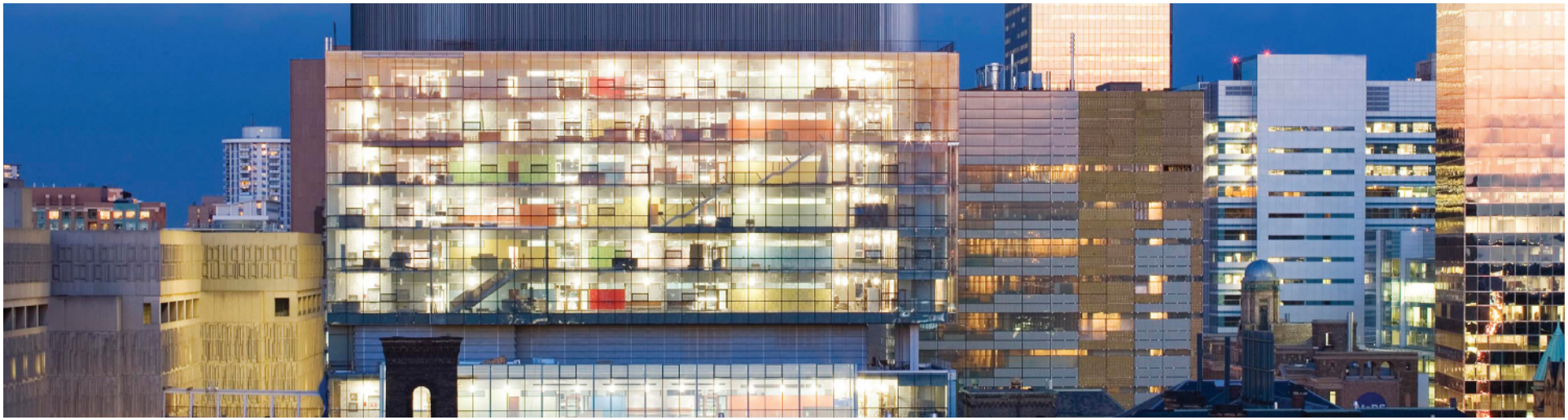
Integrative analysis of interaction networks

