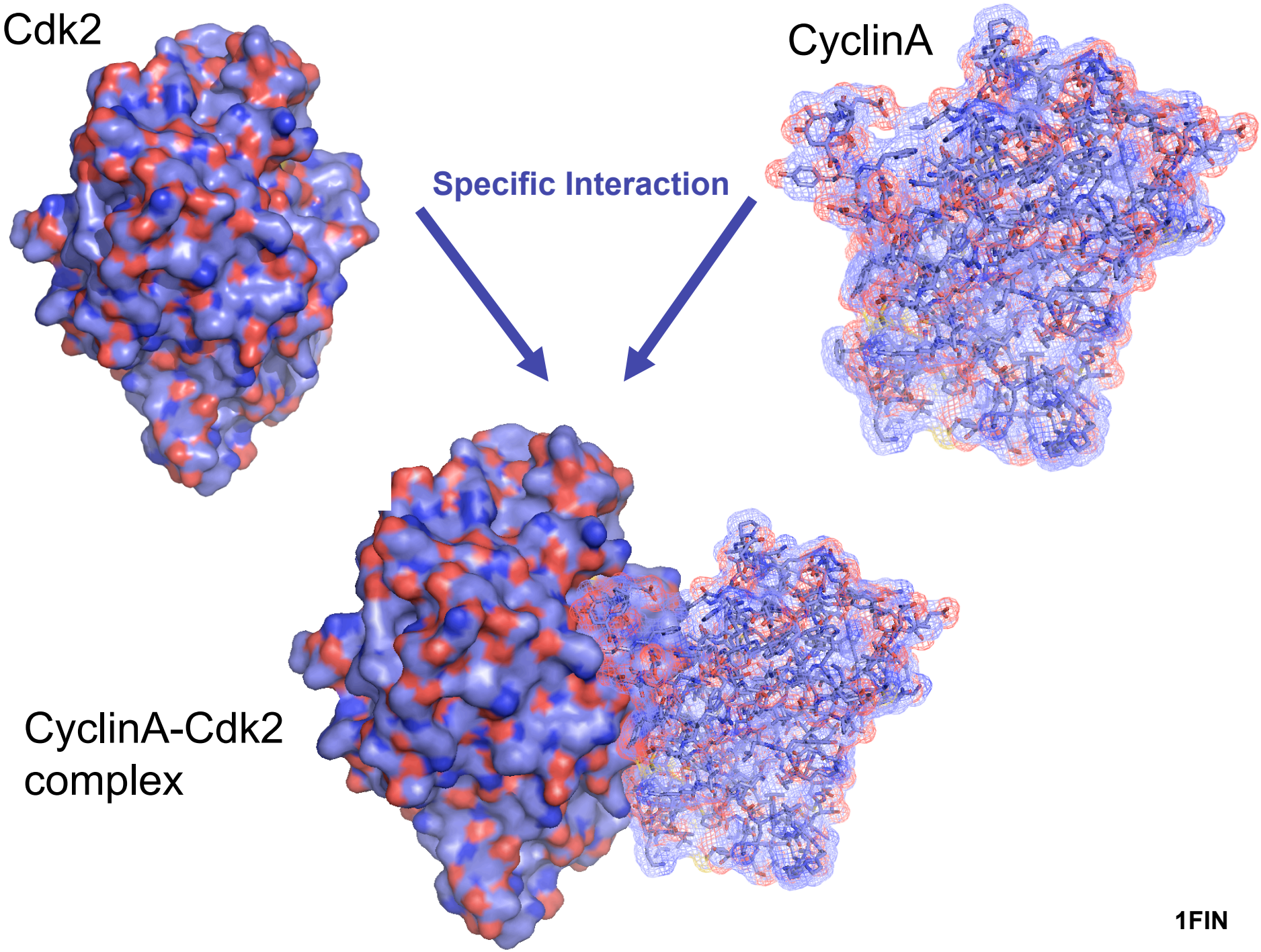
March 8, 2005



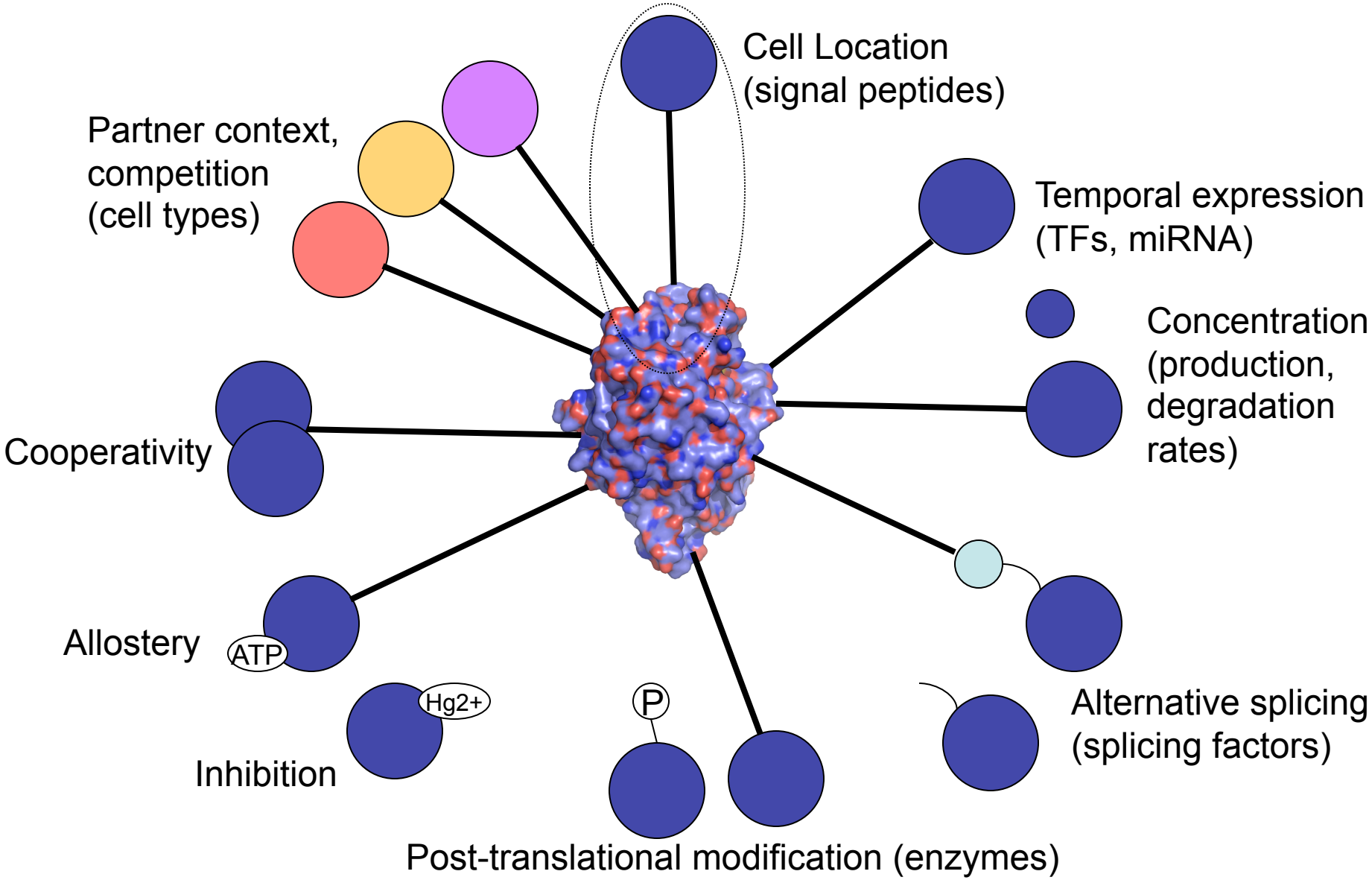
Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry



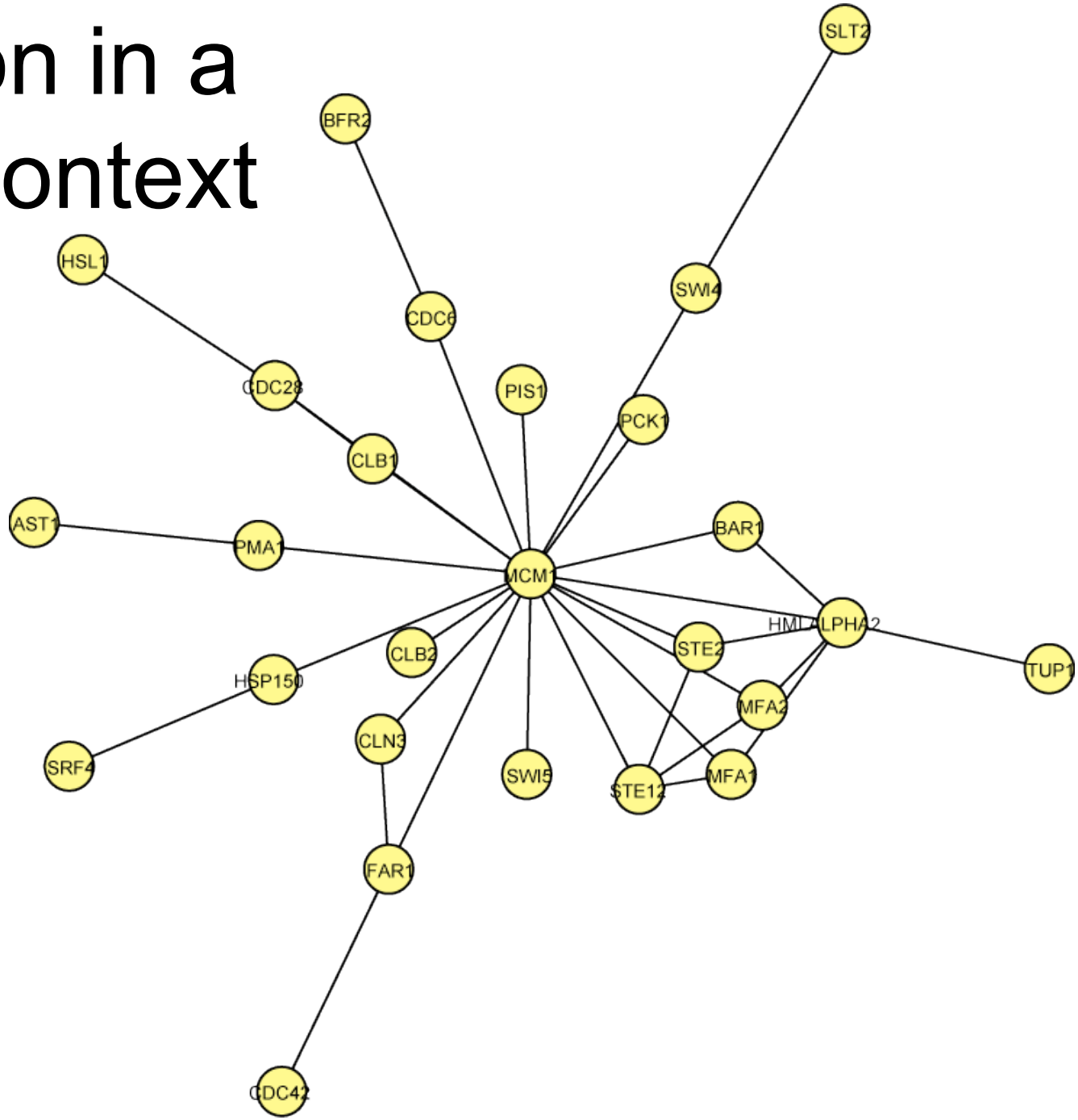
Ho et al. Nature 415(6868) 2002


Cdk2
CyclinA
Specific Interaction
CyclinA-Cdk2
complex
1FIN

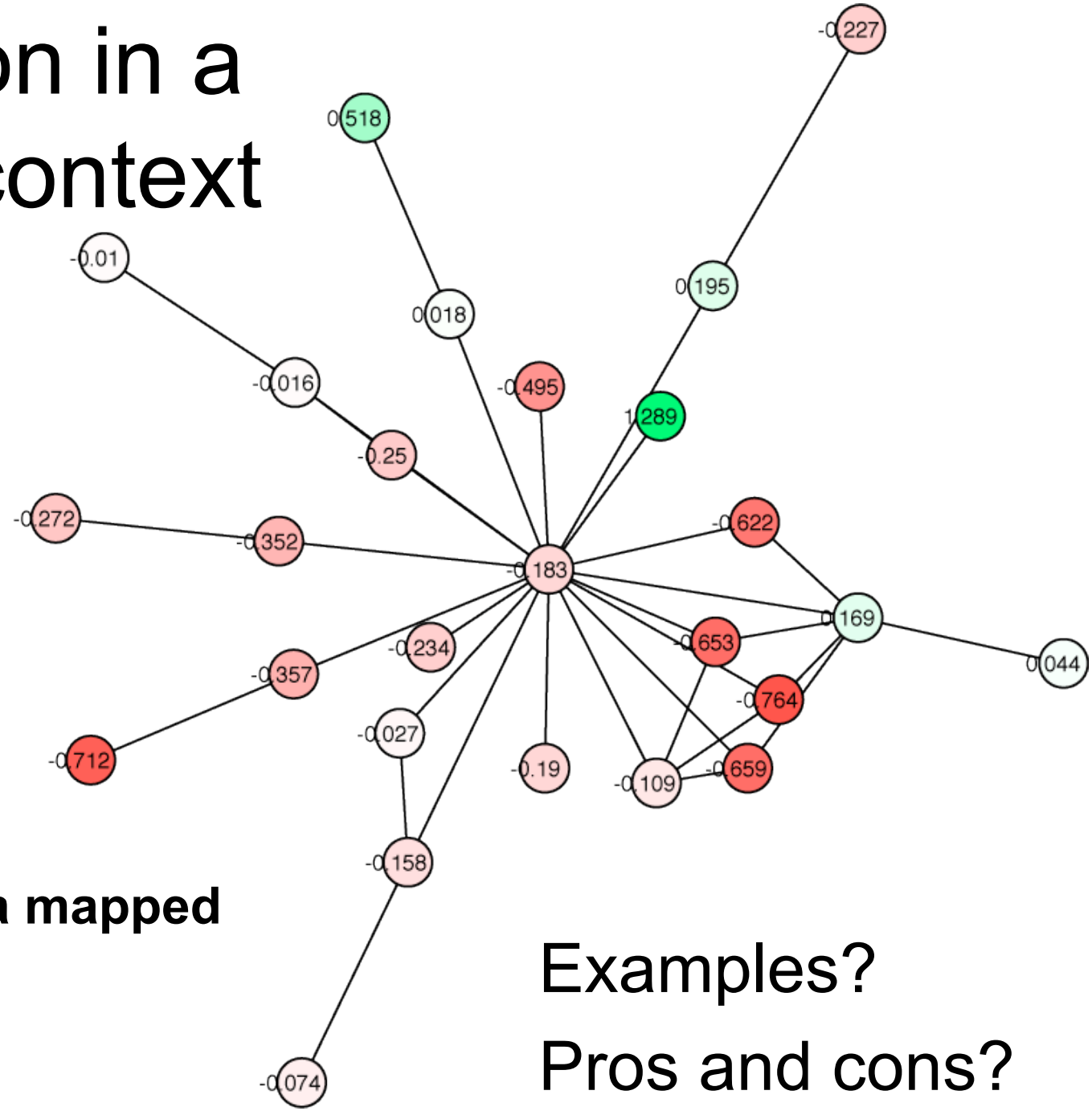
Cellular Context



Integration in a network context



Integration in a network context



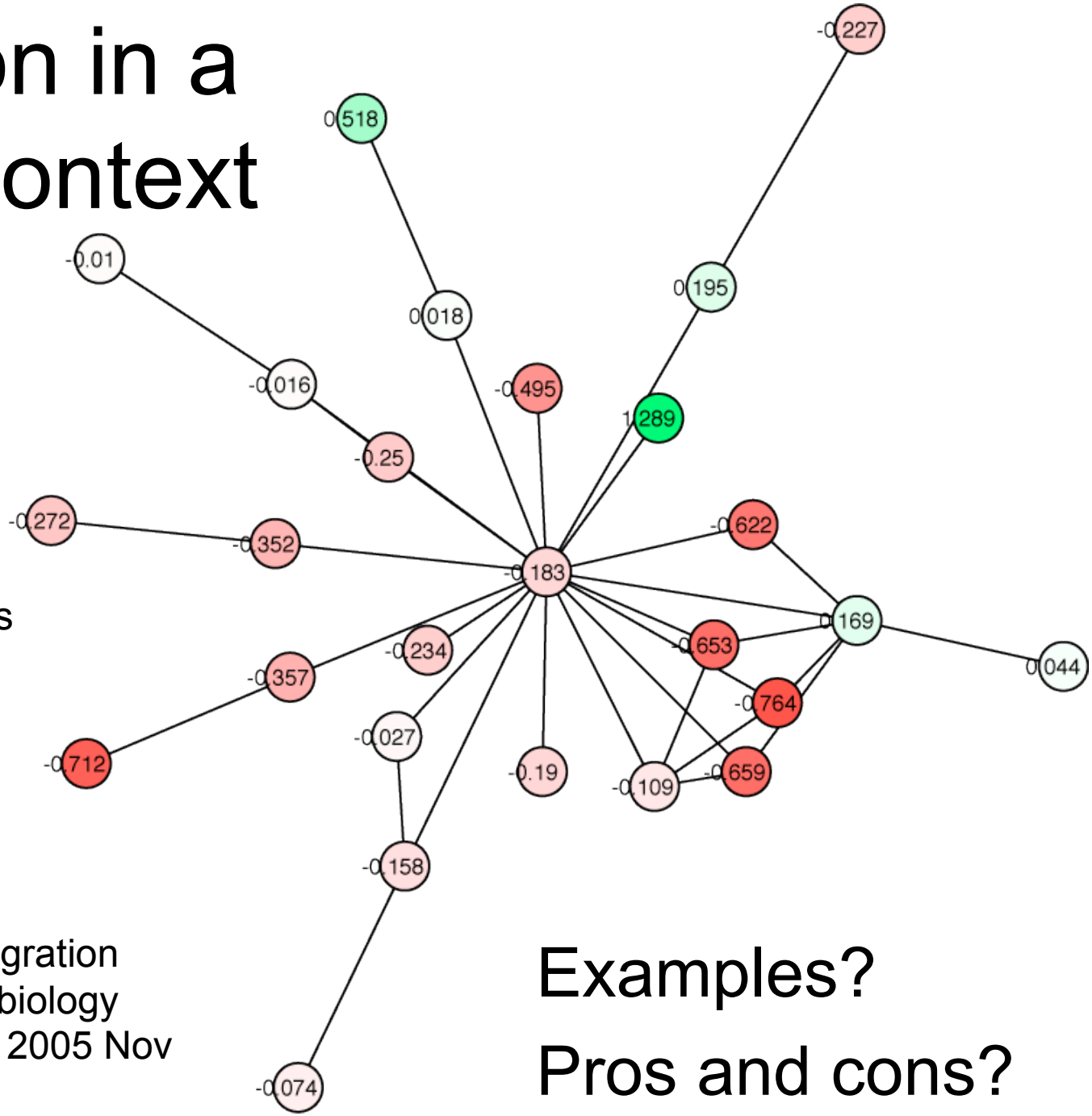
Expression data mapped to node colours

Examples?

Pros and cons?

Integration in a network context

Advantages:
-Interpretable
-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?

Data Integration

Network Analysis

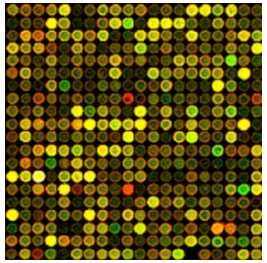
Why Network Analysis?

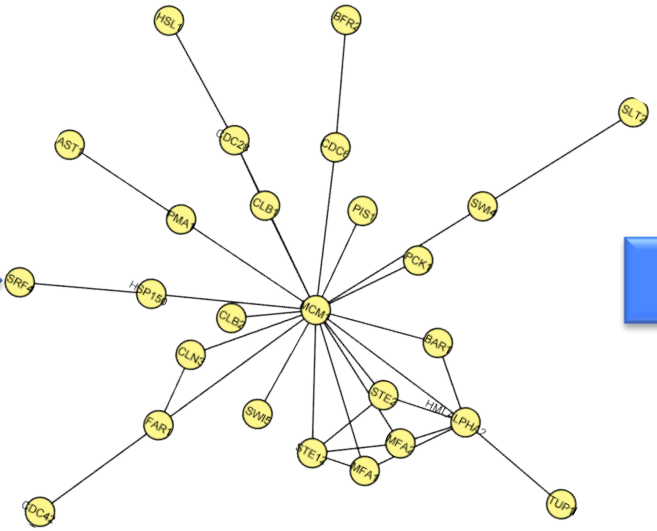
Intuitive to Biologists

- Provide a biological context for results
- More efficient than searching databases gene-by-gene
- Intuitive display for sharing data

Computationally Query to Answer Specific Questions

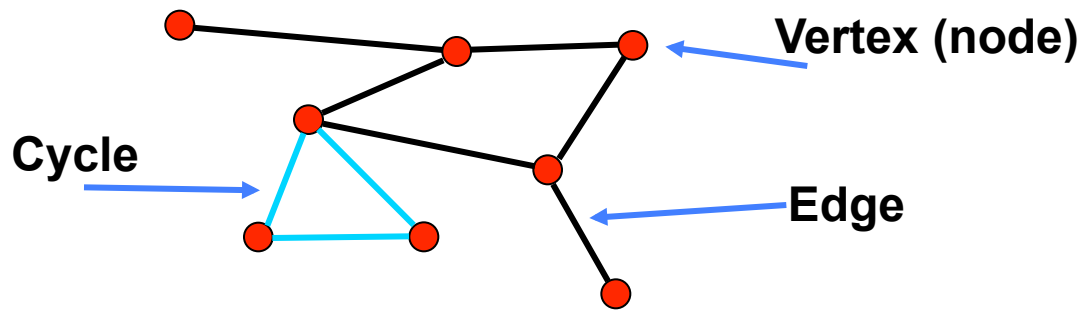
- Visualize multiple data types on a network
- Cluster, Find active pathways, Compare, Search

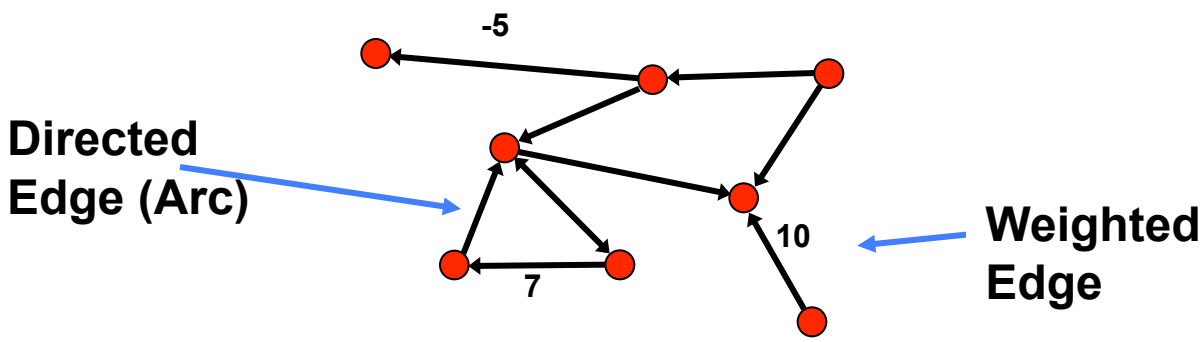




Eureka! New pathway gene!