Gary Bader http://www.baderlab.org
MoGen – Mar.14.2012

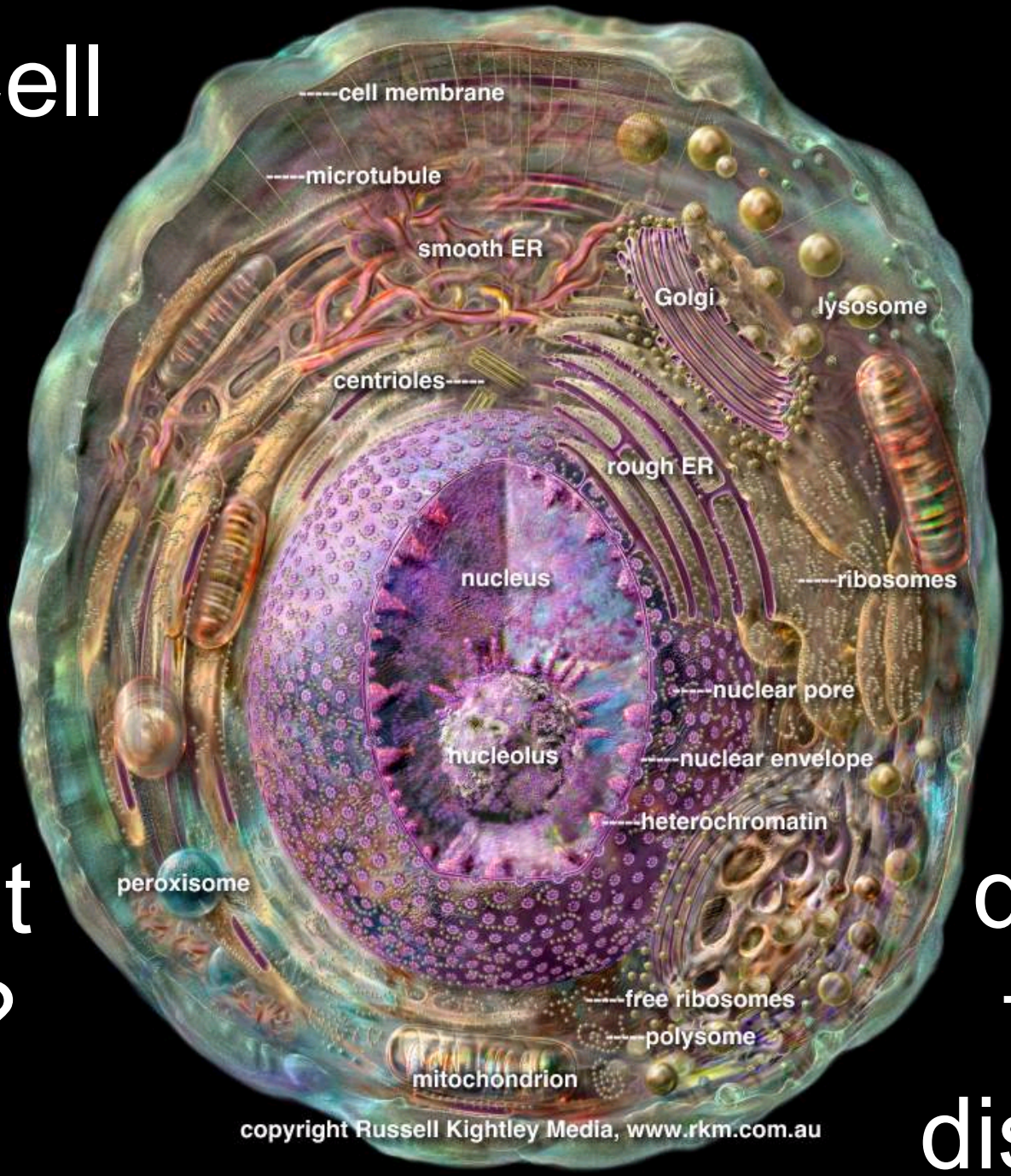


Outline

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

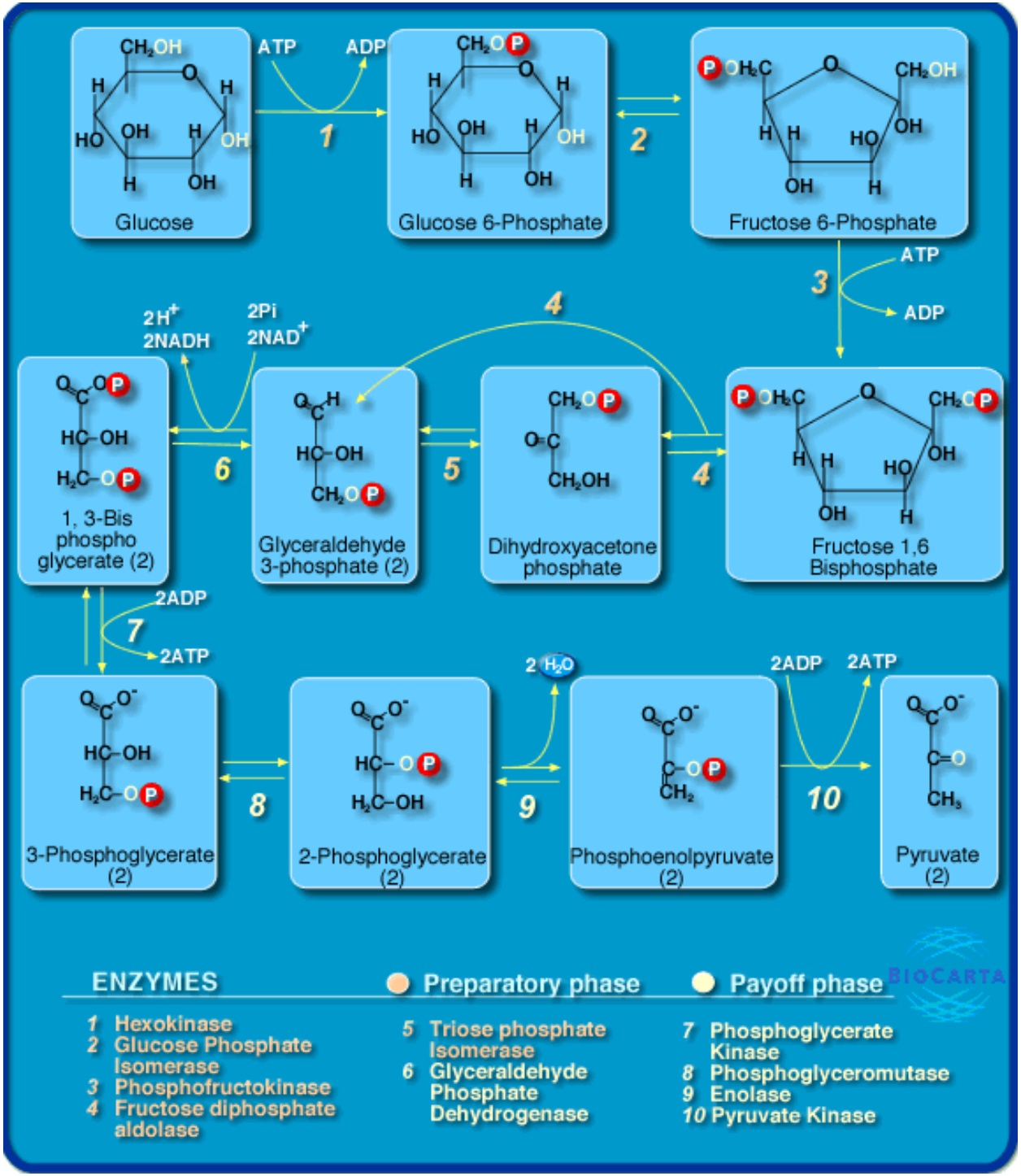
Data integration using networks

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
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

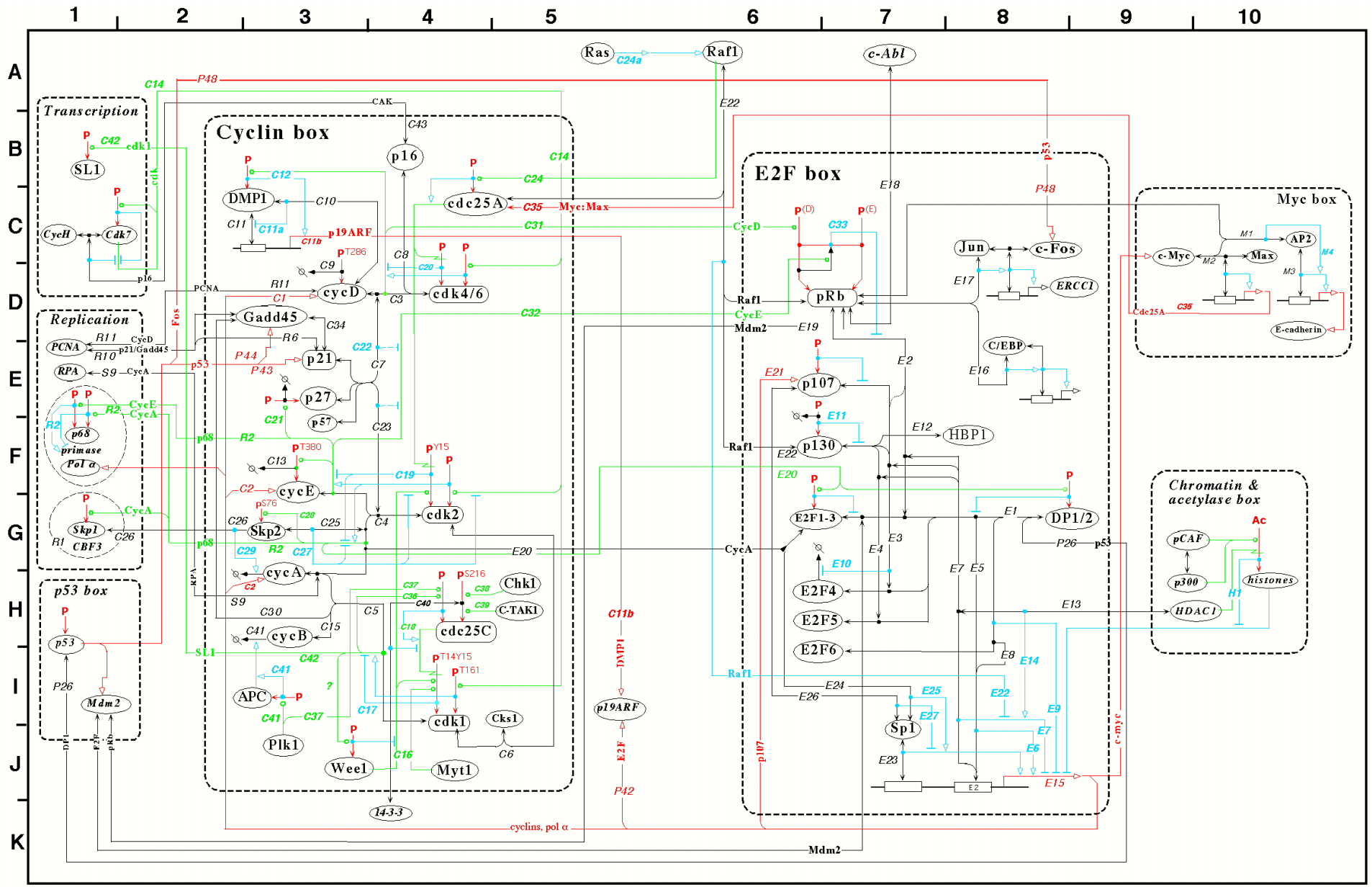
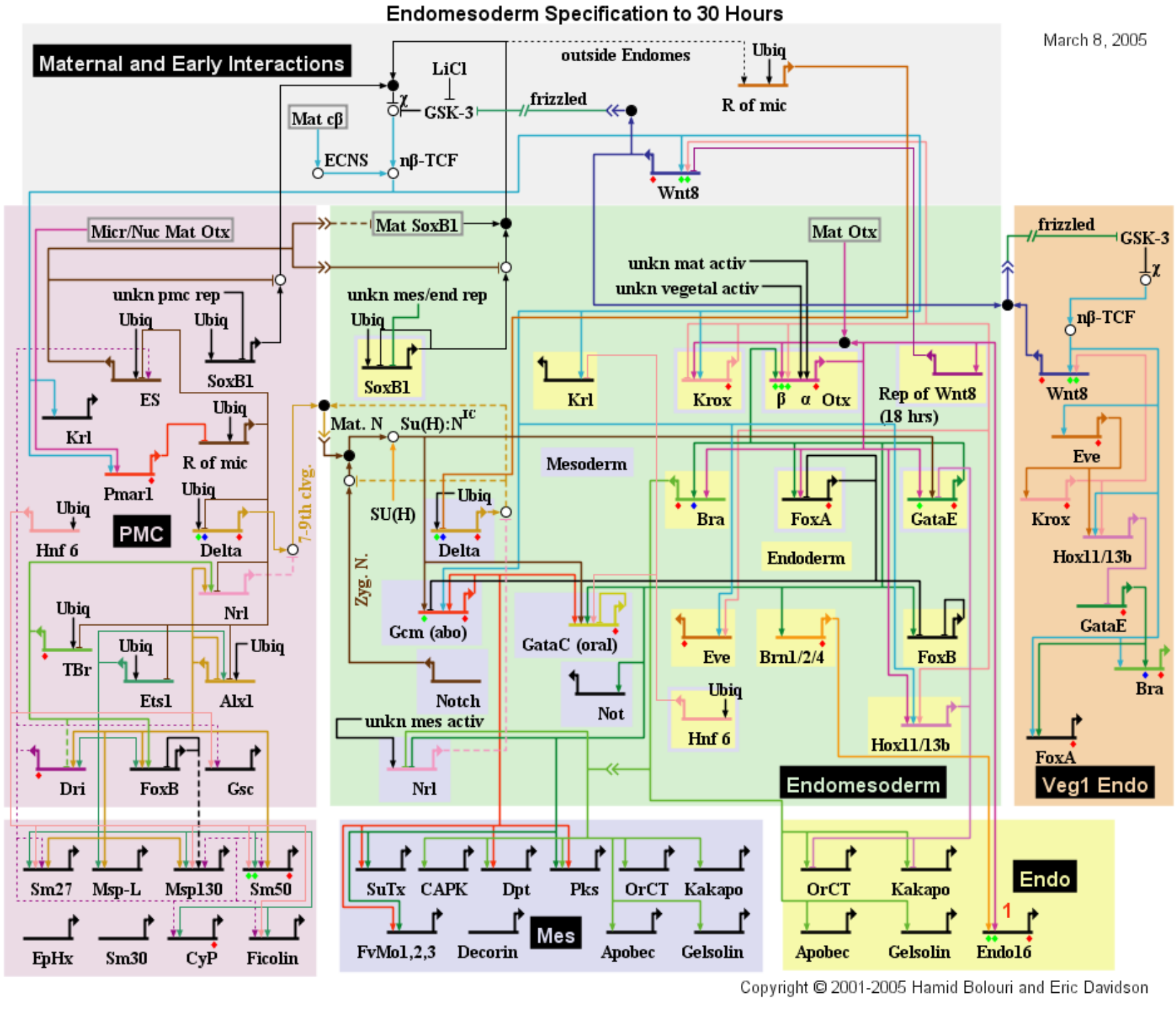
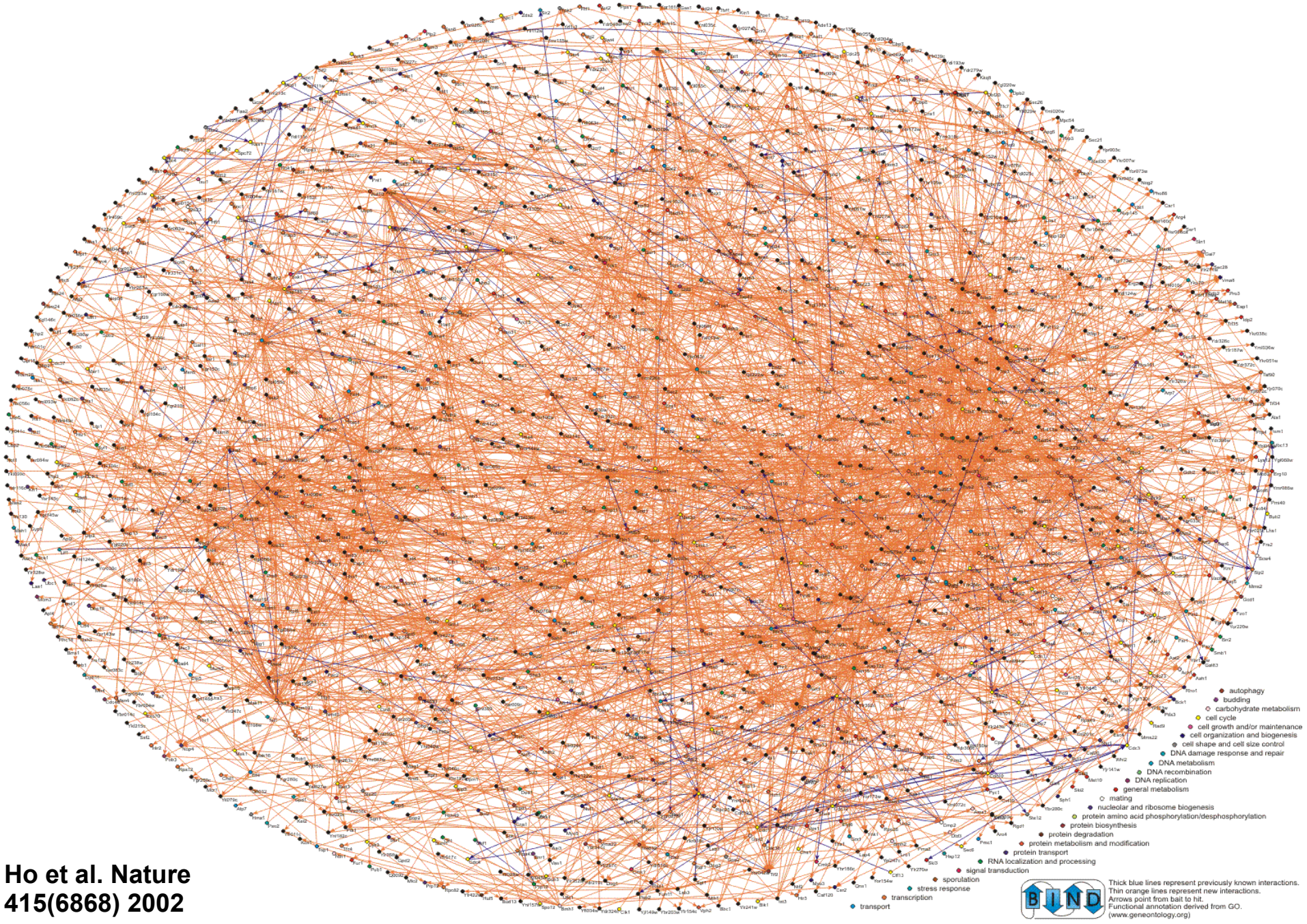