Graph Theory

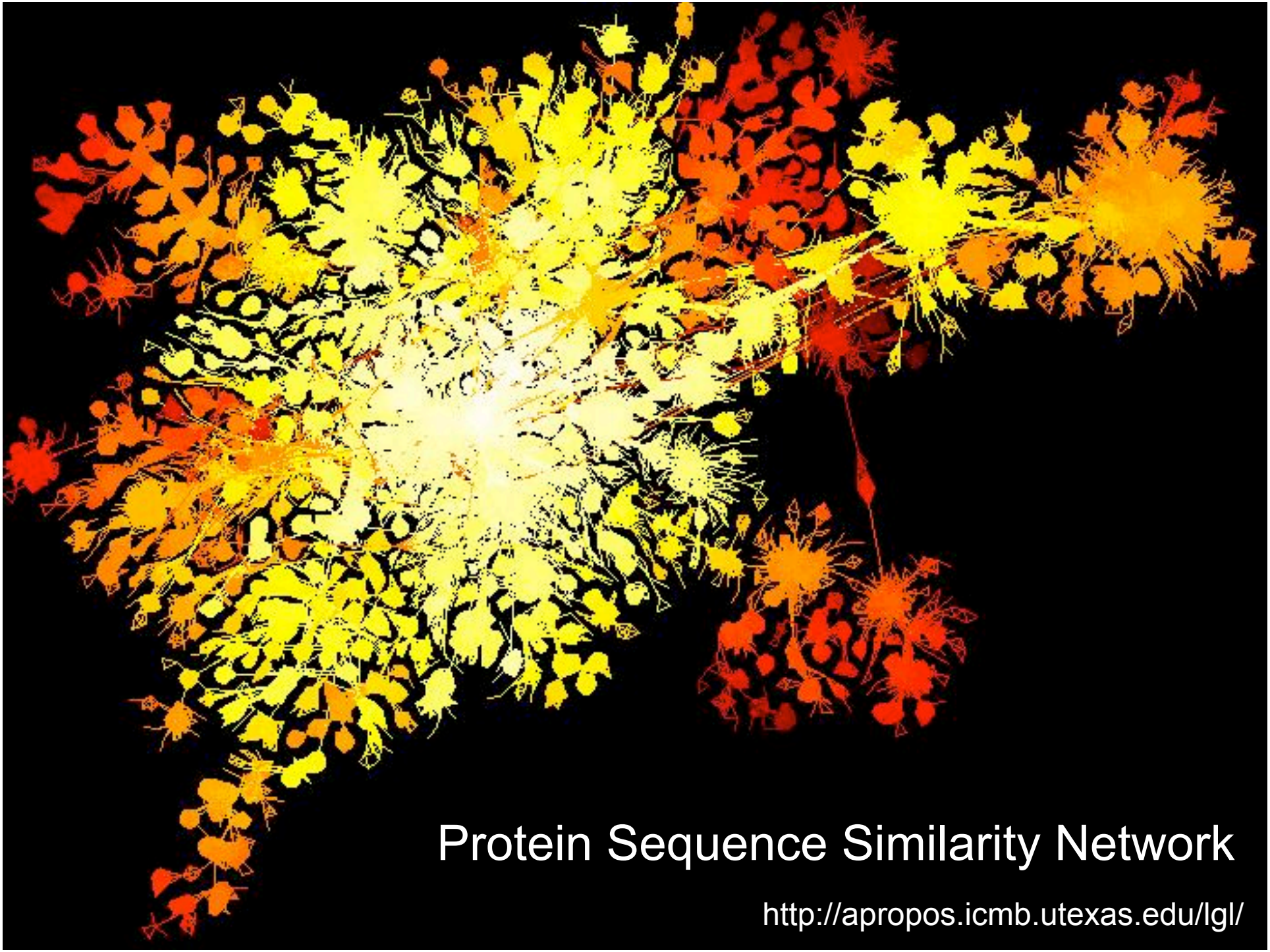




We map molecular interaction networks to graphs

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

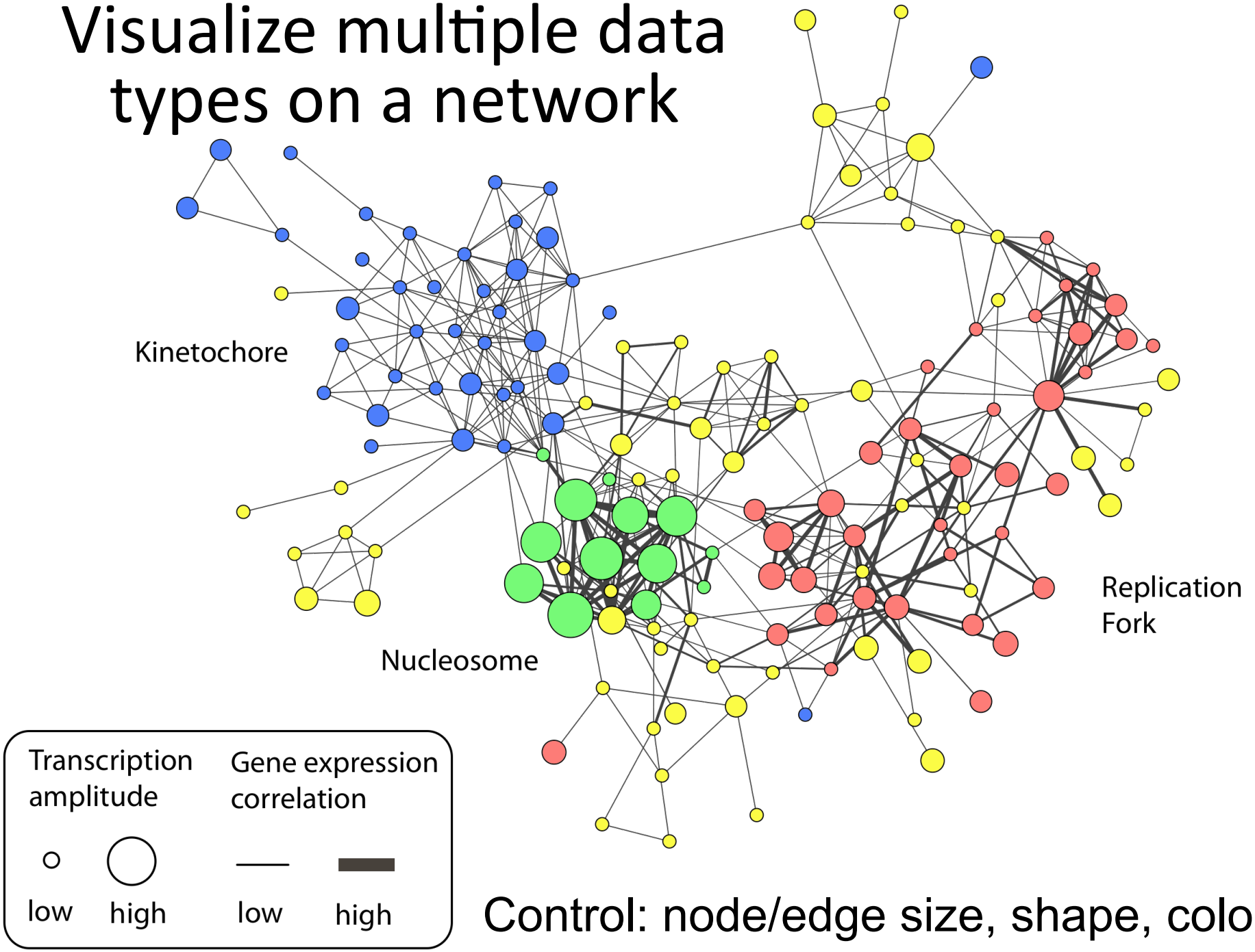

Protein Sequence Similarity Network
http://apropos.icmb.utexas.edu/lgl/

Six Degrees of Separation

- Everyone in the world is connected by at most six links
- Which path should we take?
- Shortest path by breadth first search
  - If two nodes are connected, will find the shortest path between them
- Are two proteins connected? If so, how?
- Biologically relevant?

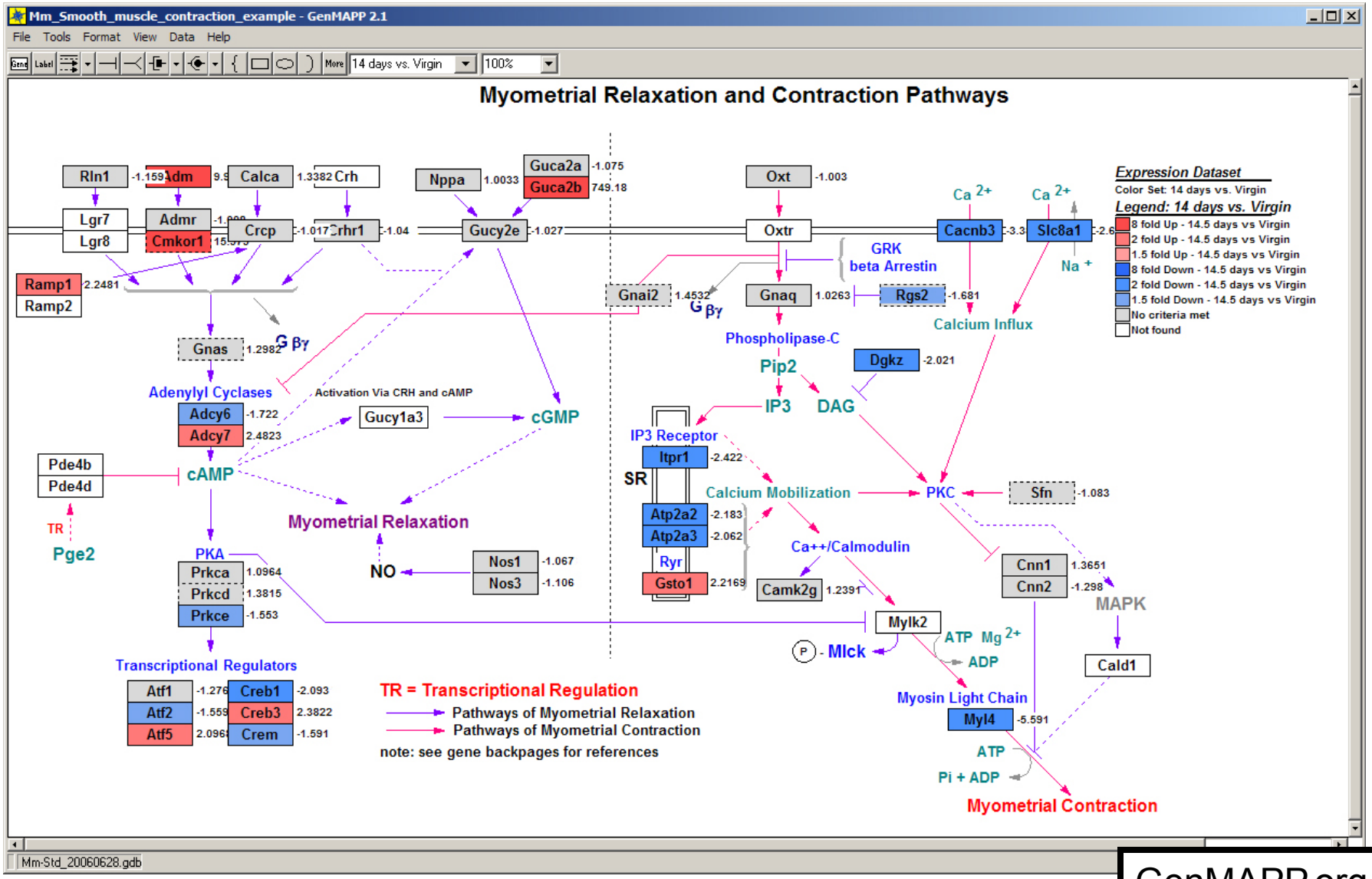
http://www.time.com/time/techtime/200406/community.html

Visualize multiple data types on a network



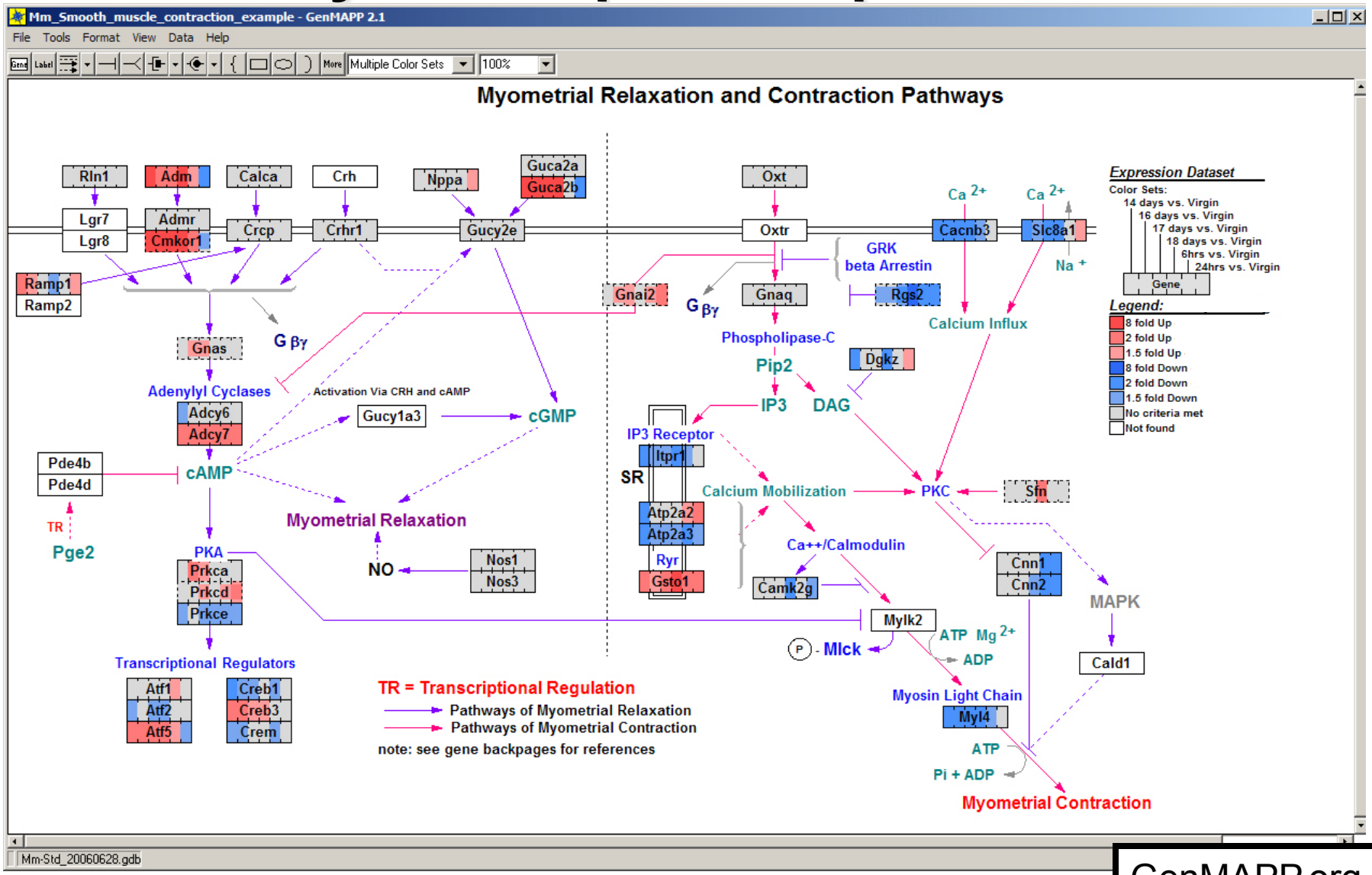
Control: node/edge size, shape, color...

Visualizing Time Course Data on Pathways: Single Comparison View



GenMAPP.org

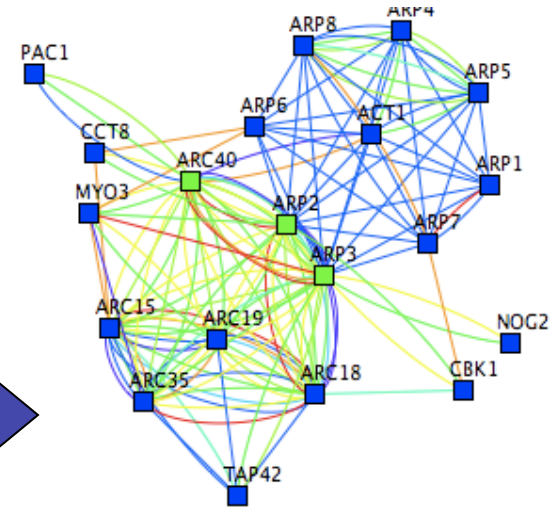
Visualizing Time Course Data on Pathways: Multiple Comparison View



GenMAPP.org

Predicting Gene Function

arp2
arp3
arc40



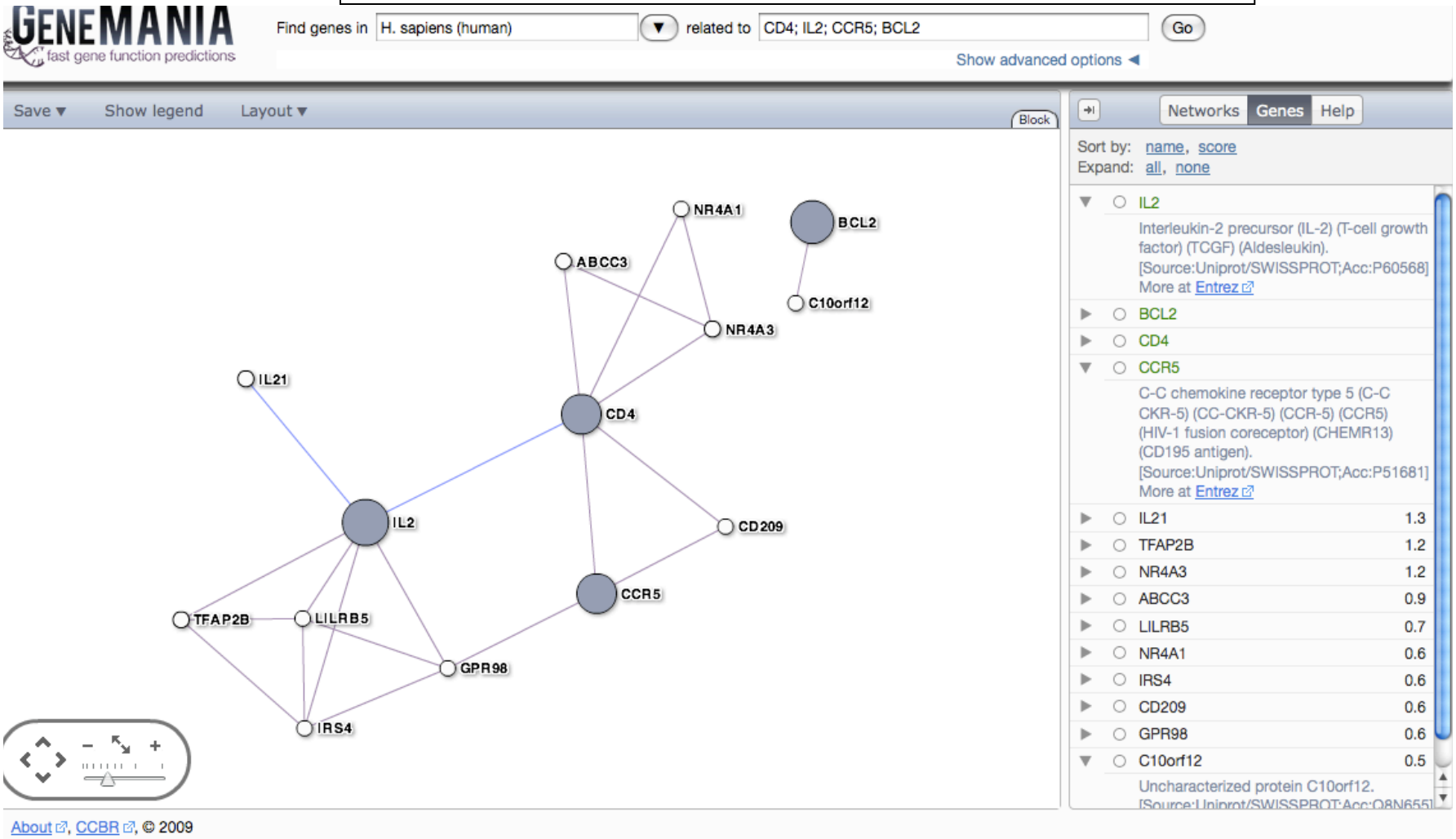
- STRING
  - http://string.embl.de/
- bioPIXIE
  - http://pixie.princeton.edu/pixie/
- GeneMania
  - http://www.genemania.org

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

http://www.genemania.org



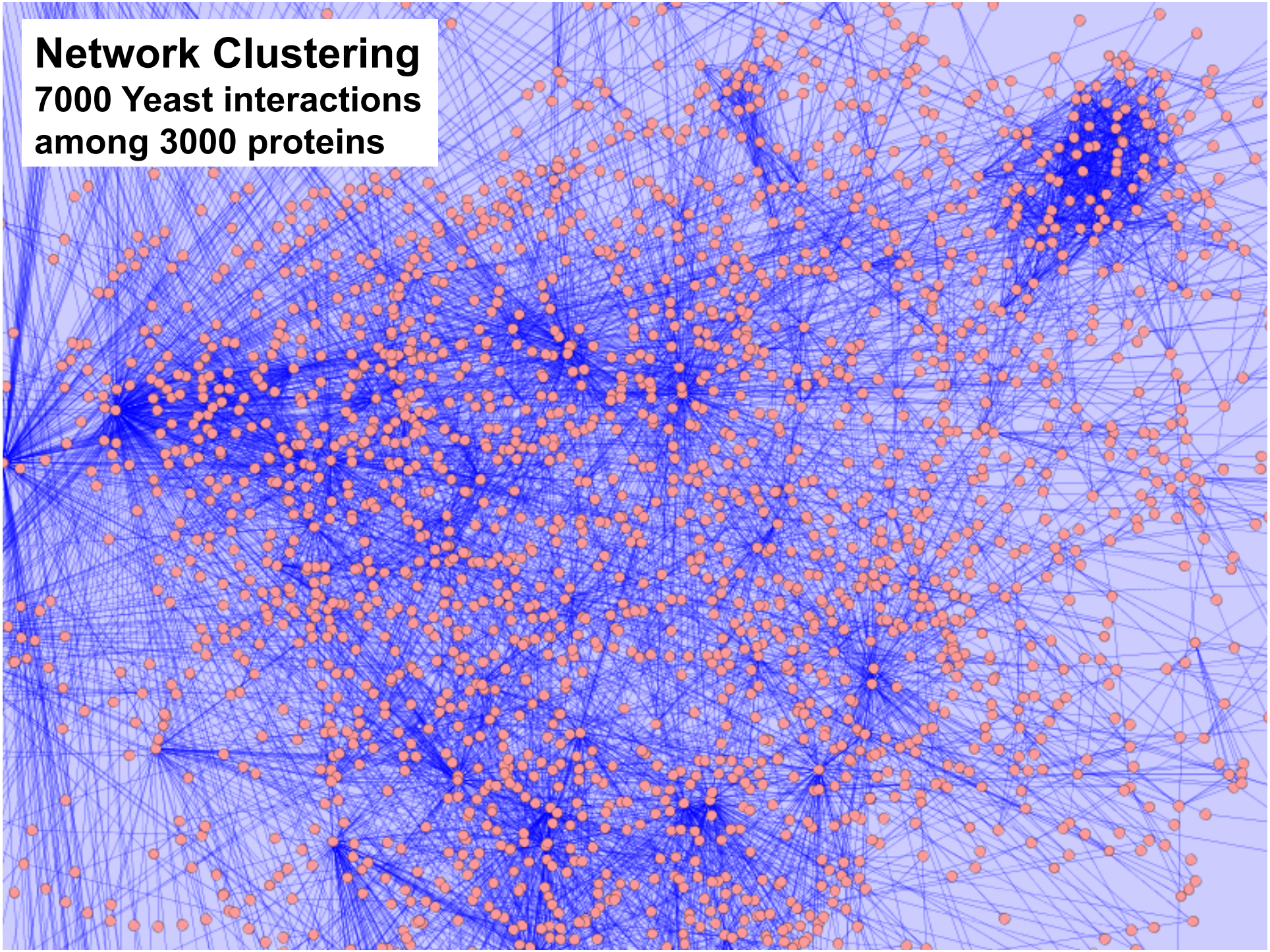
- Guilt-by-association principle
- Biological networks are combined intelligently to optimize prediction accuracy
- Algorithm is more fast and accurate than its peers