Figure 6A: The Cyclin - E2F cell cycle control system (version 3a - June 8, 1999)

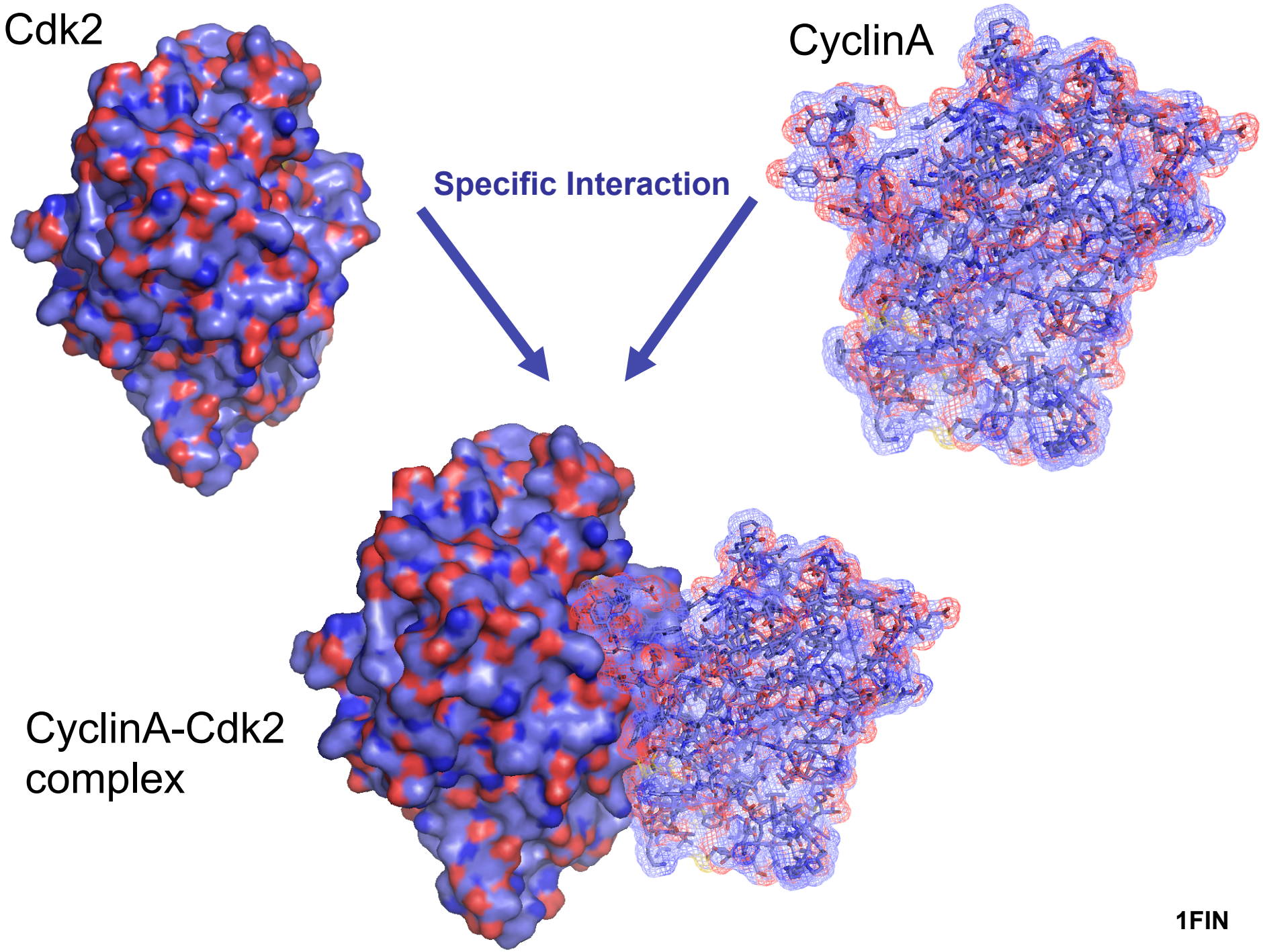
http://discover.nci.nih.gov/kohnk/interaction_maps.html