Gene Function Prediction

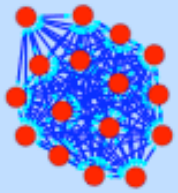
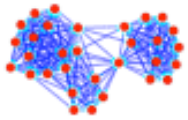
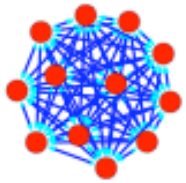
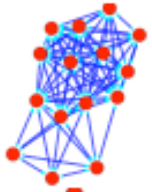
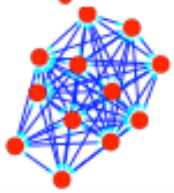
Quaid Morris (CCBR)
Rashad Badrawi, Ovi Comes, Sylva Donaldson, Christian Lopes, Farzana Kazi, Jason Montojo, Harold Rodriguez, Khalid Zuberi

Graph Clustering - MCODE Plugin

- Clusters in a protein-protein interaction network have been shown to represent protein complexes and parts of pathways
- Clusters in a protein similarity network represent protein families
- Network clustering is available through the MCODE Cytoscape plugin


Network Clustering
7000 Yeast interactions
among 3000 proteins

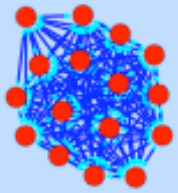
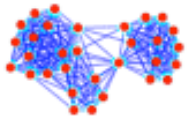
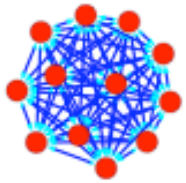
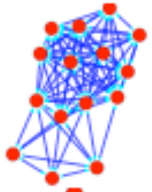
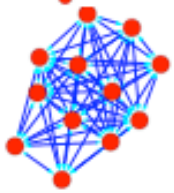
MCODE Results Summary

Rank	Score	Size	Names	Complex
1	7.25	16,116	YGR232W, YDL007W, YKL145W, YFR052W, YFR004W, YLR421C, YOR261C, YDL147W, YDR427W, YHR200W, YER021W, YOR117W, YDL097C, YOR259C, YPR108W, YDR394W	
2	6.387	31,198	YPL093W, YBL004W, YOR272W, YNL110C, YKL009W, YFL002C, YOL077C, YPL126W, YIL035C, YLR409C, YLR129W, YOR061W, YKR060W, YCR057C, YDR449C, YOR039W, YJL109C, YPL012W, YGR103W, YLR449W, YOR206W, YKL014C, YLL008W, YKL172W, YNL002C, YLR002C, YGL111W, YOL041C, YGL019W, YOR145C, YPR016C	
3	5.417	12,65	YGL011C, YOL038W, YPR103W, YMR314W, YBL041W, YOR362C, YER012W, YJL001W, YML092C, YGR253C, YER094C, YGR135W	
4	5	15,75	YPL043W, YMR290C, YER006W, YKR081C, YDR496C, YDL031W, YNL061W, YNL132W, YLR222C, YLR197W, YMR049C, YHR052W, YJL069C, YKL099C, YDL014W	
5	5	12,60	YPR187W, YPR010C, YPR110C, YNL248C, YOR341W, YNR003C, YKL144C, YOR207C, YPR190C, YNL113W, YOR116C, YBR154C	

Create a new child network. Save Done

Bader & Hogue, BMC Bioinformatics 2003 4(1):2

MCODE Results Summary

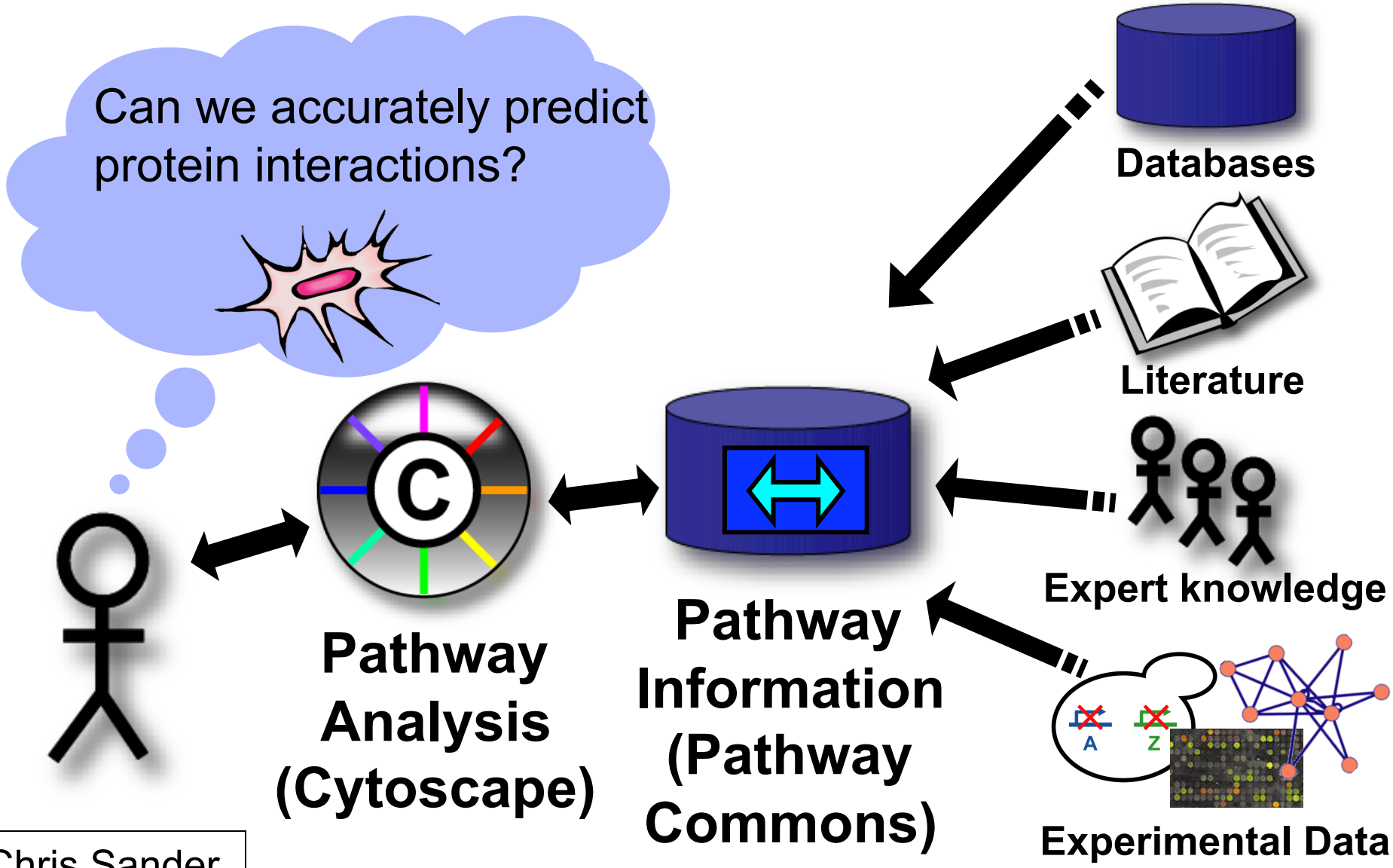
Rank	Score	Size	Names	Complex
1 Proteasome 26S	7.25	16,116	YGR232W, YDL007W, YKL145W, YFR052W, YFR004W, YLR421C, YOR261C, YDL147W, YDR427W, YHR200W, YER021W, YOR117W, YDL097C, YOR259C, YPR108W, YDR394W	
2 Ribosome	6.387	31,198	YPL093W, YBL004W, YOR272W, YNL110C, YKL009W, YFL002C, YOL077C, YPL126W, YIL035C, YLR409C, YLR129W, YOR061W, YKR060W, YCR057C, YDR449C, YOR039W, YJL109C, YPL012W, YGR103W, YLR449W, YOR206W, YKL014C, YLL008W, YKL172W, YNL002C, YLR002C, YGL111W, YOL041C, YGL019W, YOR145C, YPR016C	
3 Proteasome 20S	5.417	12,65	YGL011C, YOL038W, YPR103W, YMR314W, YBL041W, YOR362C, YER012W, YJL001W, YML092C, YGR253C, YER094C, YGR135W	
4 RNA Splicing	5	15,75	YPL043W, YMR290C, YER006W, YKR081C, YDR496C, YDL031W, YNL061W, YNL132W, YLR222C, YLR197W, YMR049C, YHR052W, YJL069C, YKL099C, YDL014W	
5 RNA Pol core	5	12,60	YPR187W, YPR010C, YPR110C, YNL248C, YOR341W, YNR003C, YKL144C, YOR207C, YPR190C, YNL113W, YOR116C, YBR154C	

Create a new child network. Save Done

Bader & Hogue, BMC Bioinformatics 2003 4(1):2

Network Data

Cell map exploration and analysis



Chris Sander

http://pathguide.org

Pathguide » the pathway resource list

Home | BioPAX | cBio | MSKCC

>300 Pathway Databases!

Navigation
- Protein-Protein Interactions
- Metabolic Pathways
- Signaling Pathways
- Pathway Diagrams
- Transcription Factors / Gene Regulatory Networks
- Protein-Compound Interactions
- Genetic Interaction Networks
- Protein Sequence Focused
- Other

Search

Organisms: All

Availability: All

Standards: All

Reset | Search

Statistics

Analyze Pathguide

Contact

Comments, Questions, Suggestions are Always Welcome!

Complete Listing of All Pathguide Resources

Pathguide contains information about **222** biological pathway resources. Click on a link to go to the resource home page or 'Details' for a description page. Databases that are free and those supporting BioPAX, CellML, PSI-M... or SBML standards are respectively indicated.

If you know of a pathway resource that is not listed here, or have other questions or comments, please send us an e-mail.

Get the Stats
Detailed Pathguide resource statistics now available

Pathguide Published
Please cite the Pathguide

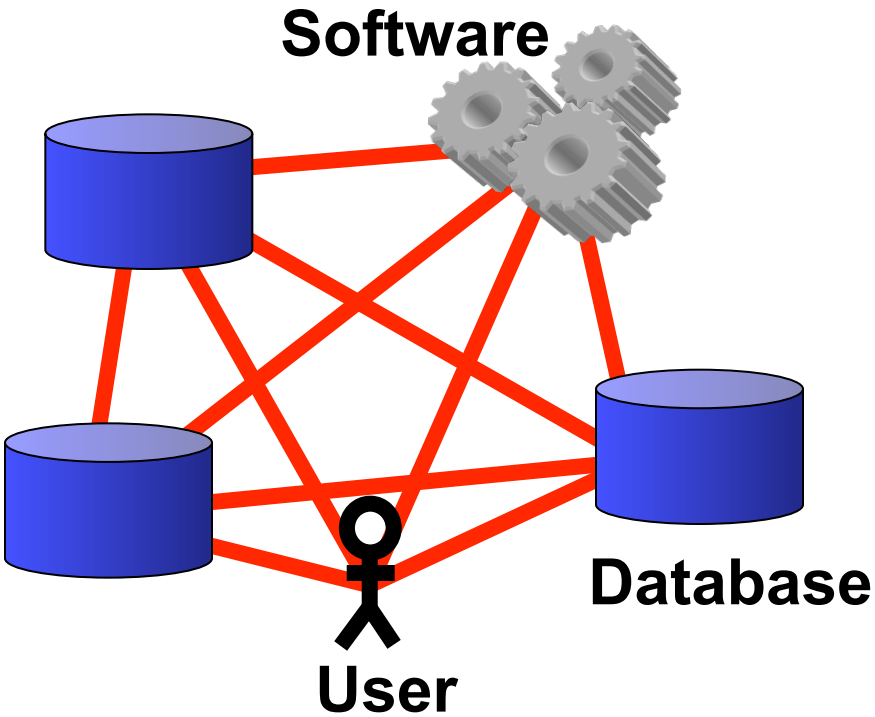
Protein-Protein Interactions

Database Name (Order: alphabetically | by web popularity)	Full Record	Availability	Standards
3DID - 3D interacting domains	Details	Free	
ABCdb - Archaea and Bacteria ABC transporter database	Details	Free	
AfCS - Alliance for Cellular Signaling Molecule Pages Database	Details	Free	
AllFuse - Functional Associations of Proteins in Complete Genomes	Details	Free	
ASEdb - Alanine Scanning Energetics Database	Details	Free	
ASPD - Artificial Selected Proteins/Peptides Database	Details	?	
BID - Binding Interface Database	Details	Free	
BIND - Biomolecular Interaction Network Database	Details	Free	PSI-MI
BindingDB - The Binding Database	Details	Free	
BioGRID - General Repository for Interaction Datasets	Details		PSI-MI
BRITE - Biomolecular Relations in Information Transmission and Expression	Details	Free	
CA1Neuron - Pathways of the hippocampal CA1 neuron	Details	Free	
Cancer Cell Map - The Cancer Cell Map	Details	Free	BioPAX
CSP - Cytokine Signaling Pathway Database	Details	Free	
CTDB - Calmodulin Target Database	Details	Free	
DDIB - Database of Domain Interactions and Bindings	Details	Free	
DIP - Database of Interacting Proteins	Details		PSI-MI
Doodle - Database of oligomeri...			
DopaNet - DopaNet			
DRC - Database of Ribosomal (			
DSM - Dynamic Signaling Maps			
FIMM - Functional Molecular Im			
FusionDB - Prokaryote Gene Fu			

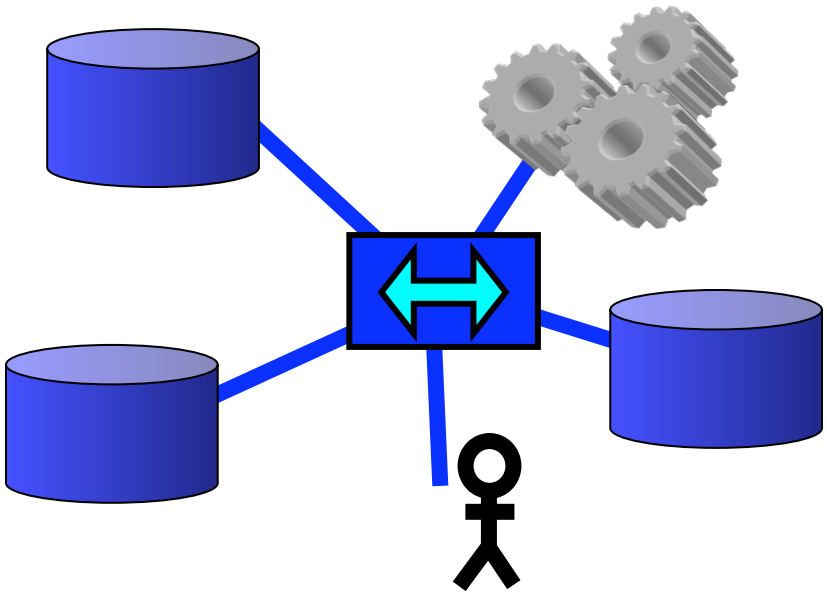
- **Varied formats, representation, coverage**
- **Pathway data extremely difficult to combine and use**

Vuk Pavlovic

Solution: Standard Exchange Formats



>100 DBs and tools
Tower of Babel



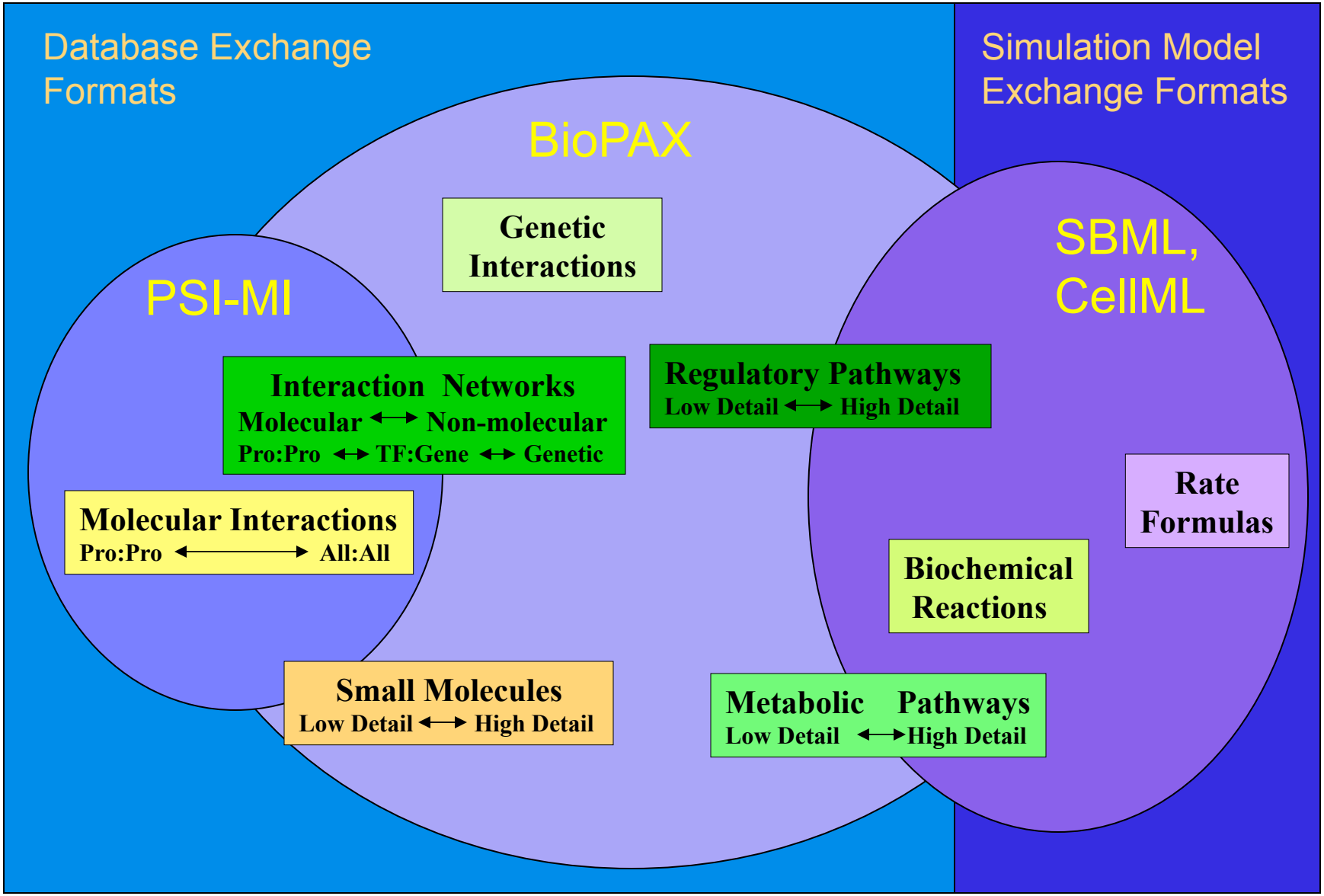
With Data
Exchange Format

Reduces work, promotes collaboration, increases accessibility

Interaction and Pathway Data Exchange Formats

- PSI-MI http://psidev.sourceforge.net
  - Molecular interactions - protein-protein interaction focus
  - Peer reviewed, HUPO community standard
- BioPAX http://www.biopax.org
  - Biological pathways
  - Community ontology in OWL, Protégé
- SBML http://www.sbml.org
  - Widely adopted for representing mathematical models of biological processes e.g. biochemical reaction networks
- CellML http://www.cellml.org
  - Math models of biological processes

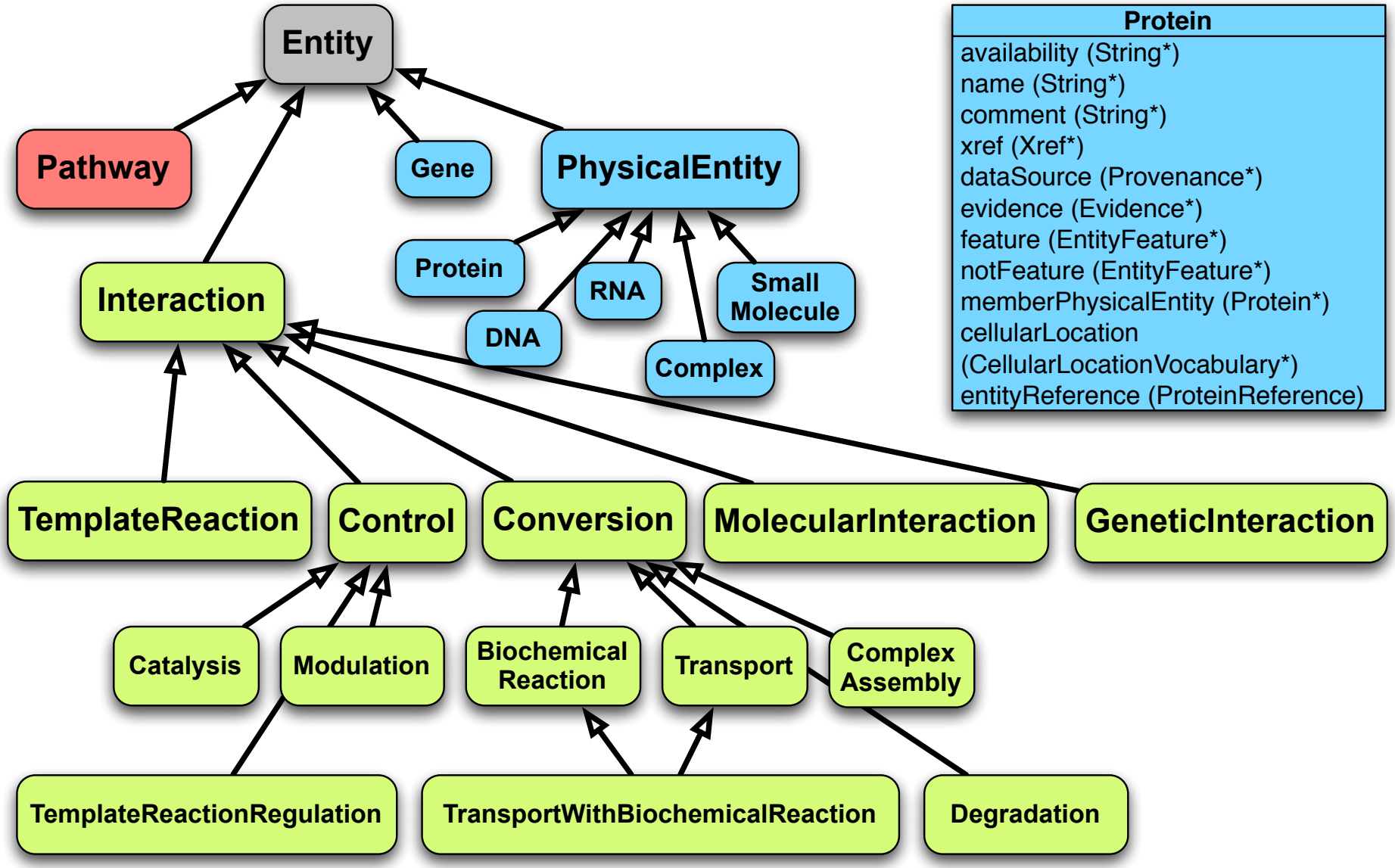
Biological Network Exchange Formats

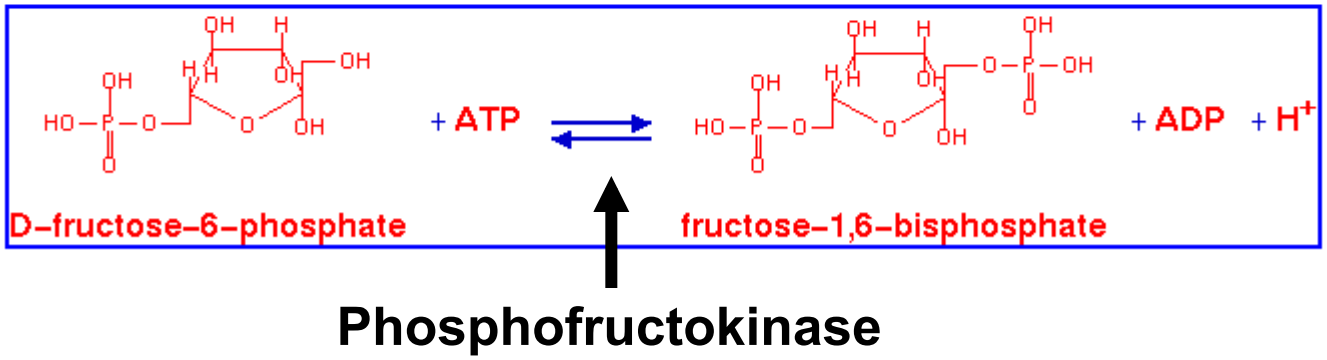


BioPAX Pathway Language

- Represent:
  - Metabolic pathways
  - Signaling pathways
  - Protein-protein, molecular interactions
  - Gene regulatory pathways
  - Genetic interactions
- Community effort: pathway databases distribute pathway information in standard format
  - Over 100 people, database groups, standard efforts