Endomesoderm Specification to 30 Hours
March 8, 2005
Maternal and Early Interactions
LiCl
GSK-3
Mat cβ
ECNS
nβ-TCF
outside Endomes
Ubiq
R of mic
frizzled
Wnt8
Micr/Nuc Mat Otx
Mat SoxB1
Mat Otx
unkn mat activ
unkn vegetal activ
unkn pmc rep
unkn mes/end rep
SoxB1
ES
Krl
Pmar1
R of mic
PMC
Hnf 6
Delta
Nrl
TBr
Ets1
Alx1
Dri
FoxB
Gsc
7-9th clvg.
Mat. N
Su(H):N^IC
SU(H)
Zyg. N.
Gcm (abo)
GataC (oral)
Notch
Not
unkn mes activ
Nrl
Krox
β α Otx
Rep of Wnt8
(18 hrs)
Mesoderm
Bra
FoxA
GataE
Endoderm
Eve
Brn1/2/4
FoxB
Hnf 6
Hox11/13b
Endomesoderm
Veg1 Endo
Sm27
Msp-L
Msp130
Sm50
EpHx
Sm30
CyP
Ficolin
SuTx
CAPK
Dpt
Pks
OrCT
Kakapo
FvMo1,2,3
Decorin
Mes
Apobec
Gelsolin
Endo
Endo16
1
Copyright © 2001-2005 Hamid Bolouri and Eric Davidson

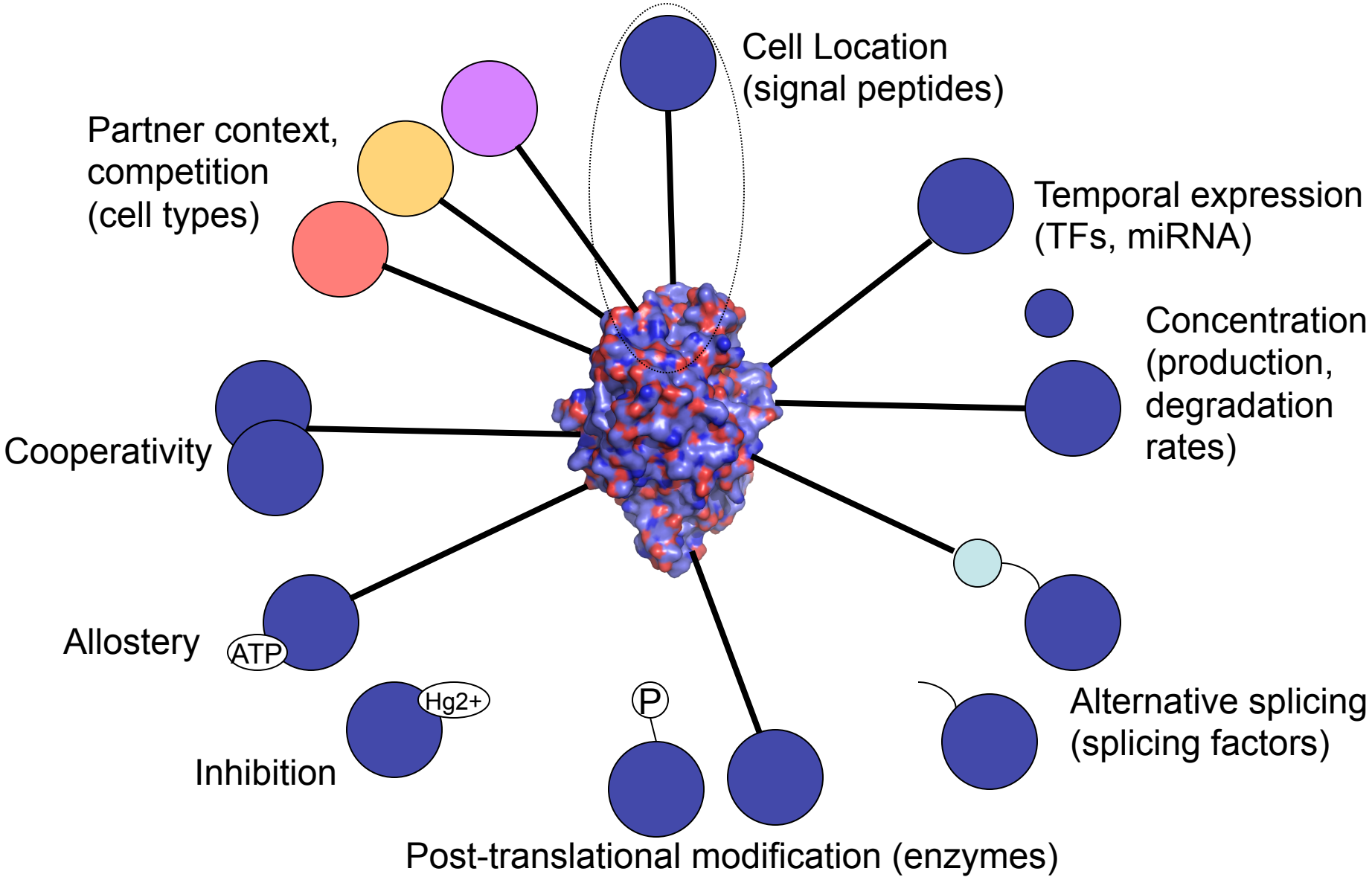
Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry



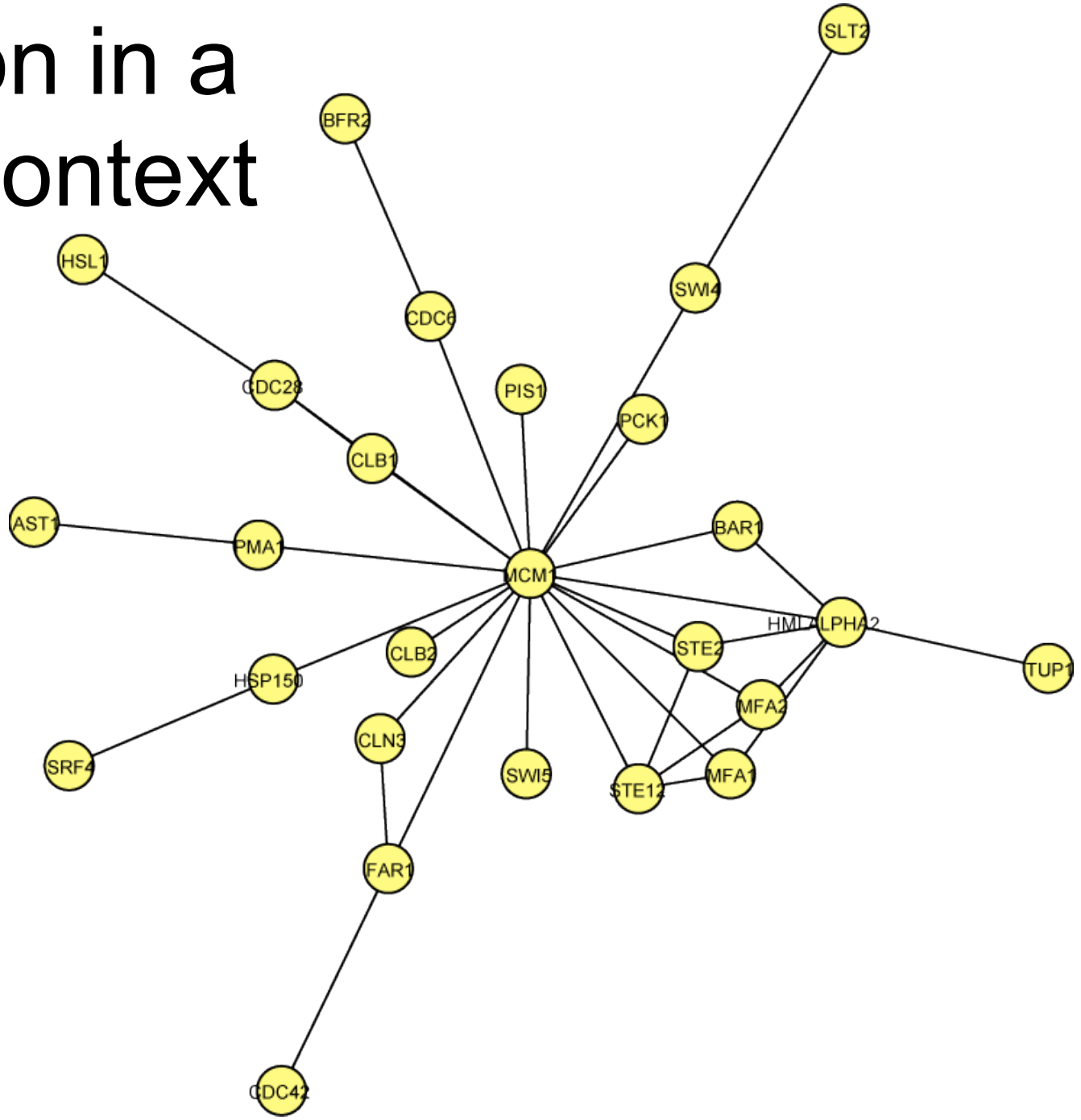
Ho et al. Nature
415(6868) 2002


Cdk2
CyclinA
Specific Interaction
CyclinA-Cdk2
complex
1FIN

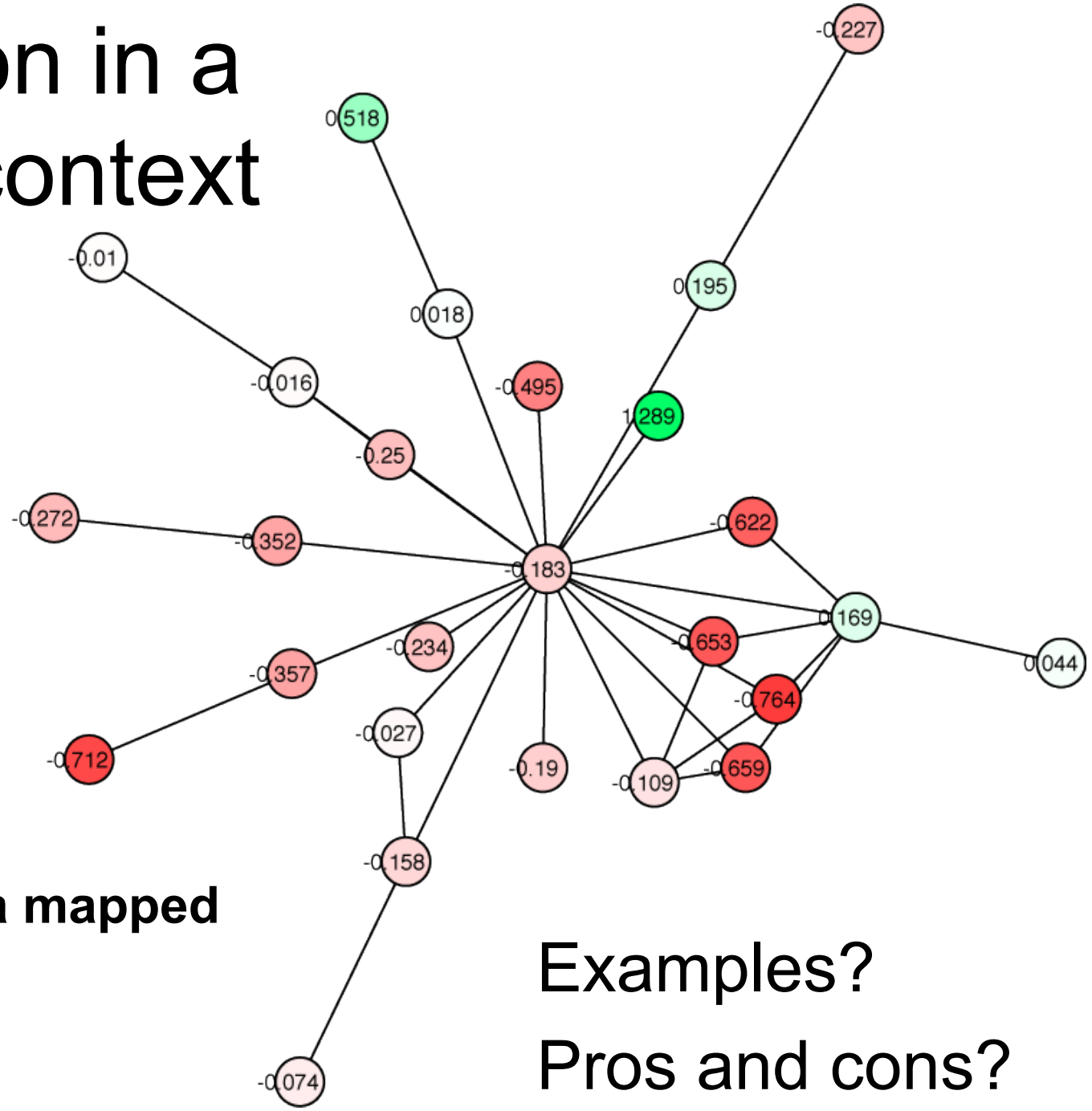
Cellular Context



Integration in a network context



Integration in a network context



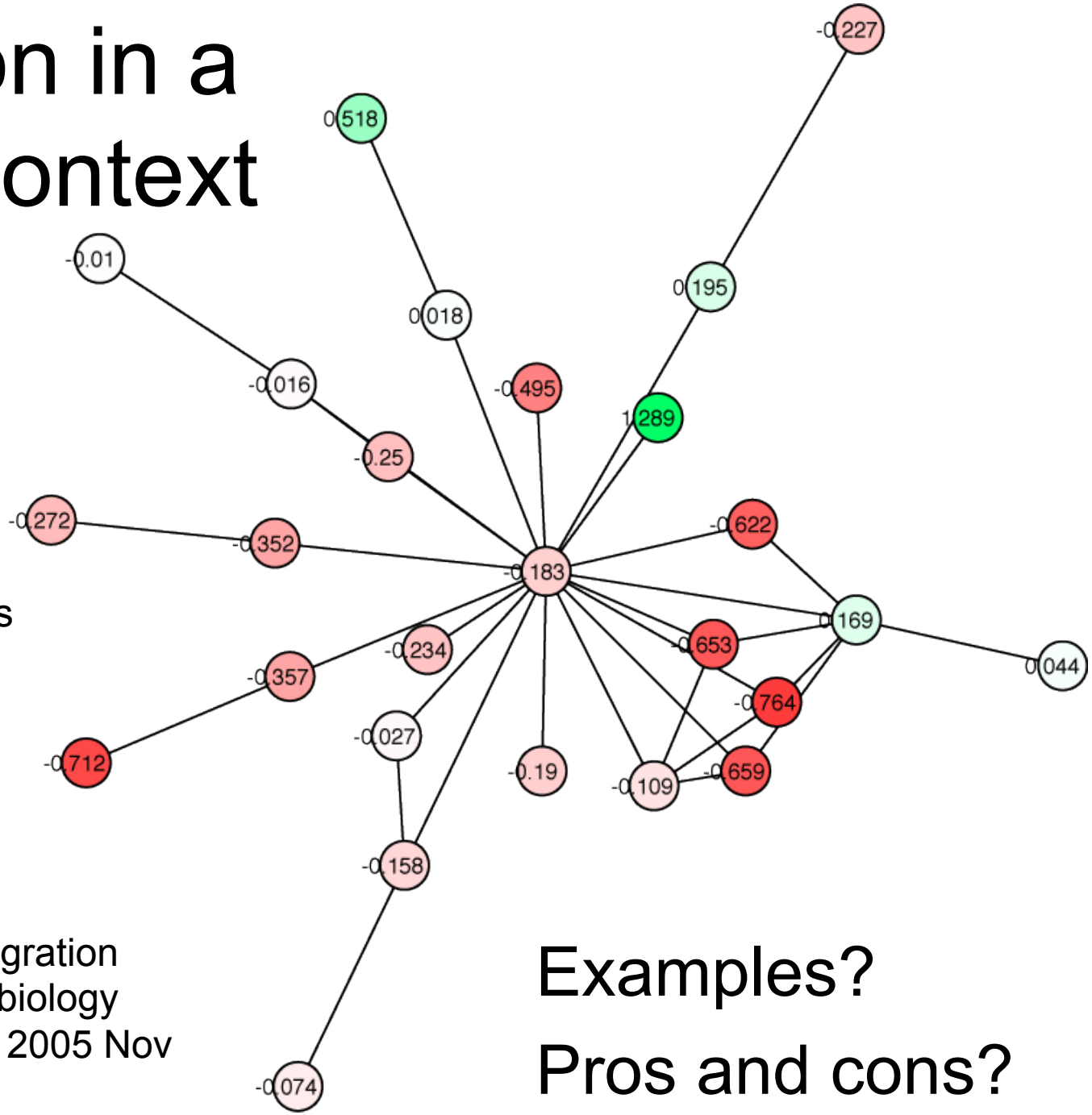
Expression data mapped to node colours