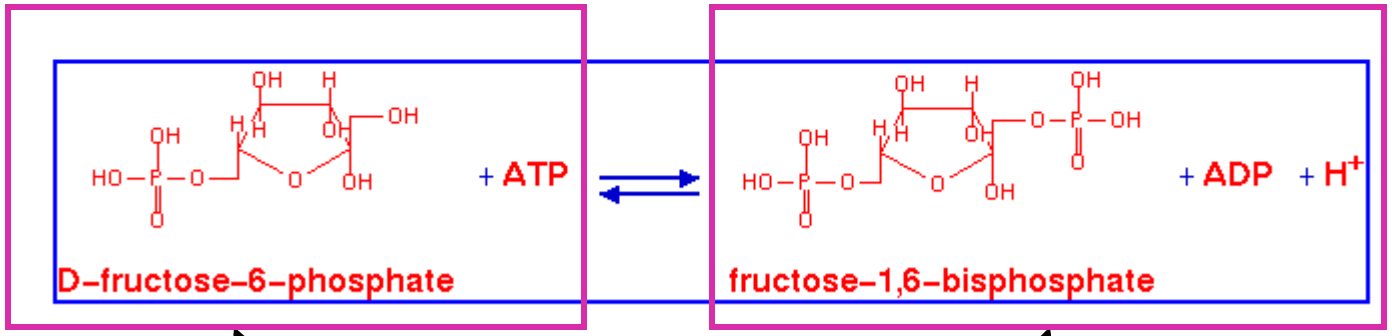
BioPAX Class Structure





Biochemical Reaction
Glycolysis Pathway

Source: BioCyc.org

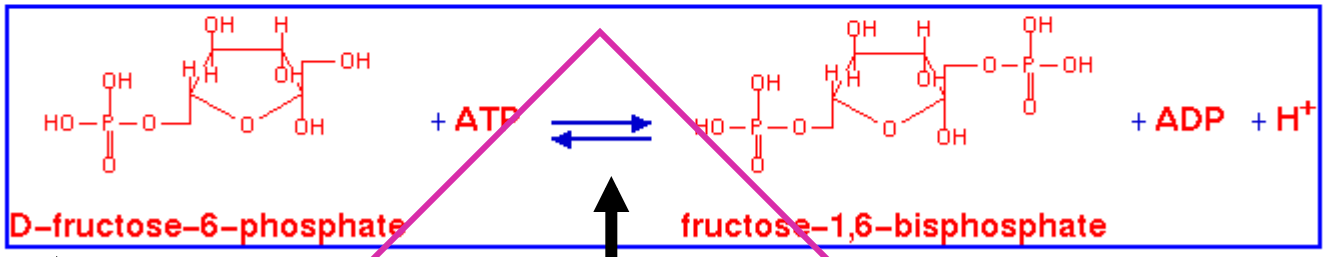


Left

Right

biochemicalReaction
PARTICIPANTS
RIGHT
SYNONYMS
SHORT-NAME
SPONTANEOUS
COMMENT
AVAILABILITY
LEFT
DATA-SOURCE
NAME
XREF
DELTA-H
DELTA-S
EC-NUMBER
KEQ
DELTA-G

EC # 2.7.1.11



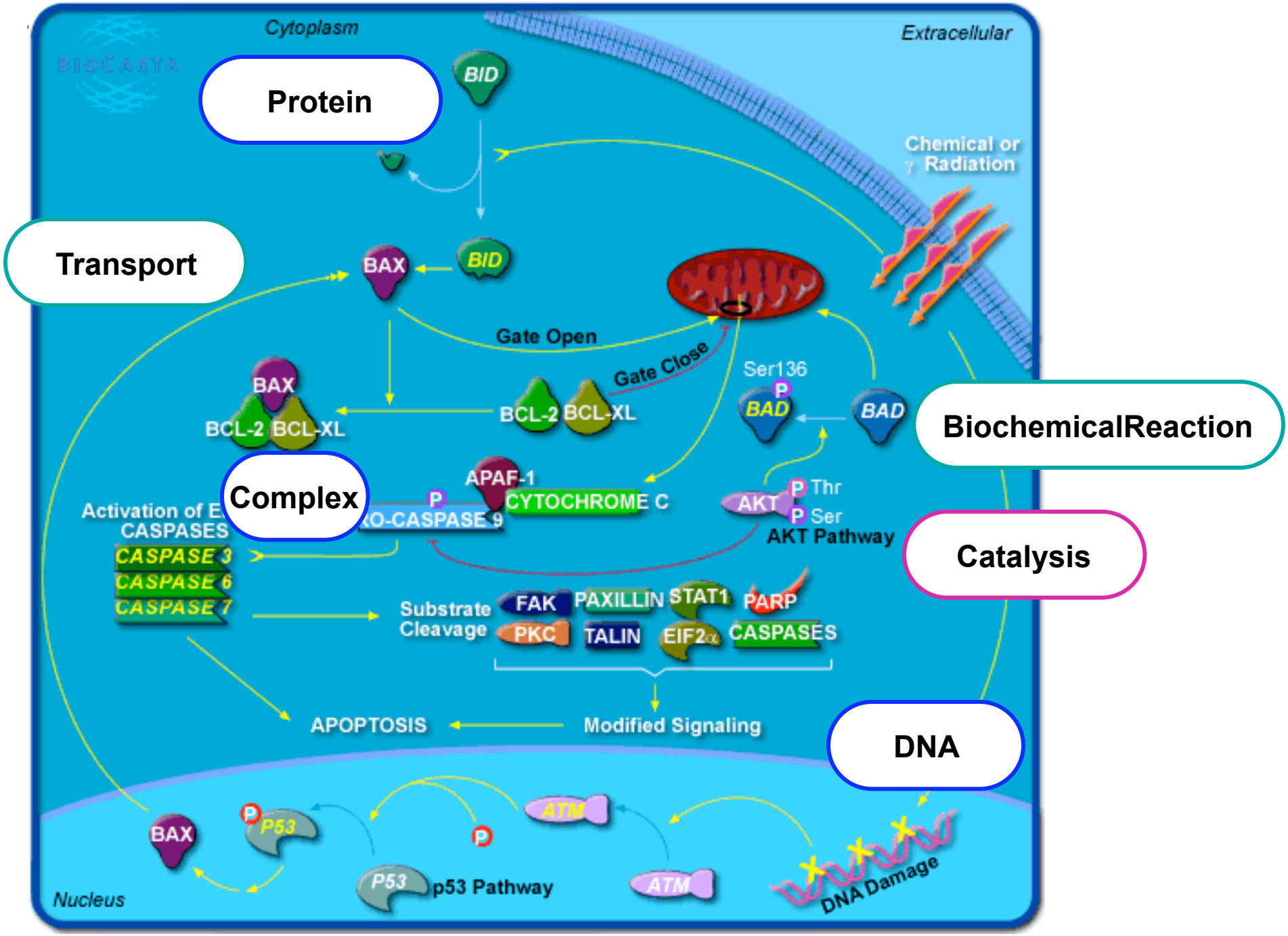
Phosphofructokinase

Controller

Controlled

catalysis	
	CONTROLLED
	COMMENT
	PARTICIPANTS
	AVAILABILITY
	CONTROL-TYPE
	DATA-SOURCE
	CONTROLLER
	SYNONYMS
	SHORT-NAME
	NAME
	XREF
	DIRECTION
	COFACTOR

Direction: reversible


Protein
Transport
BiochemicalReaction
Complex
Catalysis
DNA
Cytoplasm
Extracellular
BIOCARTA
BID
BAX
Chemical or
γ Radiation
Gate Open
Gate Close
BCL-2
BCL-XL
Ser136
P
BAD
APAF-1
CYTOCHROME C
Activation of E
CASPASES
XO-CASPASE 9
AKT
Thr
Ser
AKT Pathway
CASPASE 3
CASPASE 6
CASPASE 7
Substrate
Cleavage
FAK
PAXILLIN
STAT1
PARP
PKC
TALIN
EIF2α
CASPASES
APOPTOSIS
Modified Signaling
ATM
P53
p53 Pathway
DNA Damage
Nucleus

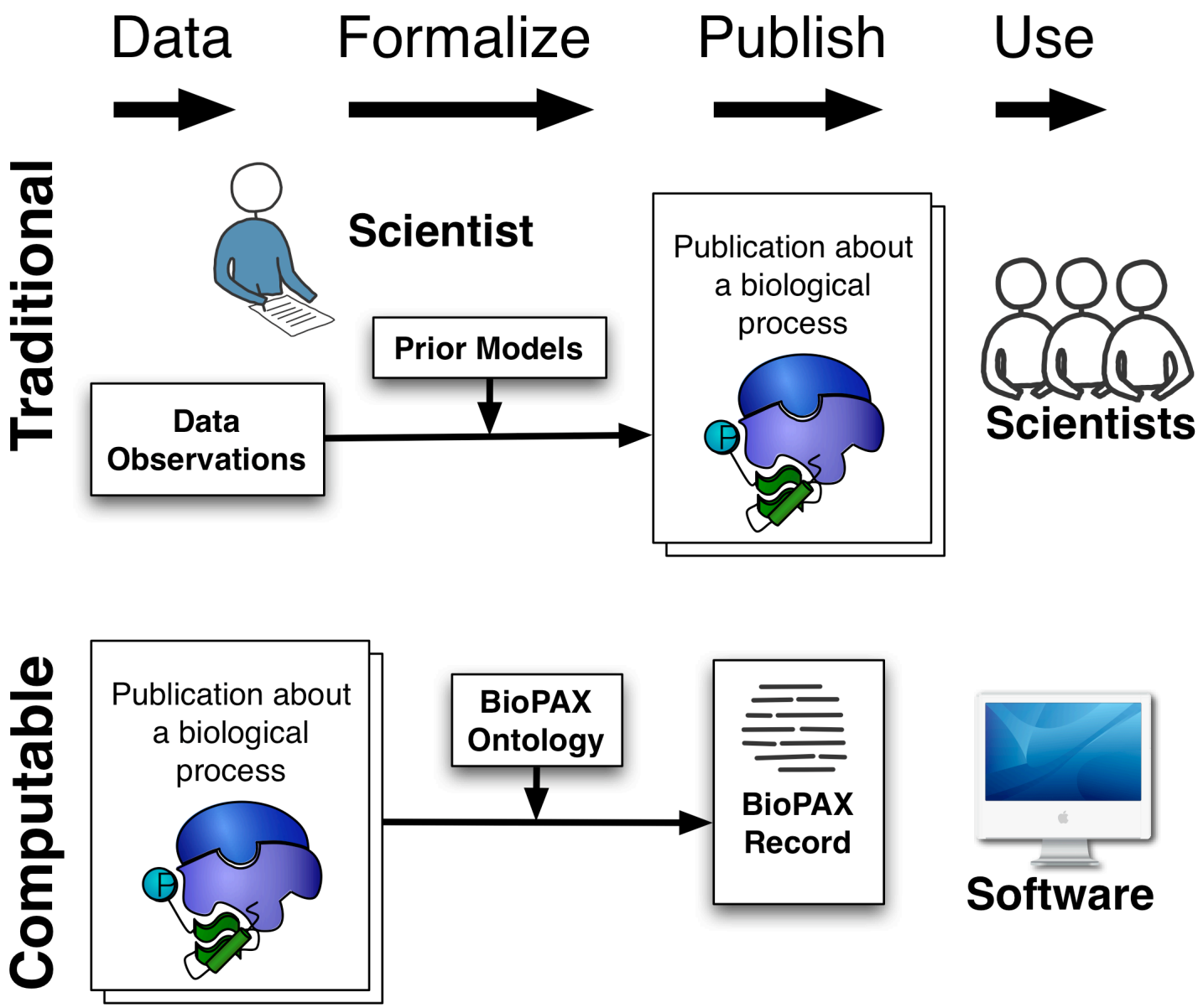
Controlled Vocabularies (CVs)

- BioPAX uses existing CVs where available via openControlledVocabulary instances
  - Cellular location: Gene Ontology (GO) component
  - PSI-MI CVs for:
    - Protein post-translational modifications
    - Interaction detection experimental methods
    - Experimental form
  - PATO phenotypic quality ontology
  - Some database providers use their own CVs
    - E.g. BioCyc evidence codes
- More at the Ontology Lookup Service
  - http://www.ebi.ac.uk/ontology-lookup/

OWL (XML) Snippet

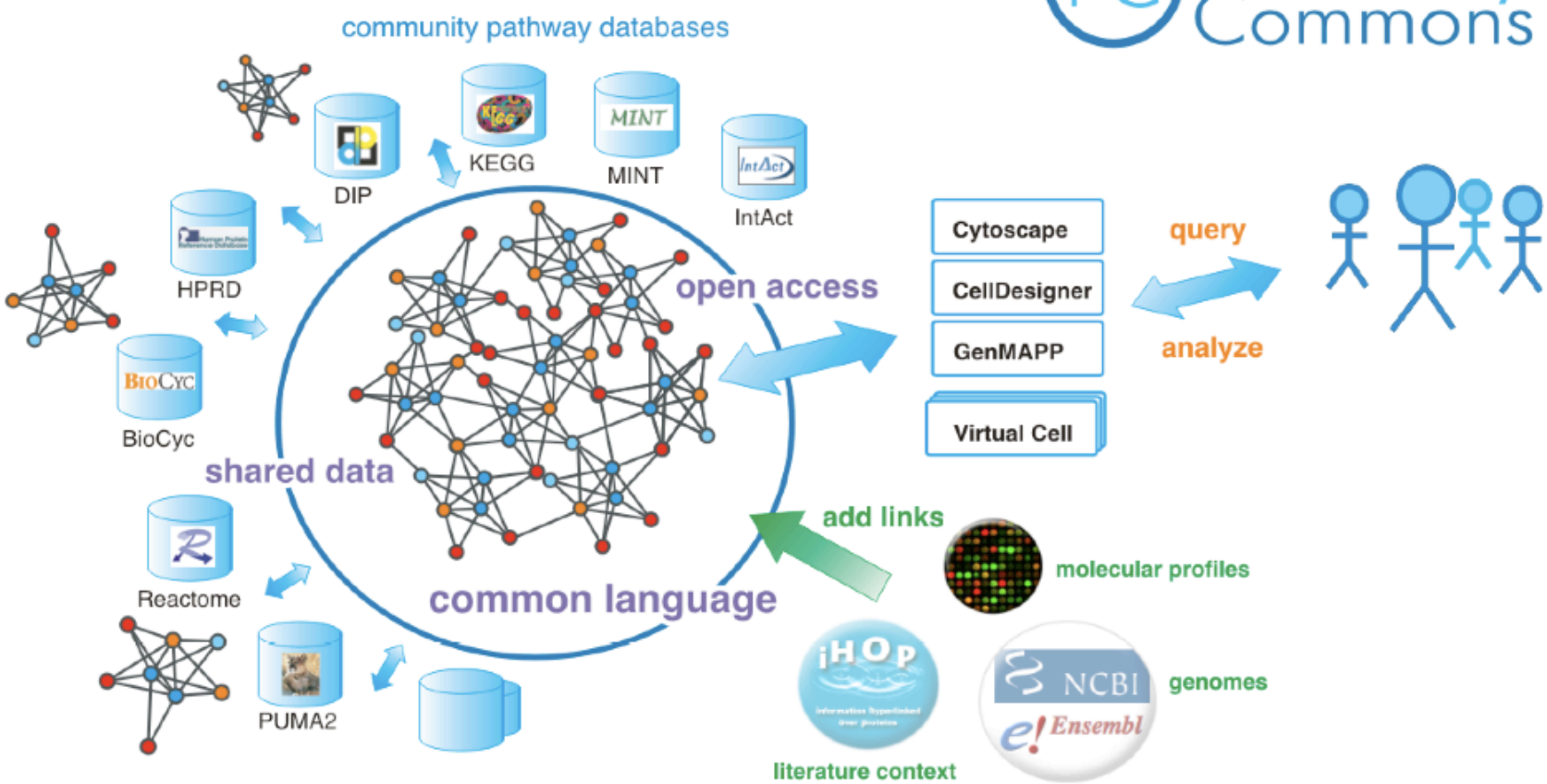

```
<bp:biochemicalReaction rdf:ID="biochemicalReaction37">
  <bp:DATA-SOURCE rdf:resource="#dataSource14"/>
  <bp:LEFT>
    <bp:physicalEntityParticipant rdf:ID="physicalEntityParticipant26">
      <bp:STOICHIOMETRIC-COEFFICIENT>1.0</bp:STOICHIOMETRIC-COEFFICIENT>
      <bp:PHYSICAL-ENTITY>
        <bp:smallMolecule rdf:ID="smallMolecule27">
          <bp:SHORT-NAME rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
          >a-D-glu-6-p</bp:SHORT-NAME>
          <bp:CHEMICAL-FORMULA rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
          >C6H13O9P</bp:CHEMICAL-FORMULA>
          <bp:SYNONYMS rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
          >&lt;FONT FACE="Symbol">a&lt;/FONT>-D-glucose-6-phoshate</bp:SYNONYMS>
          <bp:XREF>
            <bp:unificationXref rdf:ID="unificationXref30">
              <bp:ID rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
              >C00668</bp:ID>
              <bp:DB rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
              >KEGG</bp:DB>
            </bp:unificationXref>
          </bp:XREF>
          <bp:XREF rdf:resource="#unificationXref29"/>
          <bp:MOLECULAR-WEIGHT>260.14</bp:MOLECULAR-WEIGHT>
          <bp:AVAILABILITY rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
          >see http://www.amaze.ulb.ac.be/</bp:AVAILABILITY>
          <bp:SYNONYMS rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
          >glucose-6-P</bp:SYNONYMS>
          <bp:DATA-SOURCE rdf:resource="#dataSource14"/>
          <bp:SYNONYMS rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
          >alpha-D-glucose-6-p</bp:SYNONYMS>
          <bp:STRUCTURE>
            <bp:chemicalStructure rdf:ID="chemicalStructure28">
              <bp:STRUCTURE-FORMAT>SMILES</bp:STRUCTURE-FORMAT>
              <bp:STRUCTURE-DATA>C(OP(=O)(O)O)[CH]1([CH](O)[CH](O)[CH](O)[CH](O)O1)</bp:STRUCTURE-DATA>
            </bp:chemicalStructure>
          </bp:STRUCTURE>
          <bp:NAME>alpha-D-glucose 6-phosphate</bp:NAME>
          <bp:SYNONYMS rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
          >alpha-D-glucose-6-phosphate</bp:SYNONYMS>
          <bp:SYNONYMS rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
          > D-glucose-6-P</bp:SYNONYMS>
          <bp:DATA-SOURCE rdf:resource="#KB_439584_Individual_47"/>
        </bp:smallMolecule>
      </bp:PHYSICAL-ENTITY>
      <bp:CELLULAR-LOCATION rdf:resource="#openControlledVocabulary15"/>
    </bp:physicalEntityParticipant>
  </bp:LEFT>
  <bp:DELTA-G rdf:datatype="http://www.w3.org/2001/XMLSchema#double"
  >0.4</bp:DELTA-G>
  <bp:SYNONYMS rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  >alpha-D-Glucose 6-phosphate &lt;=> beta-D-Fructose 6-phosphate </bp:SYNONYMS>
  <bp:RIGHT>
    <bp:physicalEntityParticipant rdf:ID="physicalEntityParticipant38">
      <bp:CELLULAR-LOCATION rdf:resource="#openControlledVocabulary15"/>
      <bp:PHYSICAL-ENTITY>
        <bp:smallMolecule rdf:ID="smallMolecule39">
```

Pathway Information Processing



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

Pathway Commons

Search

Home | Filter | FAQ | About | Credits

Pathway Commons is a convenient point of access to biological pathway information collected from public pathway databases, which you can browse or search. Pathways include biochemical reactions, complex assembly, transport and catalysis events, and physical interactions involving proteins, DNA, RNA, small molecules and complexes. more...

Search Pathway Commons:

Search

To get started, enter a gene name, gene identifier or pathway name in the text box above. For example: p53, P38398 or mTOR.

To restrict your search to specific data sources or specific organisms, update your global filter settings.

Pathway Commons currently contains the following data sources:

Cancer Cell Map, Release: 1.0 [19-May-06]
Browse

HumanCyc, Release: 10.5 [18-Sep-06]
Browse

NCI / Nature Pathway Interaction Database
[01-Jan-07]
Browse

Reactome, Release: 19 [16-Nov-06]
Browse

Pathway Commons Quick Stats:

Number of Pathways:	921
Number of Interactions:	9,924
Number of Physical Entities:	15,515
Number of Organisms:	10

Biologists: Browse and search pathways across multiple valuable public pathway databases.

Computational biologists: Download an integrated set of pathways in BioPAX format for global analysis.

Software developers: Build software on top of Pathway Commons using our soon-to-be released web service API. Download and install the cPath software to create a local mirror.

Pathway Commons is hosted by the Computational Biology Center at Memorial Sloan-Kettering Cancer Center and the University of Toronto.

Pathway Commons

p53 Search

Pathway Commons is a work in progress. We welcome your feedback. Email us at: pc-info@pathwaycommons.org.

Home | Filter | FAQ | About | Credits | Results

Searched for: p53

Pathway Commons completed your search for "*p53*" and found **22** relevant records:

Narrow Results by Type:

- All Types (45)
- Pathway (22)
- Protein (23)

Narrow Results by Data Source:

- All Data Sources (22)
- Cancer Cell Map (2)
- NCI / Nature Pathway Interaction Database (3)
- Reactome (17)

[Update Filter Settings]

Showing Results 1 - 10 of 22 | Next 10

Pathway: Transcriptional activation of p53 responsive genes

Summary:

p53 causes G1 arrest by inducing the expression of a cell cycle inhibitor, p21 (El-Deiry et al, 1993; Harper et al, 1993; Xiong et al, 1993). P21 binds and inactivates Cyclin-Cdk complexes that mediate G1/S progression, resulting in lack of phosphorylation of Rb, E2F sequestration and cell cycle arrest at the G1/S transition. Mice with a homozygous deletion of p21 gene are deficient in their ability to undergo a G1/S arrest in response to DNA damage (Deng et al, 1995).

Data Sources:

- Reactome

- ... **p53** causes G1 arrest by inducing the expression of a cell cycle inhibitor, p21 (El-Deiry et al, 1993; Harper et al, 1993; Xiong et al, 1993).

Pathway: Stabilization of p53

- ... ATM also regulates the phosphorylation of **p53** at other sites, especially Ser-20, by activating other serine/threonine kinases in response to IR (Chehab et al, 2000 ...

Pathway: p53-Dependent G1 DNA Damage Response

- Most of the damage-induced modifications of **p53** are dependent on the ATM kinase. ... The first link between ATM and **p53** was predicted based on the earlier studies that showed that AT cells exhibit a reduced and delayed induction of **p53** following exposure to IR (Kastan et al, 1992 and Khanna and Lavin, 1993). ...pUnder normal conditions, **p53** is a short-lived protein ...

Pathway: p53-Dependent G1/S DNA damage checkpoint

- The arrest at G1/S checkpoint is mediated by the action of a widely known tumor suppressor protein, **p53**. ... Loss of **p53** functions, as a result of mutations in cancer prevent the G1/S checkpoint (Kuerbitz et al, 1992). ... **P53** is rapidly induced in response to damaged DNA.

Pathway: p53-Independent G1/S DNA damage checkpoint

- The G1 arrest induced by DNA damage has been ascribed to the transcription factor and tumor suppressor protein **p53**.

Pathway: G1/S DNA Damage Checkpoints

- In the G1 phase there are two types of DNA damage responses, the p53-dependent and the p53-independent pathways. ... The p53-dependent responses inhibit CDKs through the up-regulation of genes encoding CKIs mediated by the **p53** protein, whereas the p53-independent mechanisms inhibit CDKs through the inhibitory T14Y15 phosphorylation of Cdk2.