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

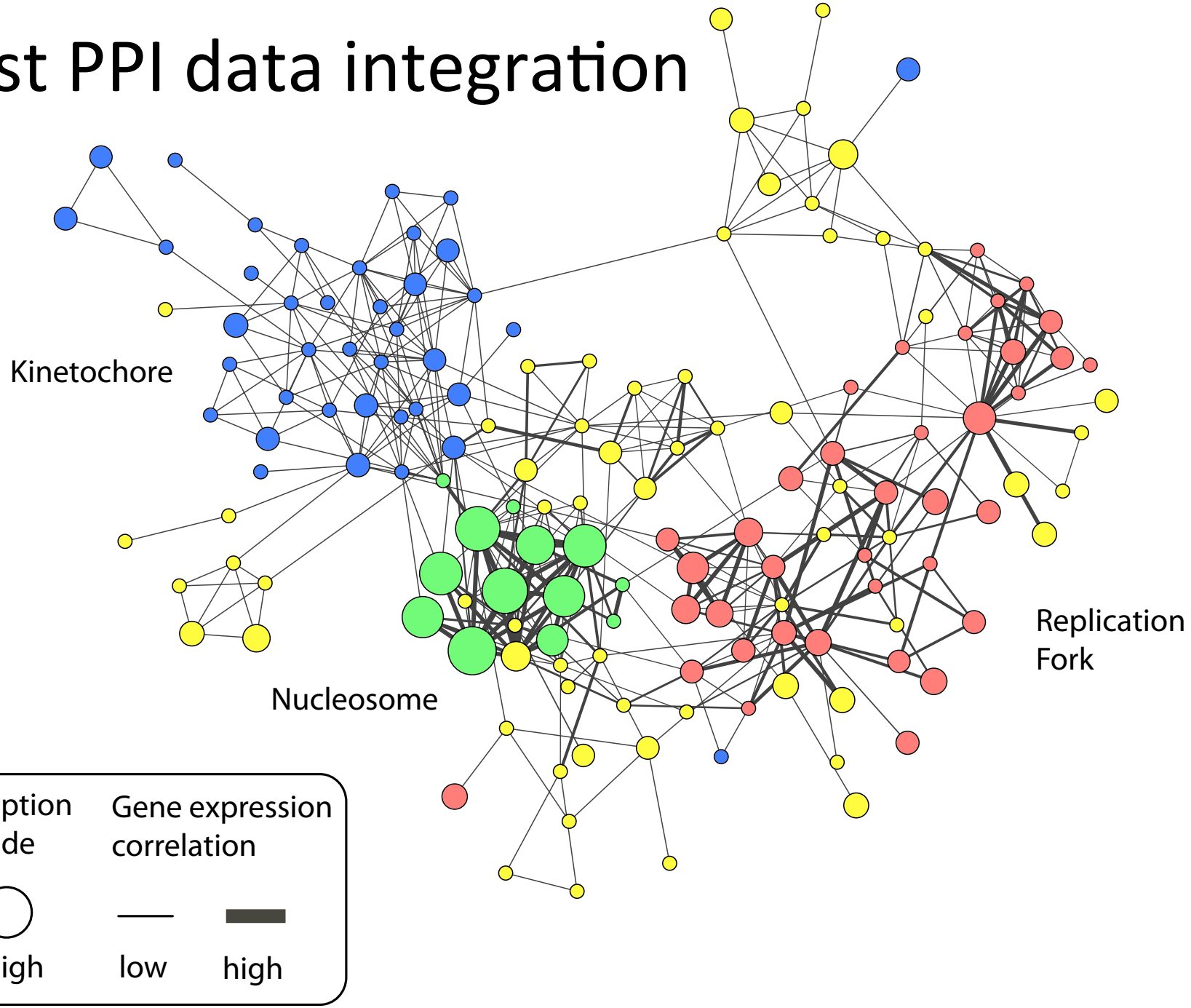


Hwang D et al. A data integration methodology for systems biology Proc Natl Acad Sci U S A. 2005 Nov 29;102(48):17296-301

Examples?

Pros and cons?

Yeast PPI data integration



Data Integration

Network Visualization and Analysis

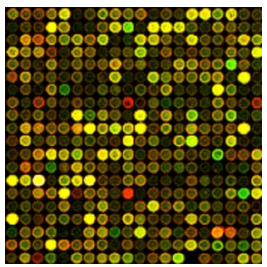