Pathway: Cell Cycle Checkpoints

http://pathwaycommons.org

Pathway Commons Status

Pathway Commons Quick Stats:

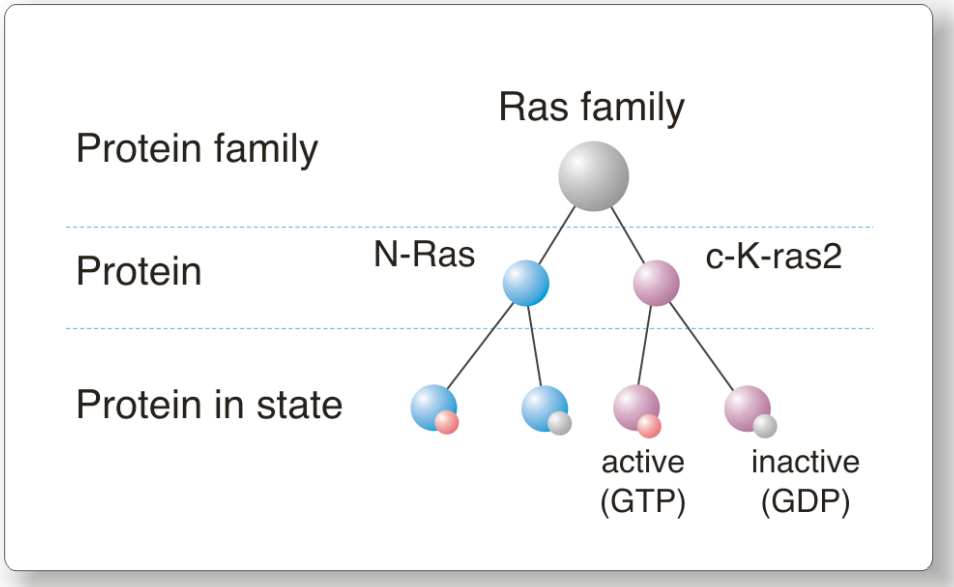
Number of Pathways:	1,449
Number of Interactions:	421,395
Number of Physical Entities:	88,509
Number of Organisms:	441

- Signaling
- Metabolism
- Molecular Interactions
- Future
  - Genetic Interactions
  - Gene Regulation

BioGRID
Memorial Sloan-Kettering Cancer Center
Human Protein Reference Database
HumanCyc
SBCNY
IntAct
MINT
National Cancer Institute
nature
Reactome

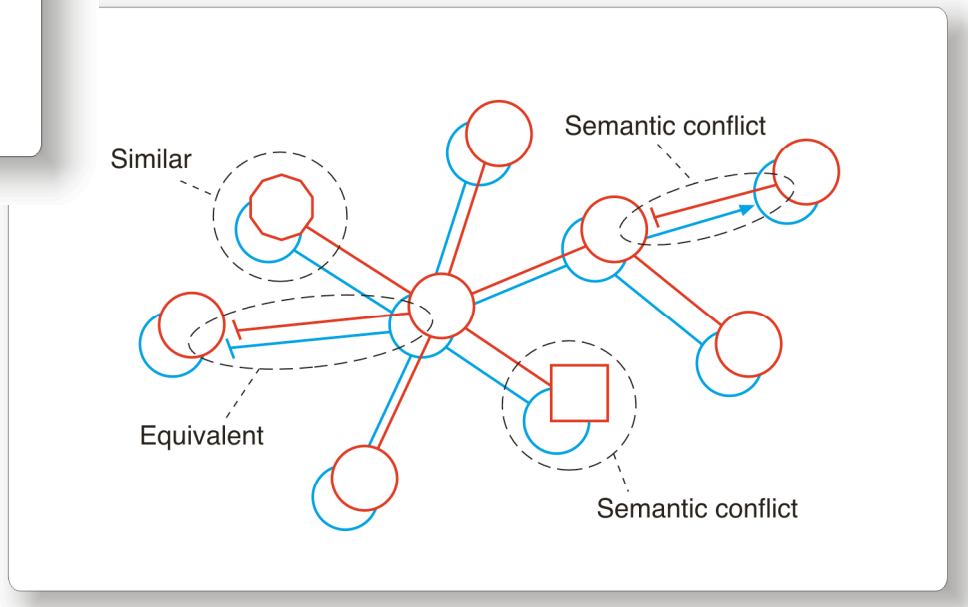
Towards an Integrated Cell Map

- Semantic pathway integration is difficult



Physical entities

Relationships

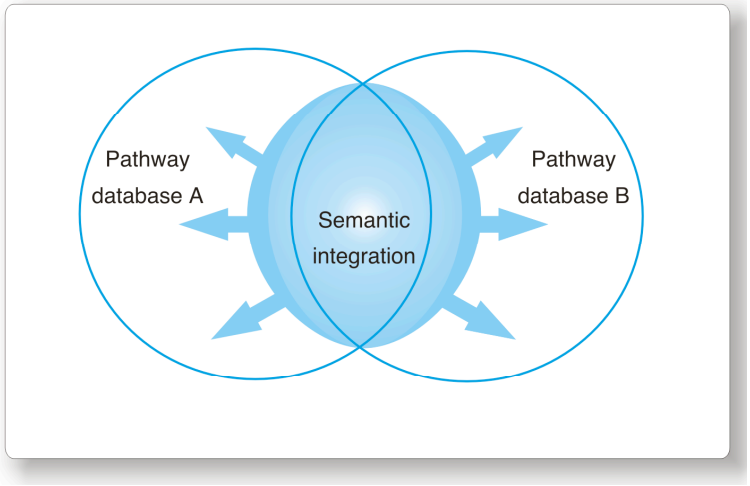


Determining equivalent entities is critical

Practical Semantic Integration

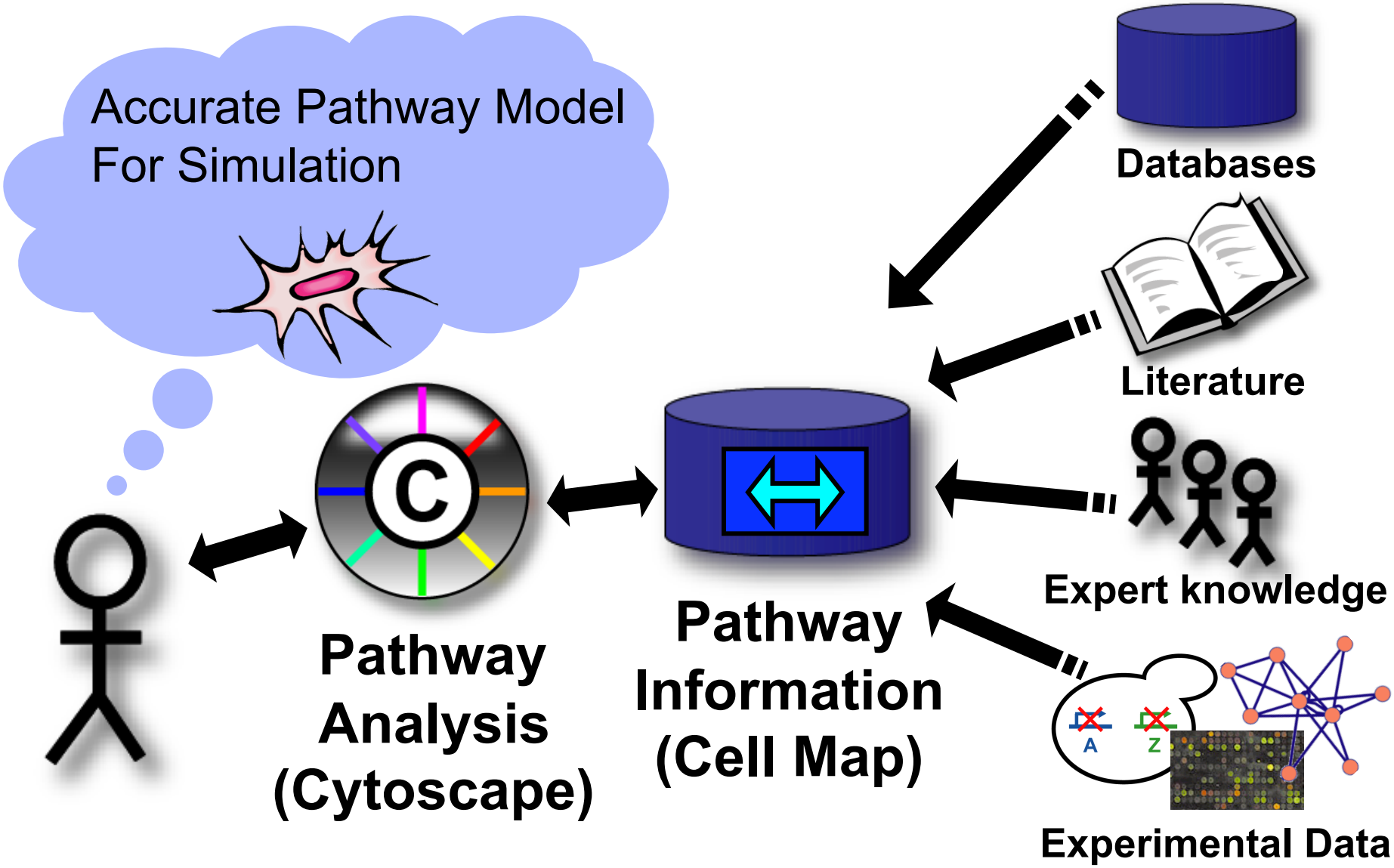
- Minimize errors
  - Integrate only where possible with high accuracy
  - Detect and flag conflicts, errors for users, no revision
  - Promote best-practices to minimize future errors
  - Interaction confidence algorithms
  - Validation software
  - Allow users to filter and select trusted sources
-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

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

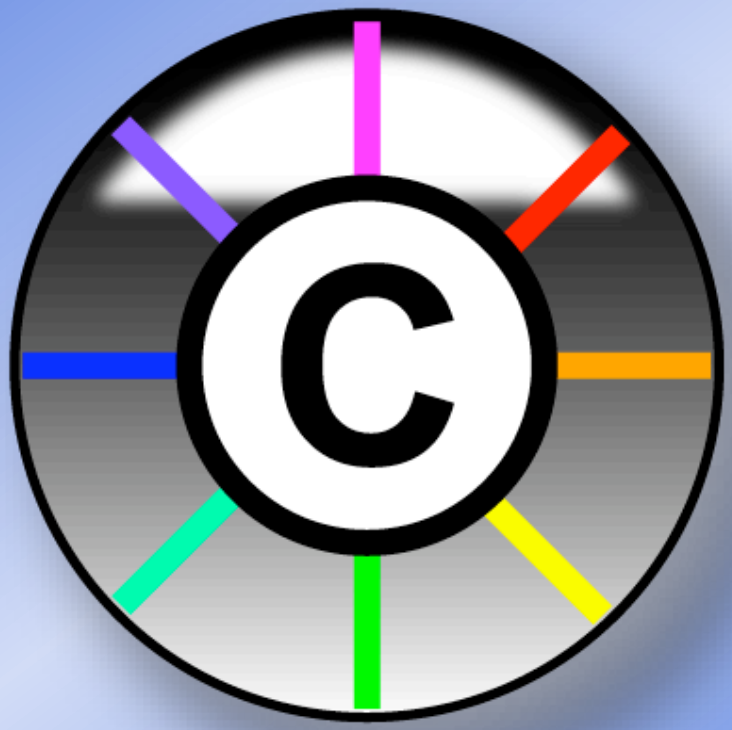


Network Visualization and Analysis

Using Pathway Information



Cytoscape



UNIVERSITY OF TORONTO

UCSF

NCIBI
National Center for Integrative Biomedical Informatics

Unilever

MEMORIAL SLOAN-KETTERING CANCER CENTER · ESTABLISHED 1884 ·

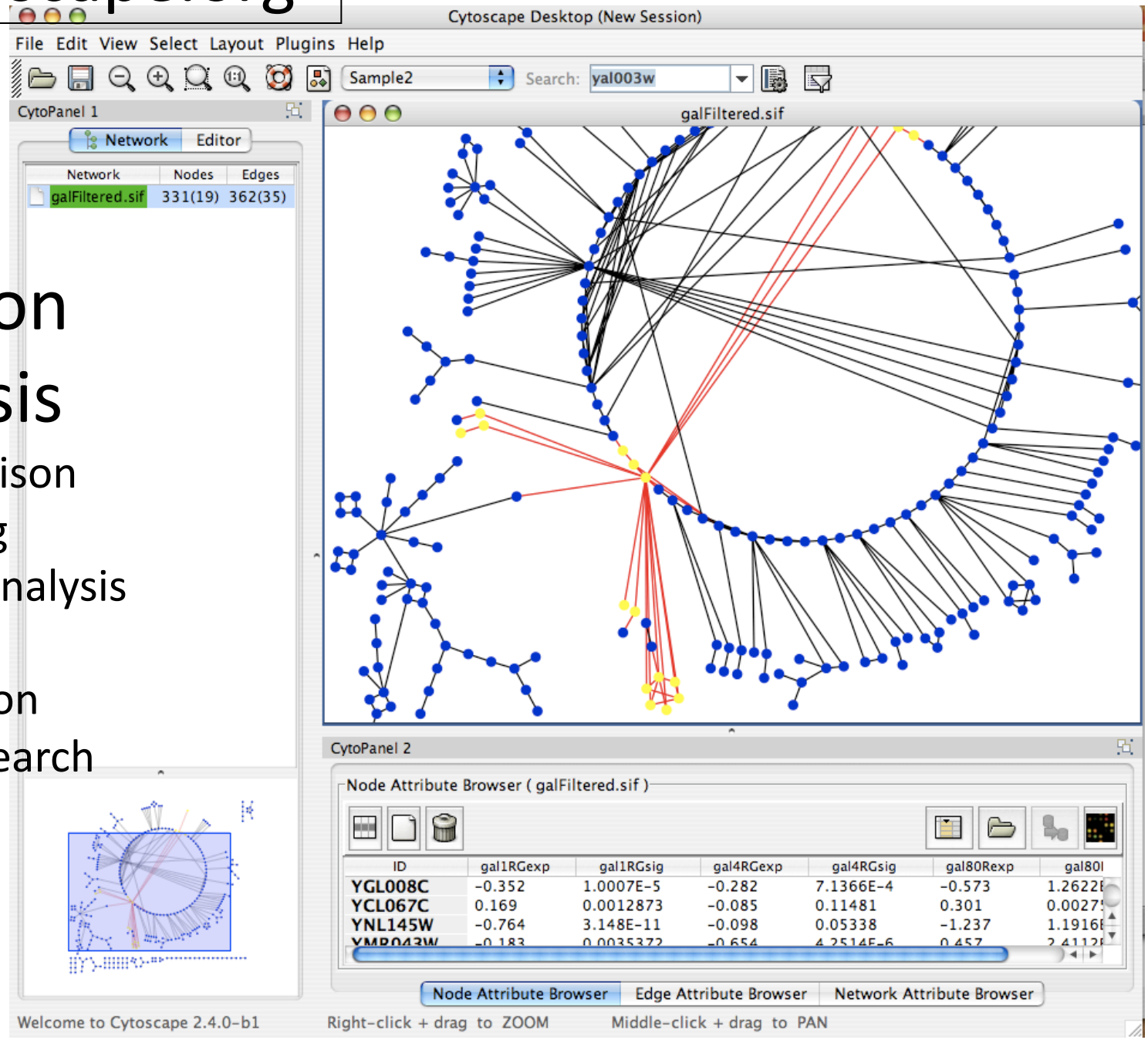
Agilent Technologies

http://cytoscape.org

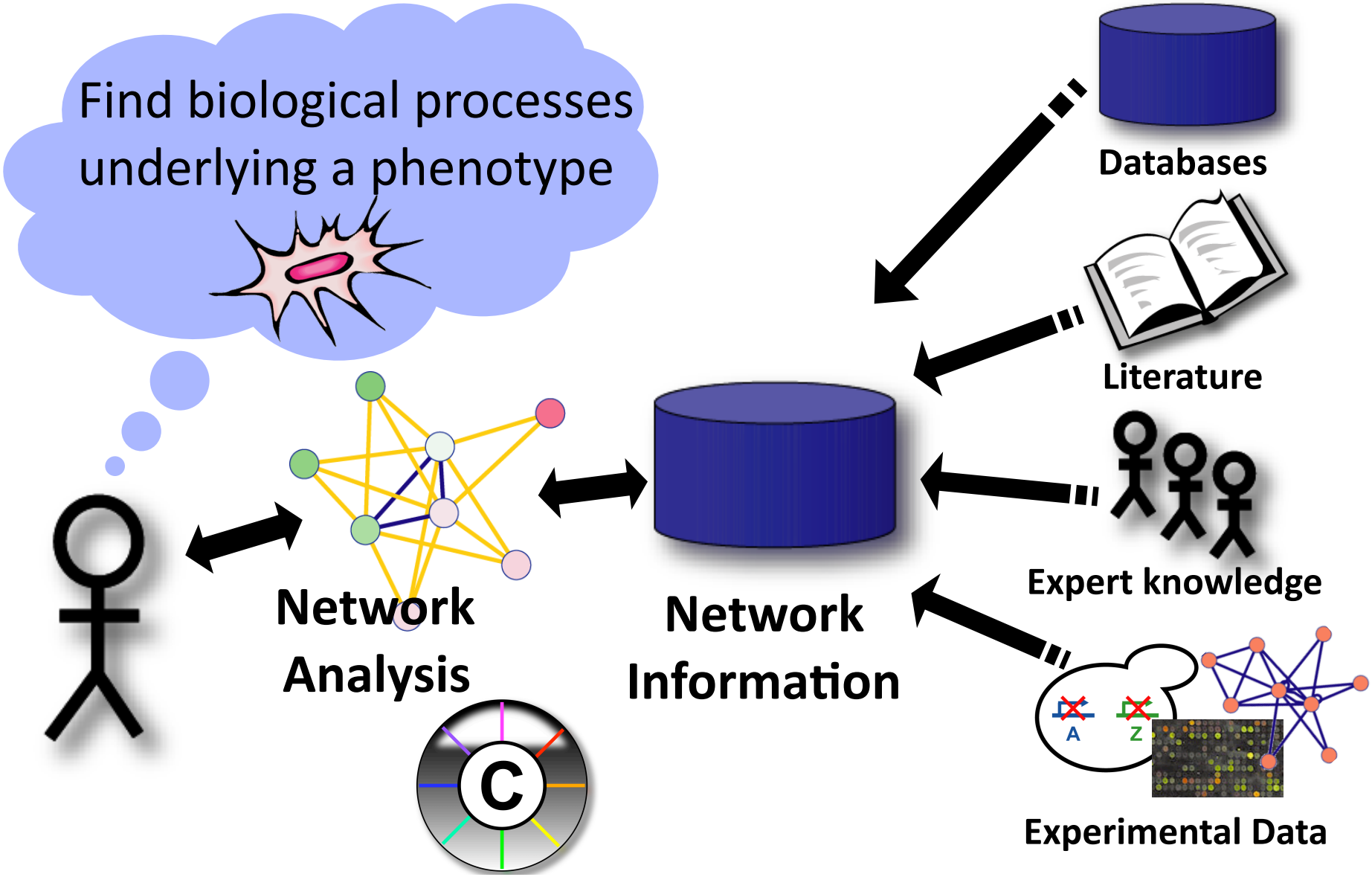
Network visualization and analysis

Pathway comparison
Literature mining
Gene Ontology analysis
Active modules
Complex detection
Network motif search

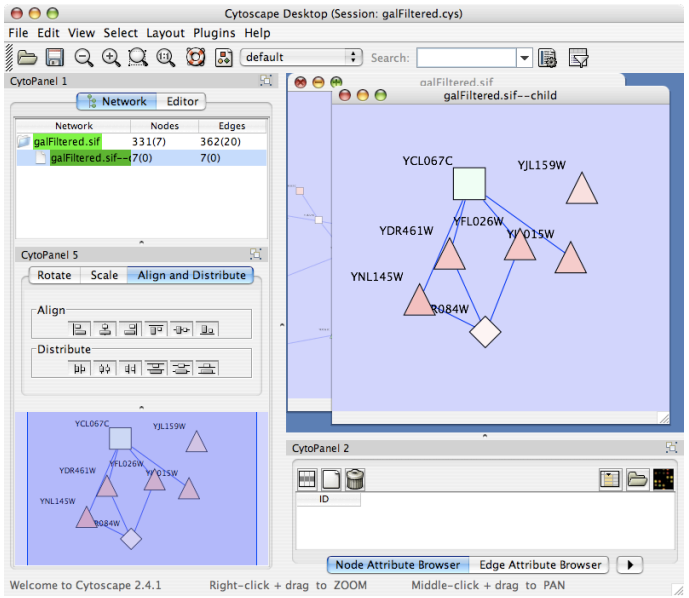
UCSD, ISB, Agilent, MSKCC, Pasteur, UCSF, Unilever, UToronto, U Michigan



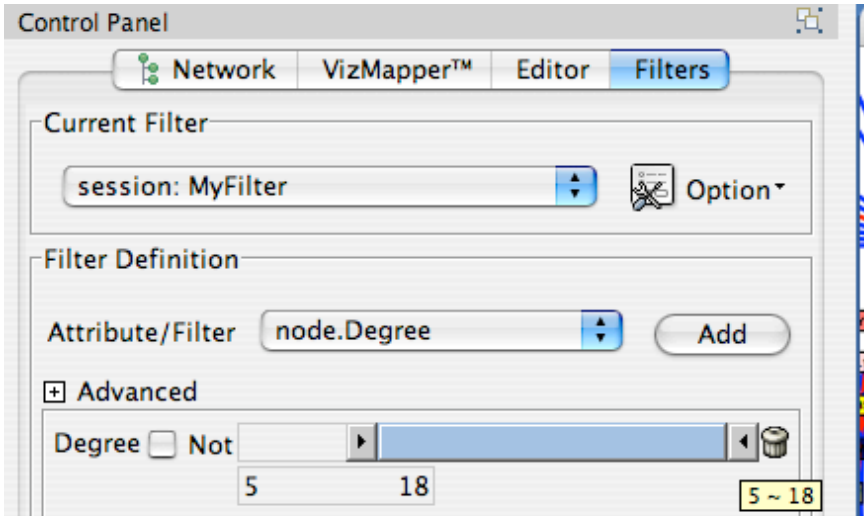
Network Analysis using Cytoscape



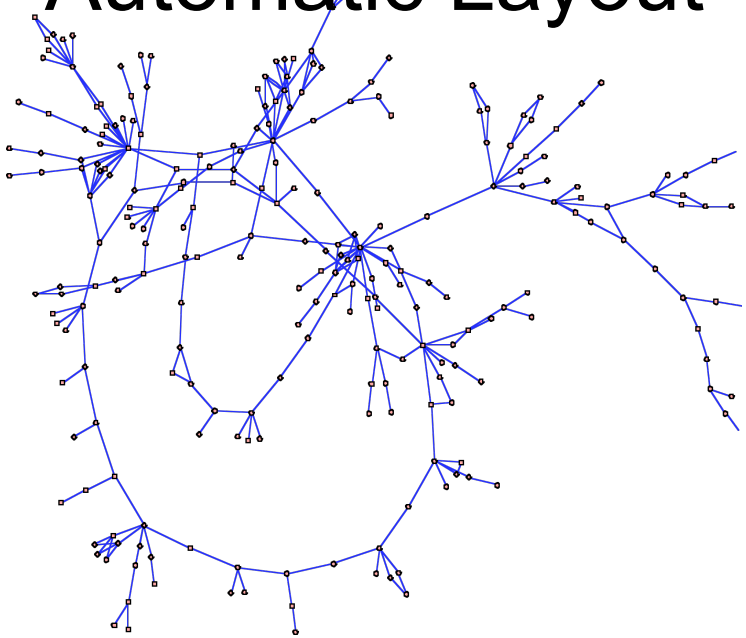
Manipulate Networks



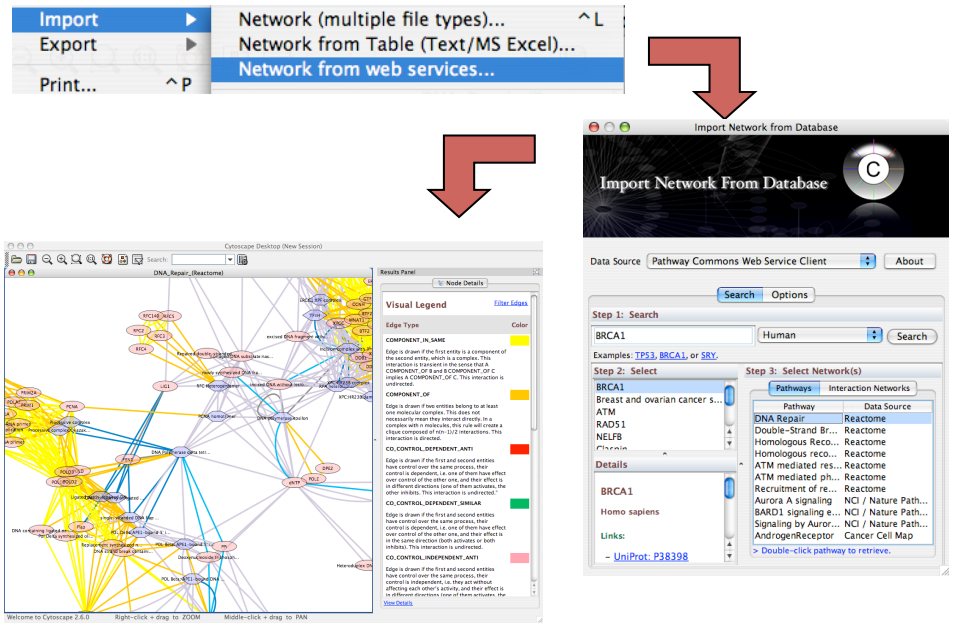
Filter/Query



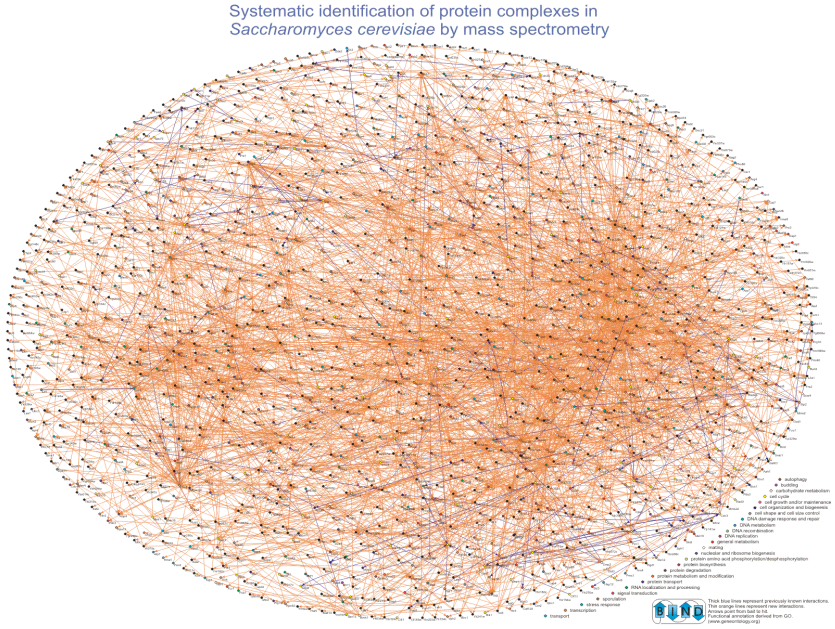
Automatic Layout



Interaction Database Search



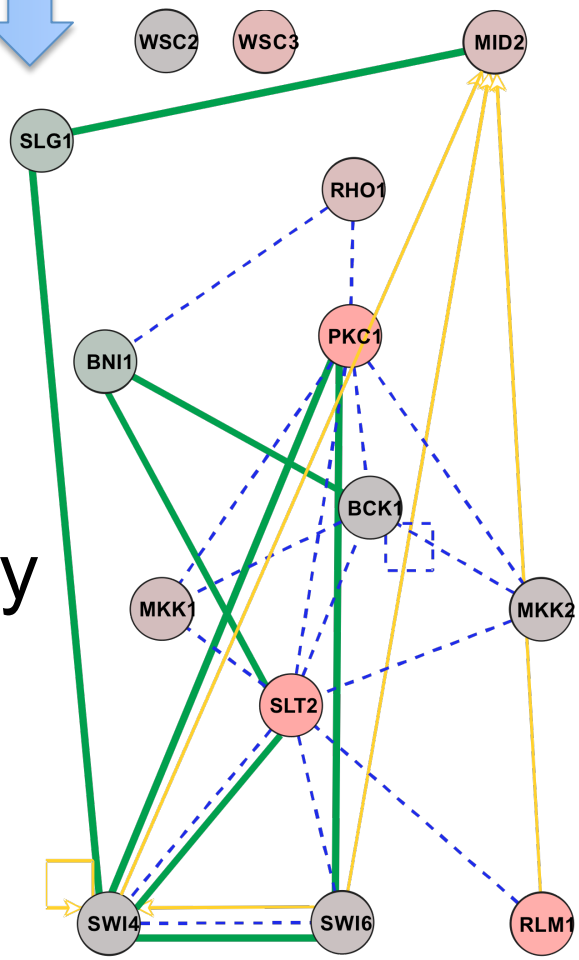
Overview



Zoom

Focus

PKC
Cell
Wall
Integrity



Synthetic Lethal

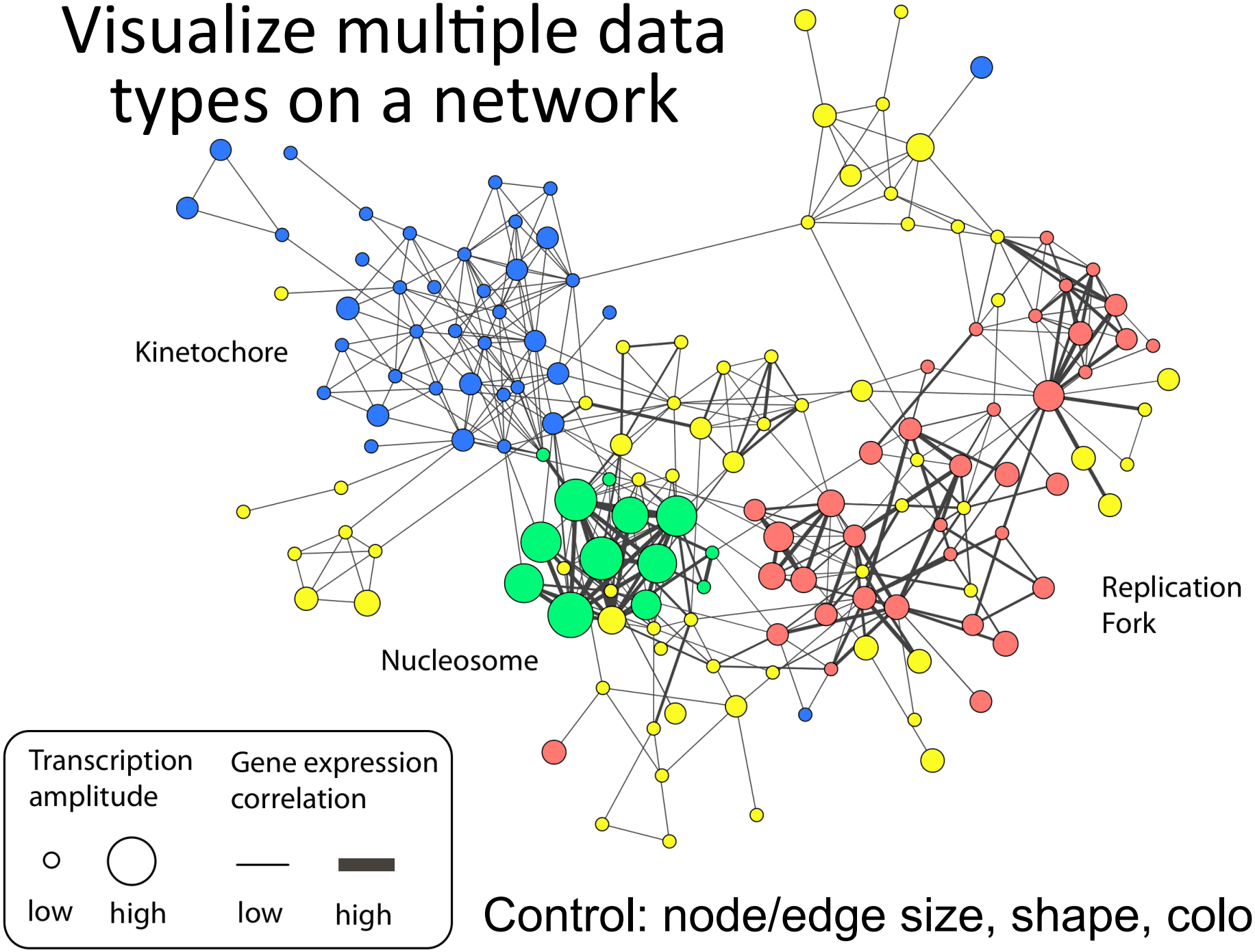
Transcription Factor Regulation

Protein-Protein Interaction

Up Regulated Gene Expression

Down Regulated Gene Expression

Visualize multiple data types on a network



Control: node/edge size, shape, color...

Active Community

http://www.cytoscape.org

- Help
  - 8 tutorials, >10 case studies
  - Mailing lists for discussion
  - Documentation, data sets
- 10,000s users, 2500 downloads/month
- >40 Plugins Extend Functionality
  - Build your own, requires programming
    - e.g. Retina Workbench

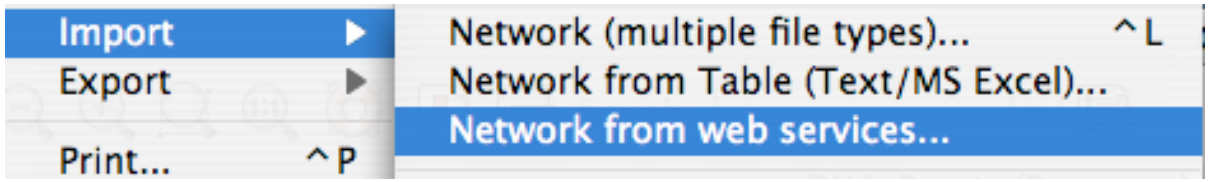
Cline MS et al. Integration of biological networks and gene expression data using Cytoscape Nat Protoc. 2007;2(10):2366-82

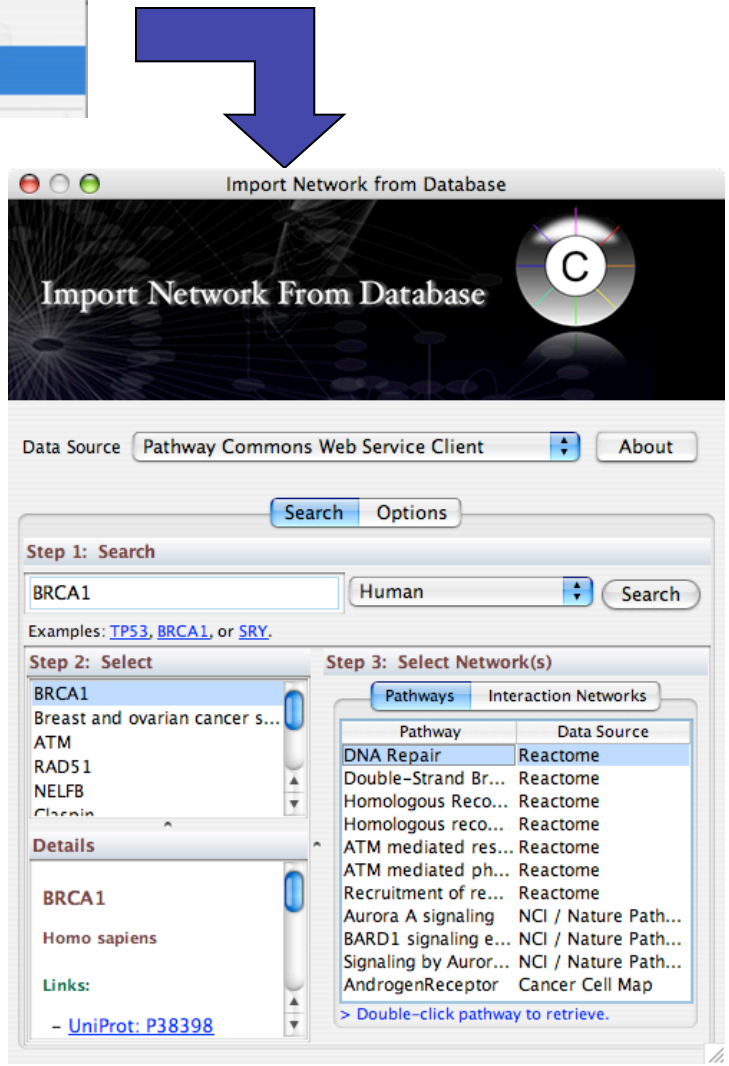
Analyzing Molecular Profiles

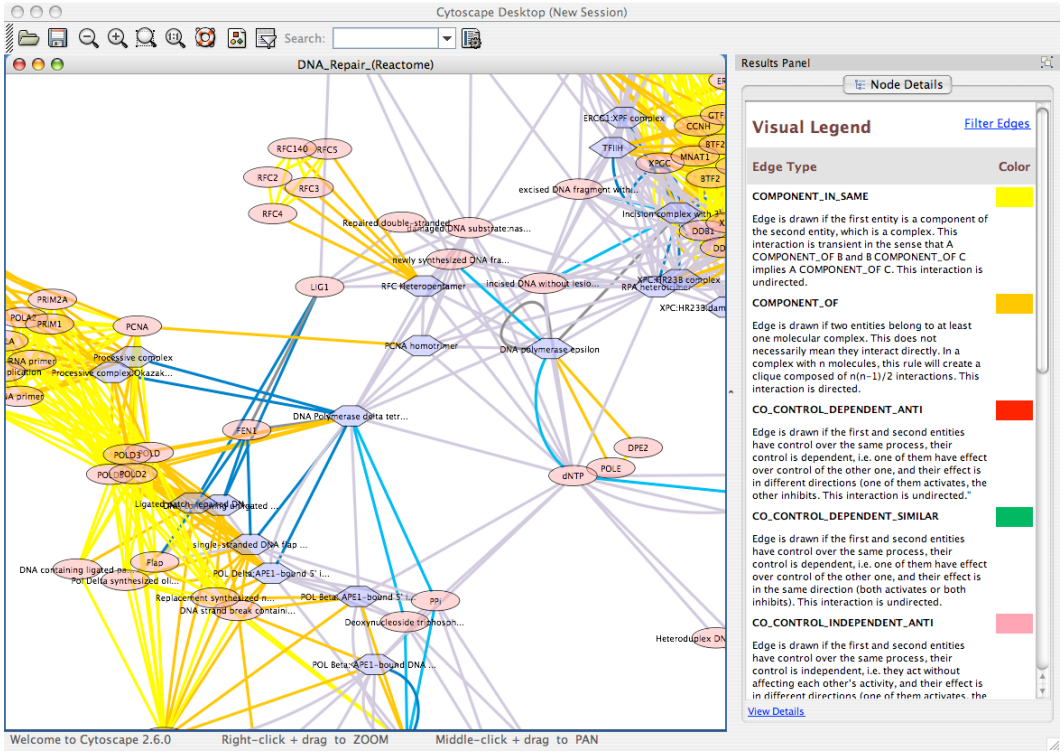
Analyzing gene expression data in a network context

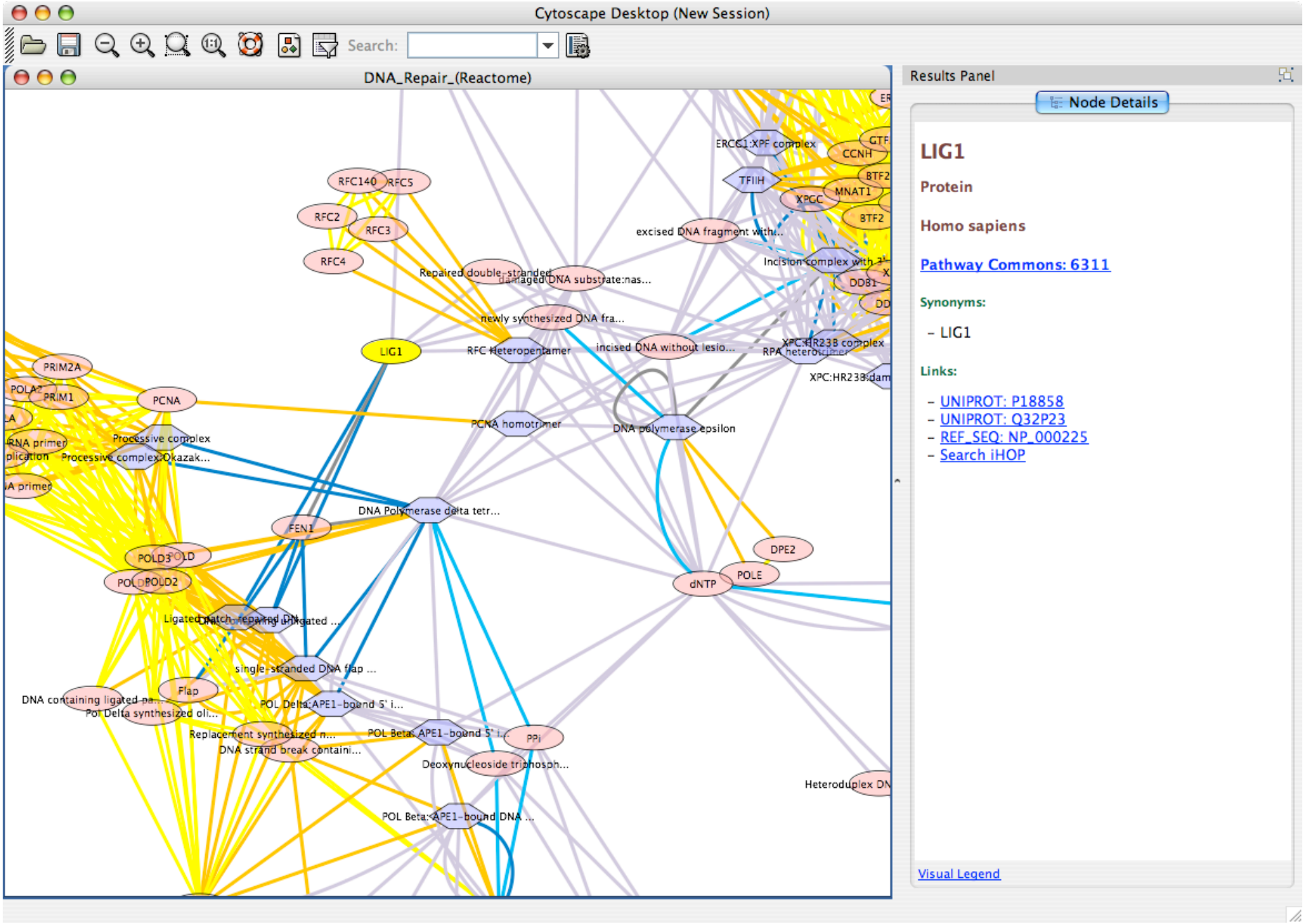
- Input
  - Gene expression data
  - Network data
- Output
  - Visual diagram of expression data on network
  - Active network regions
- Outline
  - Where to find network data?
    - Interaction database (cPath)
    - Literature associations via text mining
  - Load expression data
  - Identify active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

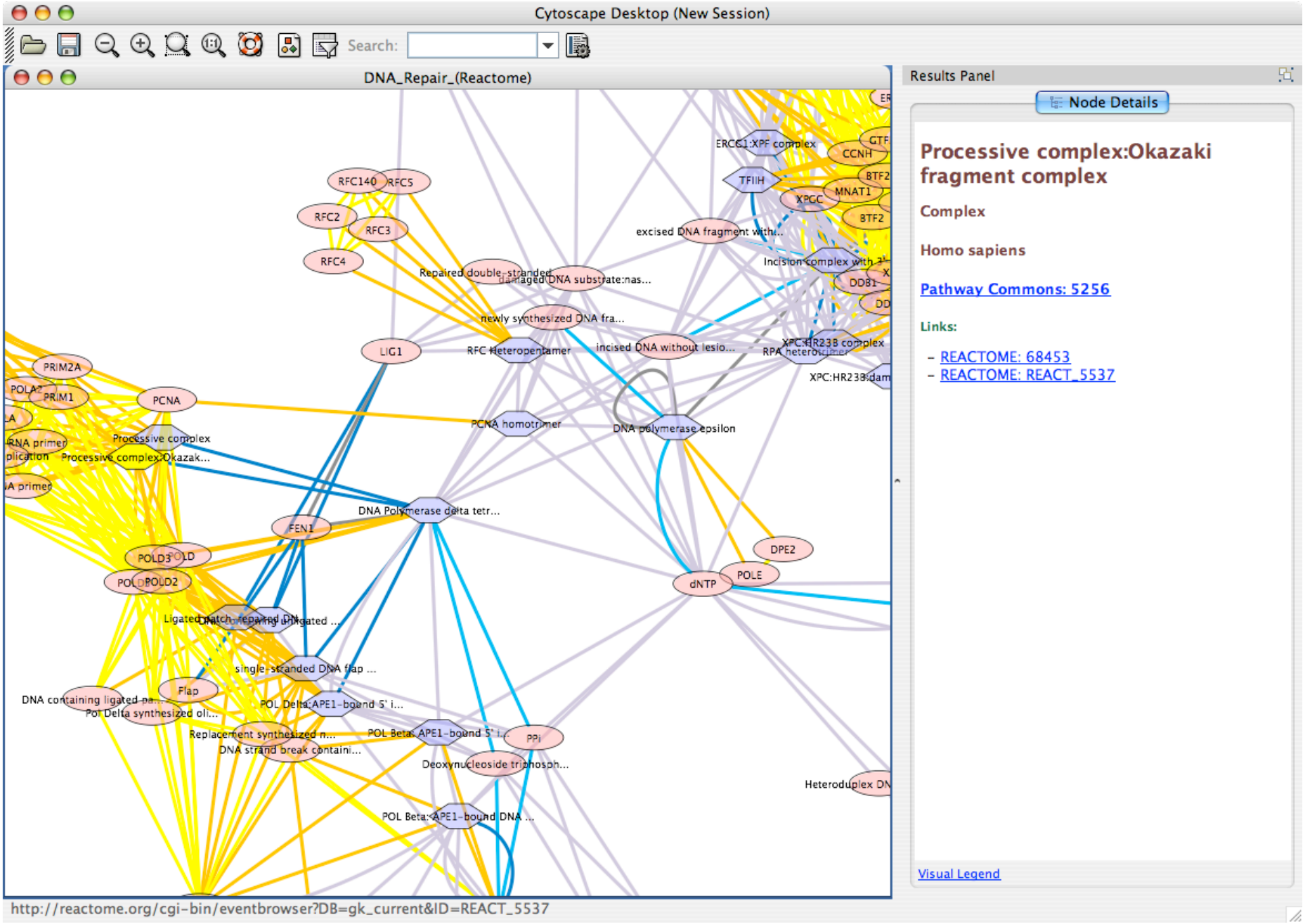
Interaction Database Search





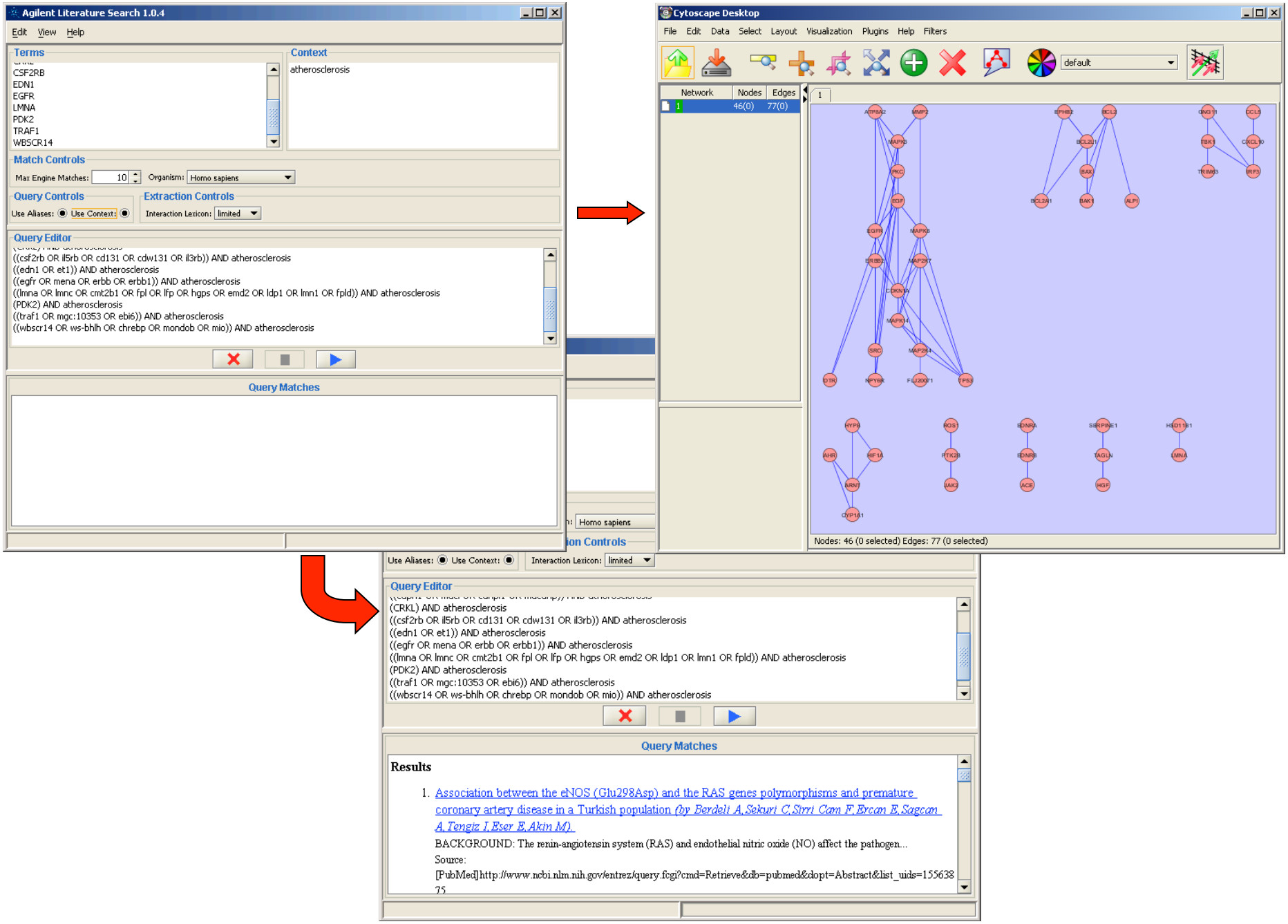



Cytoscape Desktop (New Session)
Search:
DNA_Repair_(Reactome)
RFC140
RFC5
RFC2
RFC3
RFC4
LIG1
PRIM2A
PRIM1
PCNA
Processive complex
Processive complex:Okazak...
FEN1
POLD3
POLD2
Flap
DNA Polymerase delta tetr...
PCNA homotrimer
DNA polymerase epsilon
RFC Heteropentamer
newly synthesized DNA fra...
incised DNA without lesio...
Repaired double-stranded damaged DNA substrate:nas...
excised DNA fragment with...
ERCC1:XPF complex
TFIIH
XPGC
MNAT1
CCNH
BTF2
DDB1
XPC:HR23B complex
RPA heterotrimer
DPE2
POLE
dNTP
single-stranded DNA flap ...
POL Delta:APE1-bound 5' i...
POL Beta:APE1-bound 5' i...
Replacement synthesized n...
DNA strand break containi...
Deoxynucleoside triphosph...
POL Beta:APE1-bound DNA ...
Heteroduplex DN
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

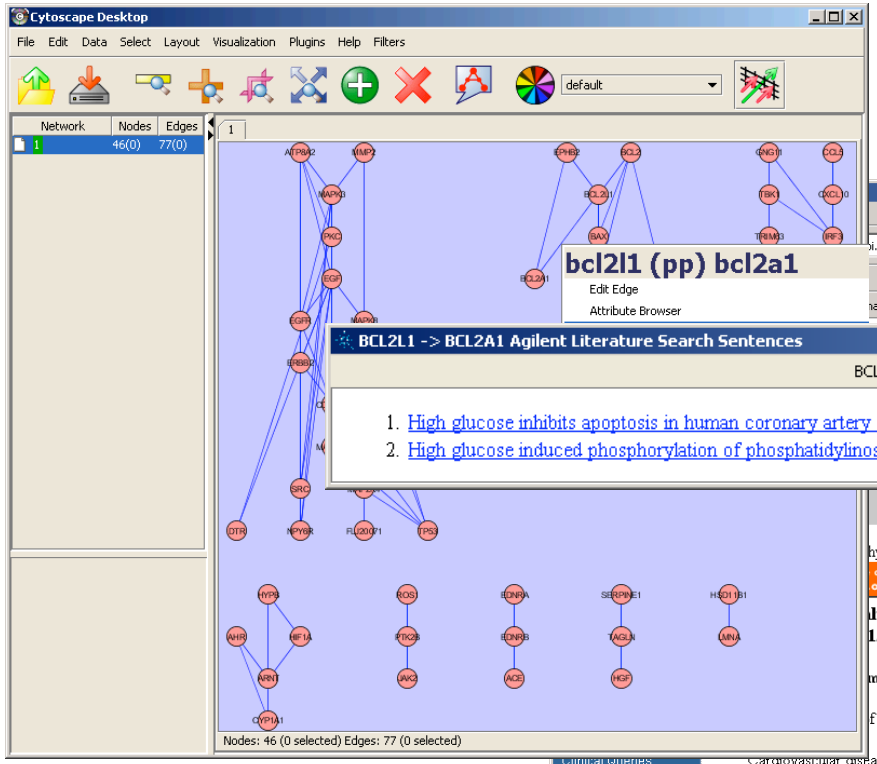

Cytoscape Desktop (New Session)
Search:
DNA_Repair_(Reactome)
Results Panel
Node Details
Processive complex:Okazaki fragment complex
Complex
Homo sapiens
Pathway Commons: 5256
Links:
- REACTOME: 68453
- REACTOME: REACT_5537
Visual Legend
RFC140
RFC5
RFC2
RFC3
RFC4
LIG1
PRIM2A
PRIM1
PCNA
Processive complex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

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/

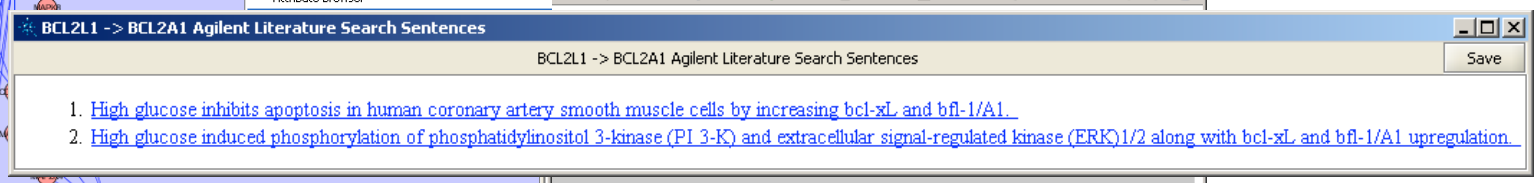

Agilent Literature Search 1.0.4
Edit View Help
Terms
CSF2RB
EDN1
EGFR
LMNA
PDK2
TRAF1
WBSCR14
Context
atherosclerosis
Match Controls
Max Engine Matches: 10
Organism: Homo sapiens
Query Controls
Use Aliases: Use Context:
Extraction Controls
Interaction Lexicon: limited
Query Editor
((csf2rb OR il5rb OR 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)
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 Cam 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

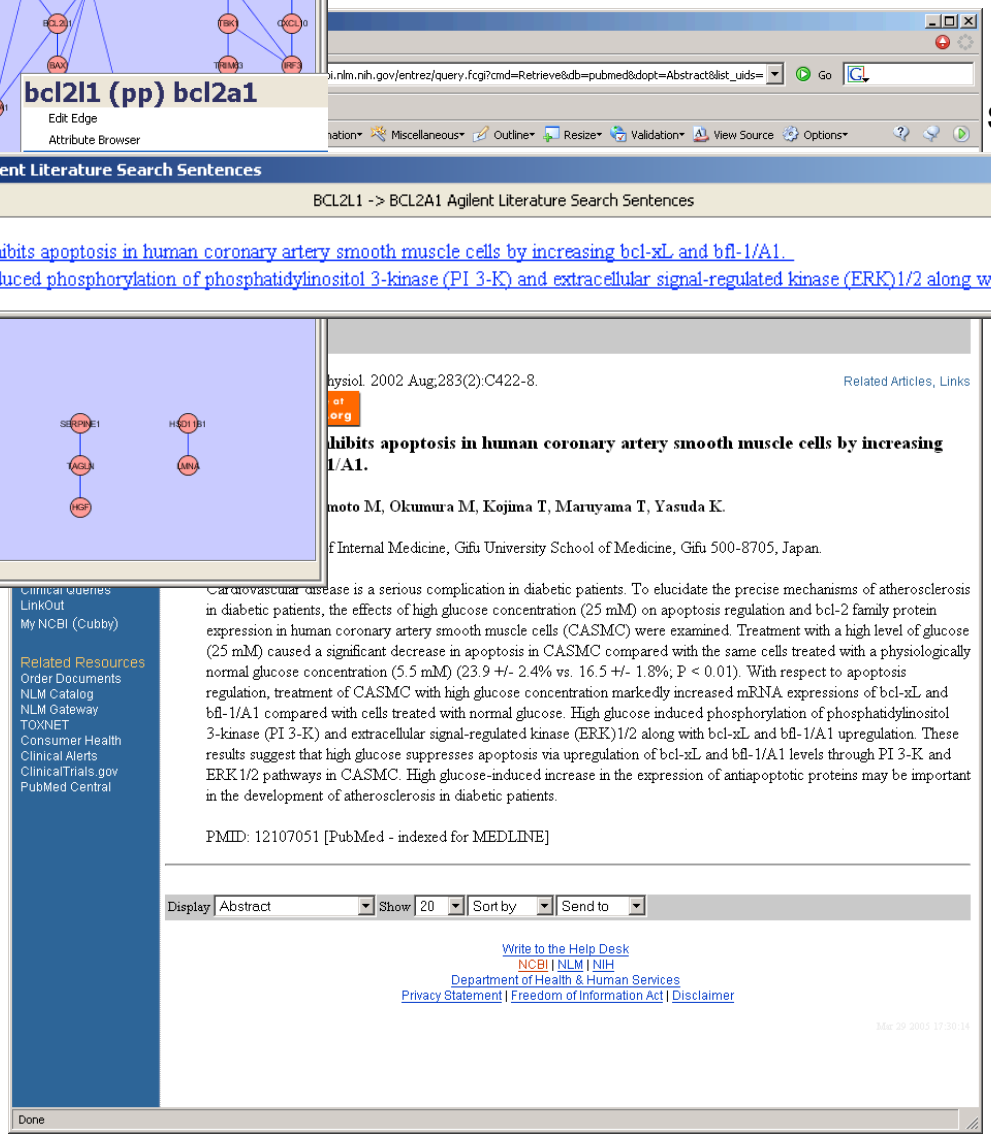
Cytoscape Network produced by Literature Search.



Abstract from the scientific literature

Sentences for an edge





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:

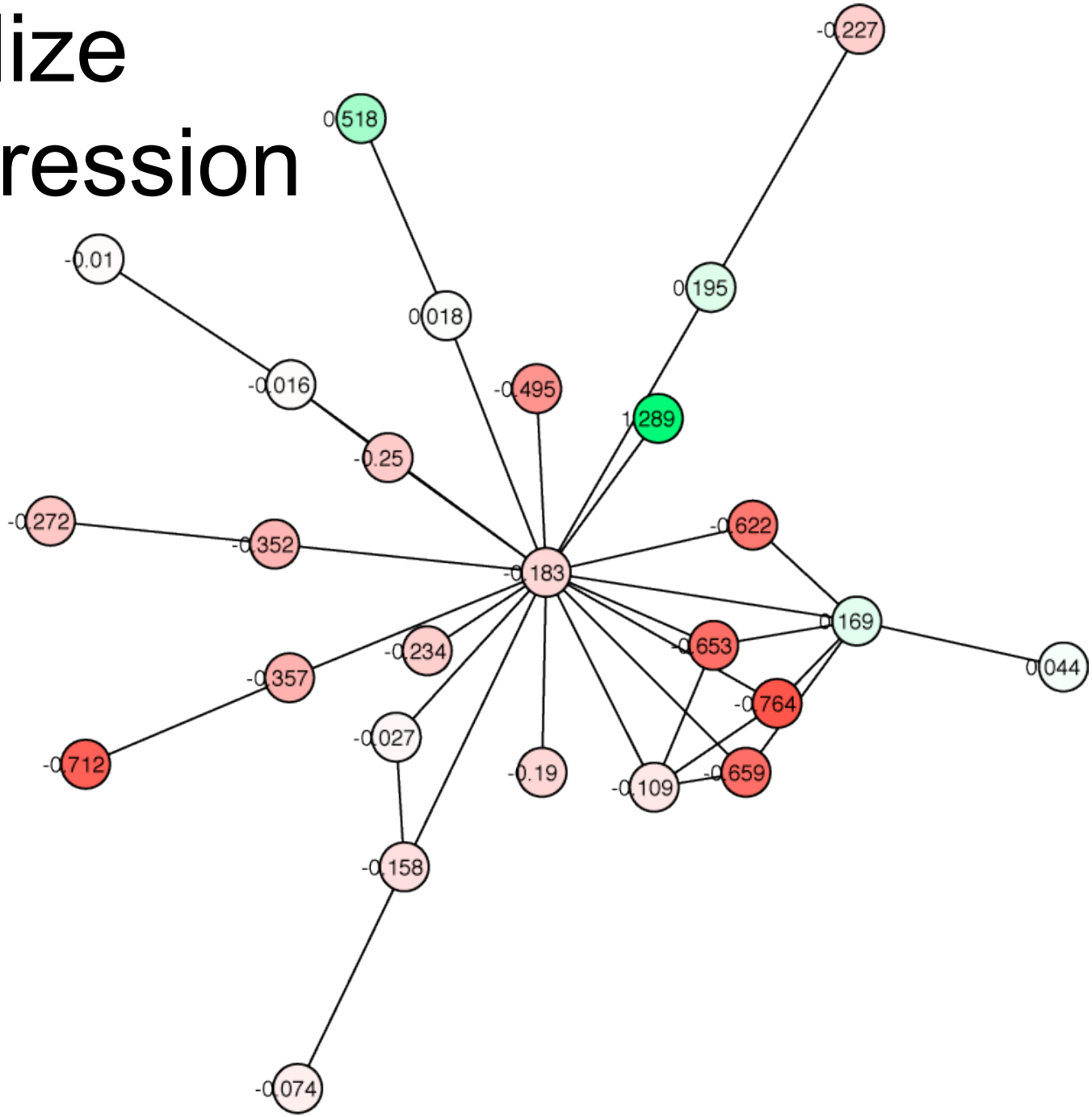
Submit

*	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

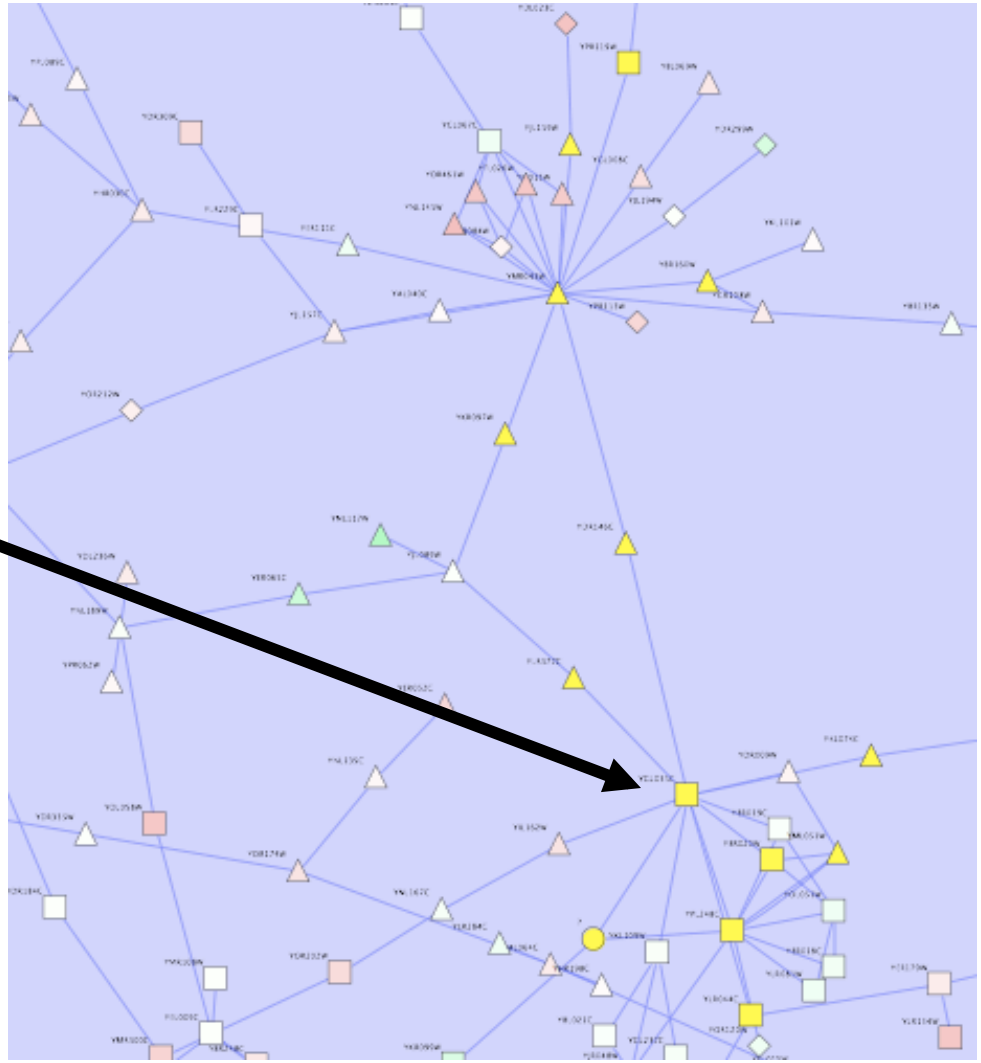
Active Module Results

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

Conditions vs. Pathways

Network	Size	Score	gal1RGsig	gal4RGsig	gal80Rsig
1	14	3.78			
2	26	3.584			
3	10	2.994			
4	7	2.934			
5	4	2.636			

Save Dismiss



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

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

Bonus Slides

Gene and Protein Identifiers

- Identifiers (IDs) are names or numbers that help track database records
  - E.g. Social Insurance Number, Entrez Gene ID 41232
- Gene and protein information stored in many databases
  - → Genes have many IDs
- Records for: Gene, DNA, RNA, Protein
  - Important to use the correct record type
  - E.g. Entrez Gene records don't store sequence. They link to DNA regions, RNA transcripts and proteins.

Common Identifiers

Gene
Ensembl ENSG00000139618
Entrez Gene 675
Unigene Hs.34012

RNA transcript
GenBank BC026160.1
RefSeq NM_000059
Ensembl ENST00000380152

Protein
Ensembl ENSP00000369497
RefSeq NP_000050.2
UniProt BRCA2_HUMAN or A1YBP1_HUMAN
IPI IPI00412408.1
EMBL AF309413
PDB 1MIU

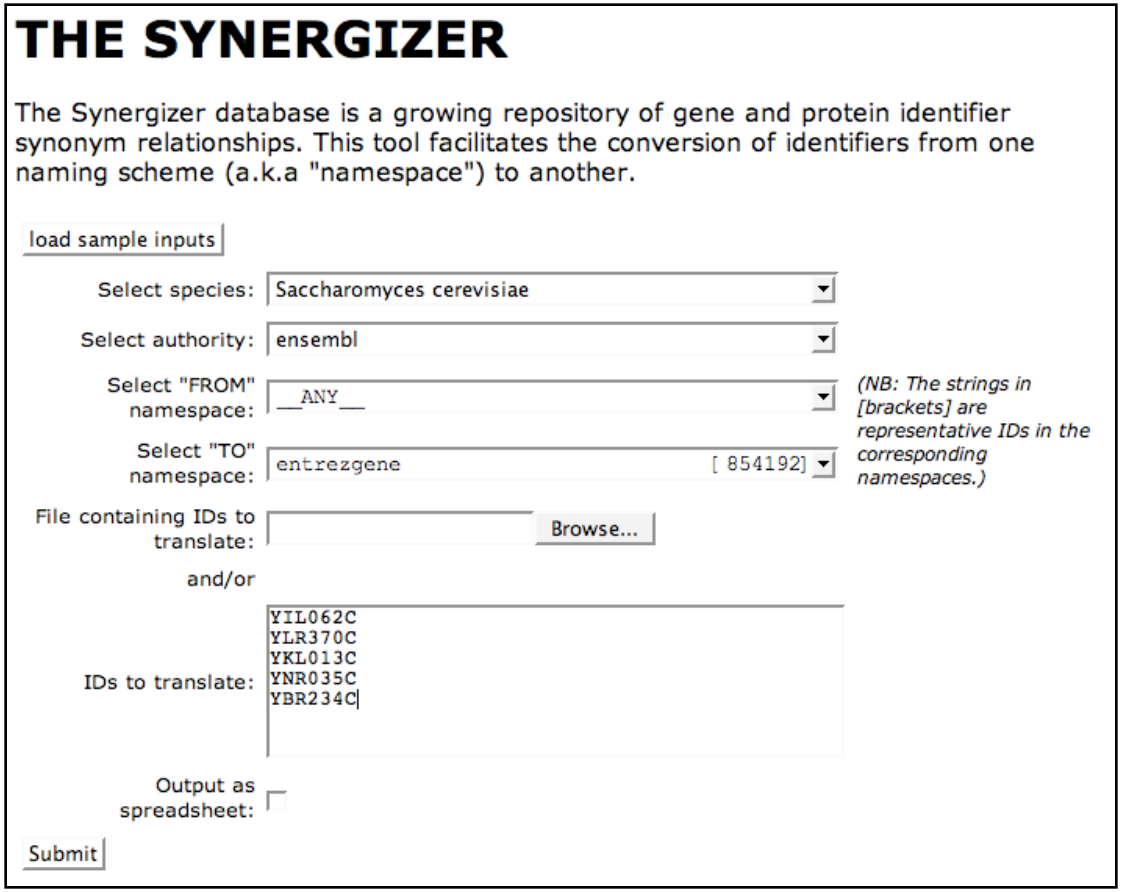
Species-specific
HUGO HGNC BRCA2
MGI MGI:109337
RGD 2219
ZFIN ZDB-GENE-060510-3
FlyBase CG9097
WormBase WBGene00002299 or ZK1067.1
SGD S000002187 or YDL029W

Annotations
InterPro IPR015252
OMIM 600185
Pfam PF09104
Gene Ontology GO:0000724
SNPs rs28897757

Experimental Platform
Affymetrix 208368_3p_s_at
Agilent A_23_P99452
CodeLink GE60169
Illumina GI_4502450-S

Red = Recommended

ID Mapping Services



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

- Synergizer
  - http://llama.med.harvard.edu/cgi/synergizer/translate
- Ensembl BioMart
  - http://www.ensembl.org
- PIR
  - http://pir.georgetown.edu/pirwww/search/idmapping.shtml

ID Mapping Challenges

- Gene name ambiguity
  - Not a good ID, but official gene symbol is ok e.g. HGNC/HUGO gene symbol
- Excel error-introduction
  - OCT4 is changed to October-4
- Problems reaching 100% coverage
  - E.g. due to version issues
  - Use multiple sources to increase coverage

Zeeberg BR et al. Mistaken identifiers: gene name errors can be introduced inadvertently when using Excel in bioinformatics BMC Bioinformatics. 2004 Jun 23;5:80

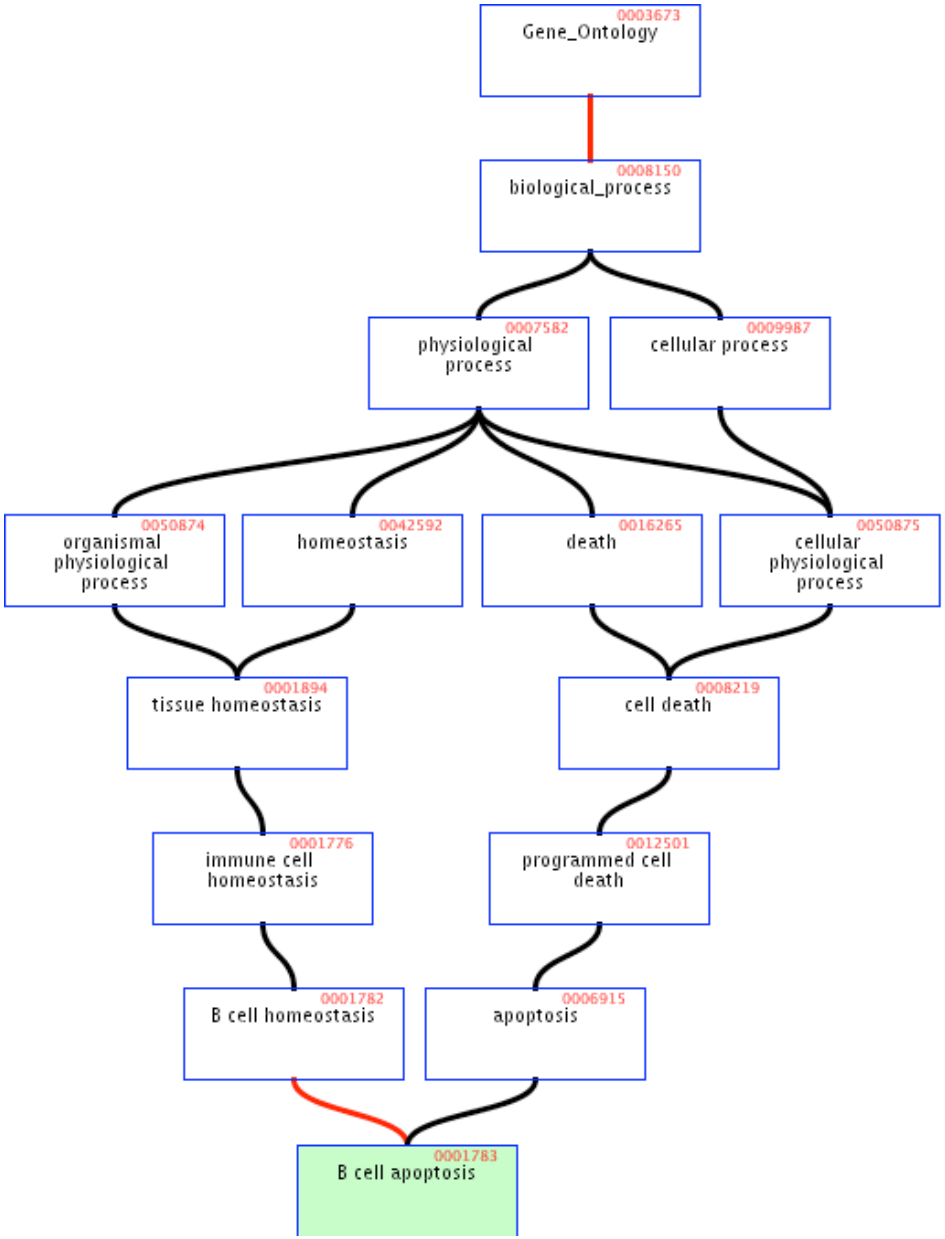
Additional Plugins

- Bingo: over-representation analysis
- ClusterMaker: clusters networks, includes MCL
- NetworkAnalyzer: calculates statistics about a network
- (You may have to use an earlier version of Cytoscape to get some plugins to run)

The Gene Ontology (GO)

- Describes gene function

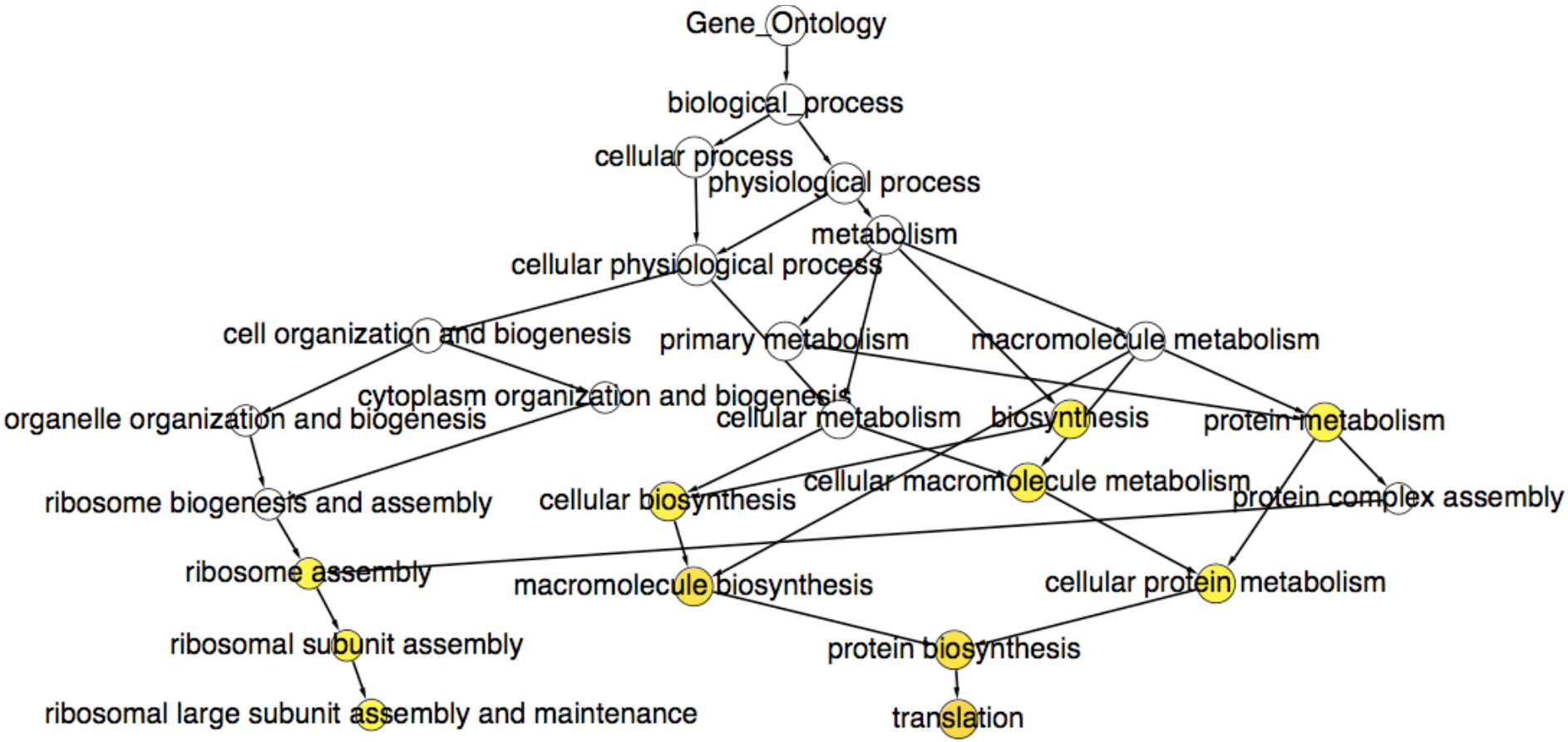
1. Agreed upon terms (controlled vocabulary)
   - Biological process
   - Cellular component
   - Molecular function
2. Genome annotation



www.geneontology.org

BiNGO

Hypergeometric p-value
Multiple testing correction
(Benjamini-Hochberg FDR)



Caveats: Gene identifiers must match;
low GO term coverage, GO bias

Maere, S., Heymans, K. and Kuiper, M
Bioinformatics 21, 3448-3449, 2005

NetMatch

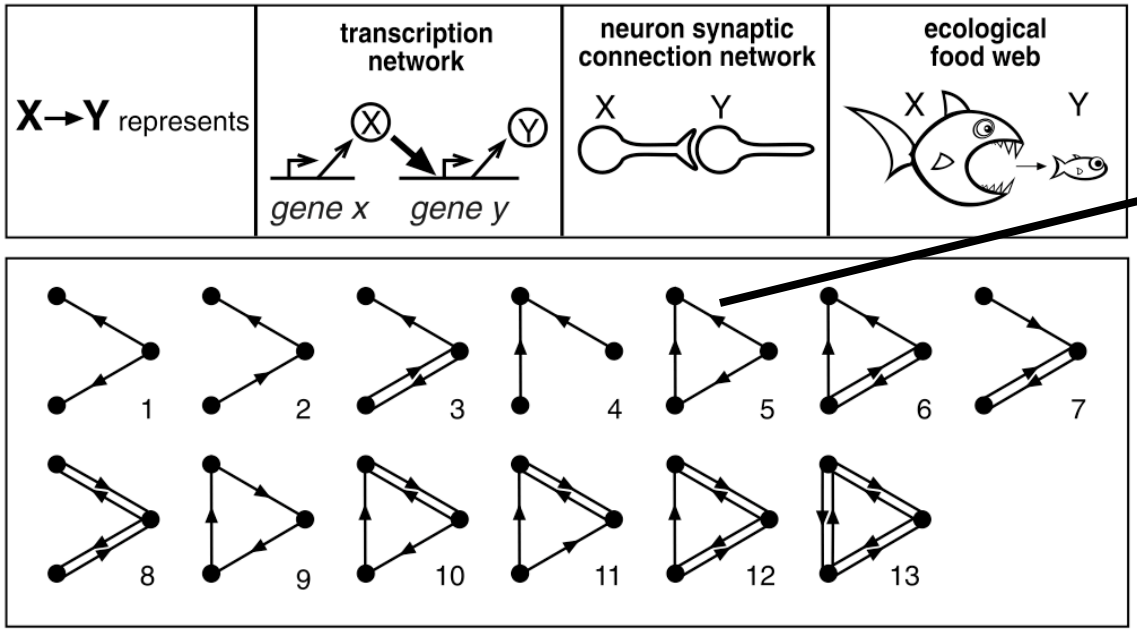
- Query a network for topological matches
- Input: query and target networks, optional node/edge labels
- Output: Topological query matches as subgraphs of target network
- Supports: subgraph matching, node/edge labels, label wildcards, approximate paths
- http://alpha.dmi.unict.it/~ctnyu/netmatch.html

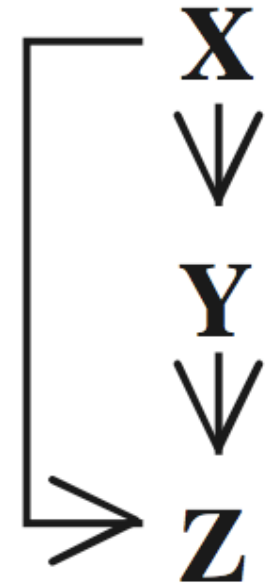
Ferro A, Giugno R, Pigola G, Pulvirenti A, Skripin D, Bader GD, Shasha D
Bioinformatics 2007 Feb 3

Extends state space representation based search from Cordella et al. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2004, 26, 10, 1367--1372

Find Feed-Forward Motifs

- Graph motifs over-represented in many network types



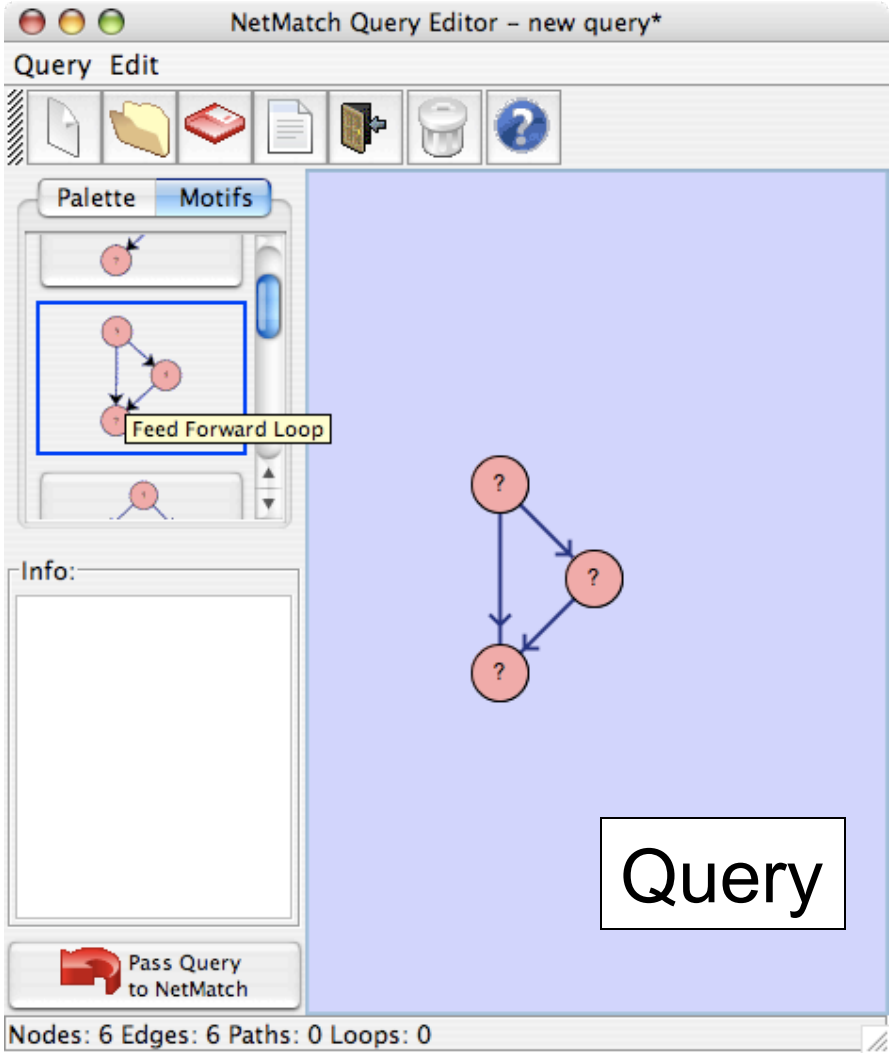


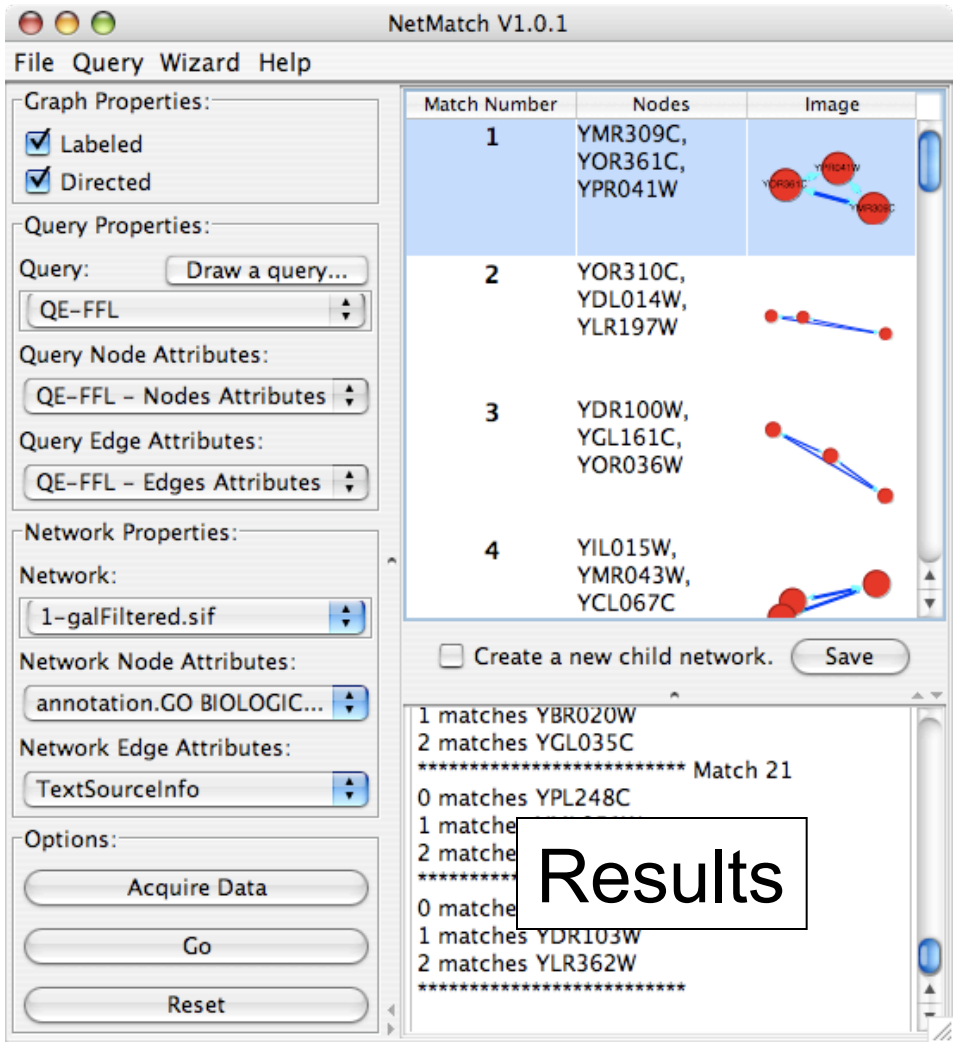
Feed-forward loop

Gene regulation
Neurons
Electronic circuits

Milo et al. Science 298, 2002

Find Feed-Forward Motifs

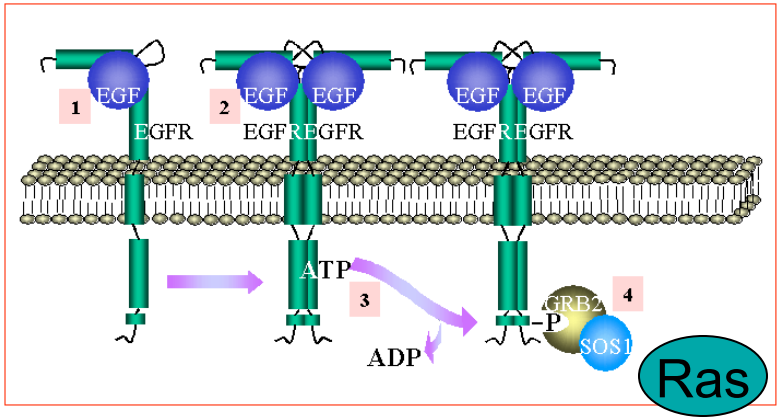




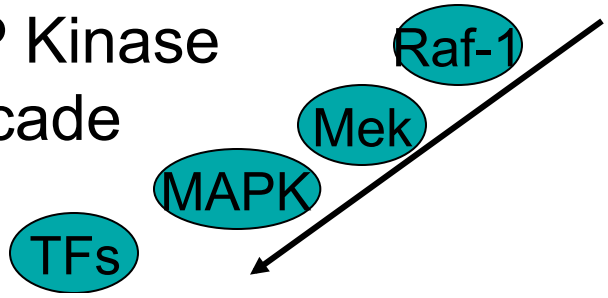
Find Signaling Pathways

- Potential signaling pathways from plasma membrane to nucleus via cytoplasm

Signaling pathway example

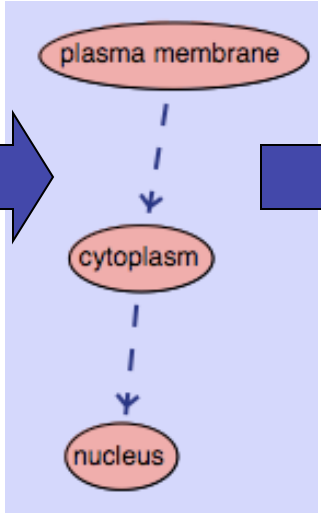


MAP Kinase Cascade



Nucleus - Growth Control
Mitogenesis

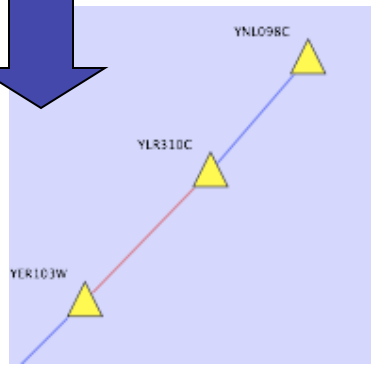
NetMatch query



NetMatch Results

Match Number	Nodes	Image
	YGL008C	
4	YJL157C, YMR043W, YLR229C	
5	YJL157C, YAL040C, YLR229C	
6	YLR310C, YER103W, YNL098C	

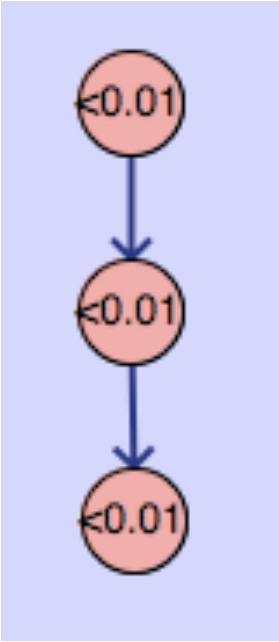
Shortest path between subgraph matches



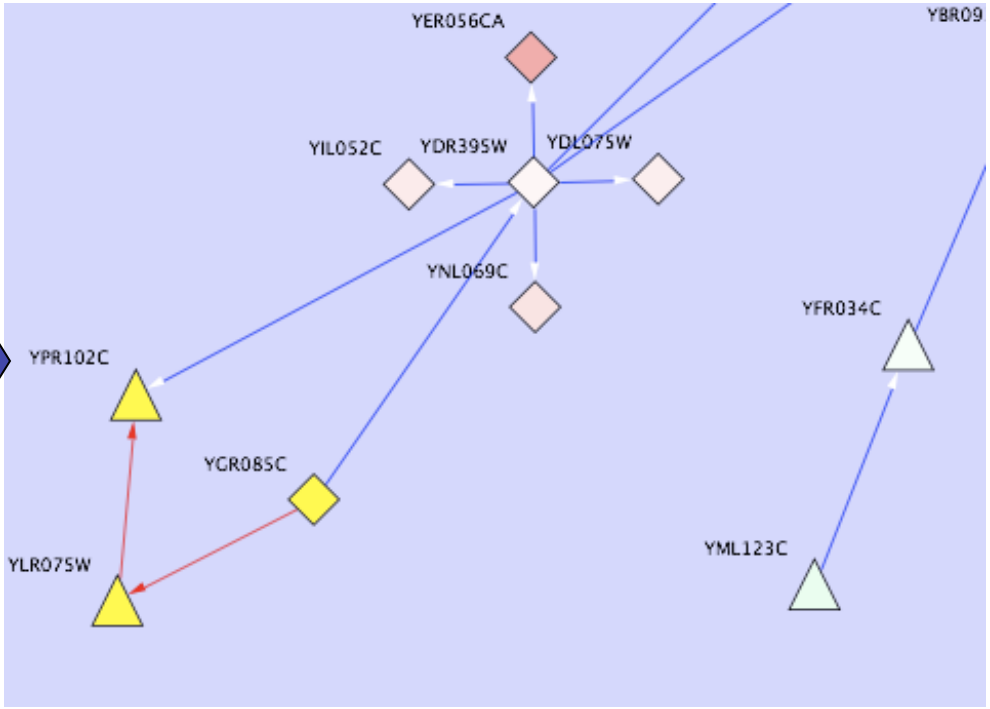
Find Expressed Motifs

Find specific subgraphs where certain nodes are significantly differentially expressed

NetMatch query



NetMatch Results



Protein	Differential Expression Significance
YLR075W	1.7255E-4
YGR085C	2.639E-4
YPR102C	3.7183E-4

Systems Biology Graphical Notation

http://sbgn.org