Why Network Analysis?

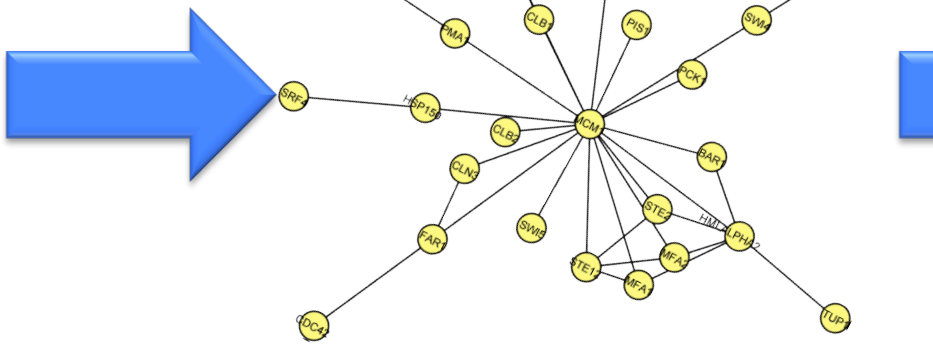
Intuitive to Biologists

- 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

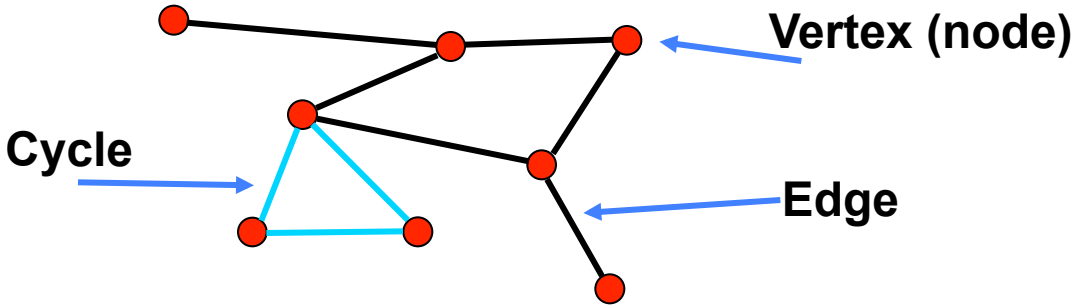
- Visualize multiple data types on a network
- Cluster, Find active pathways, Compare, Search

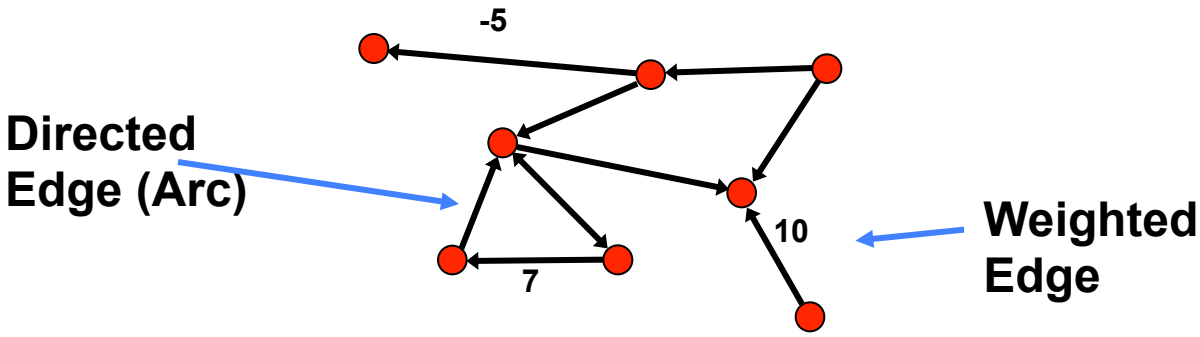




Eureka! New pathway gene!

Graph Theory

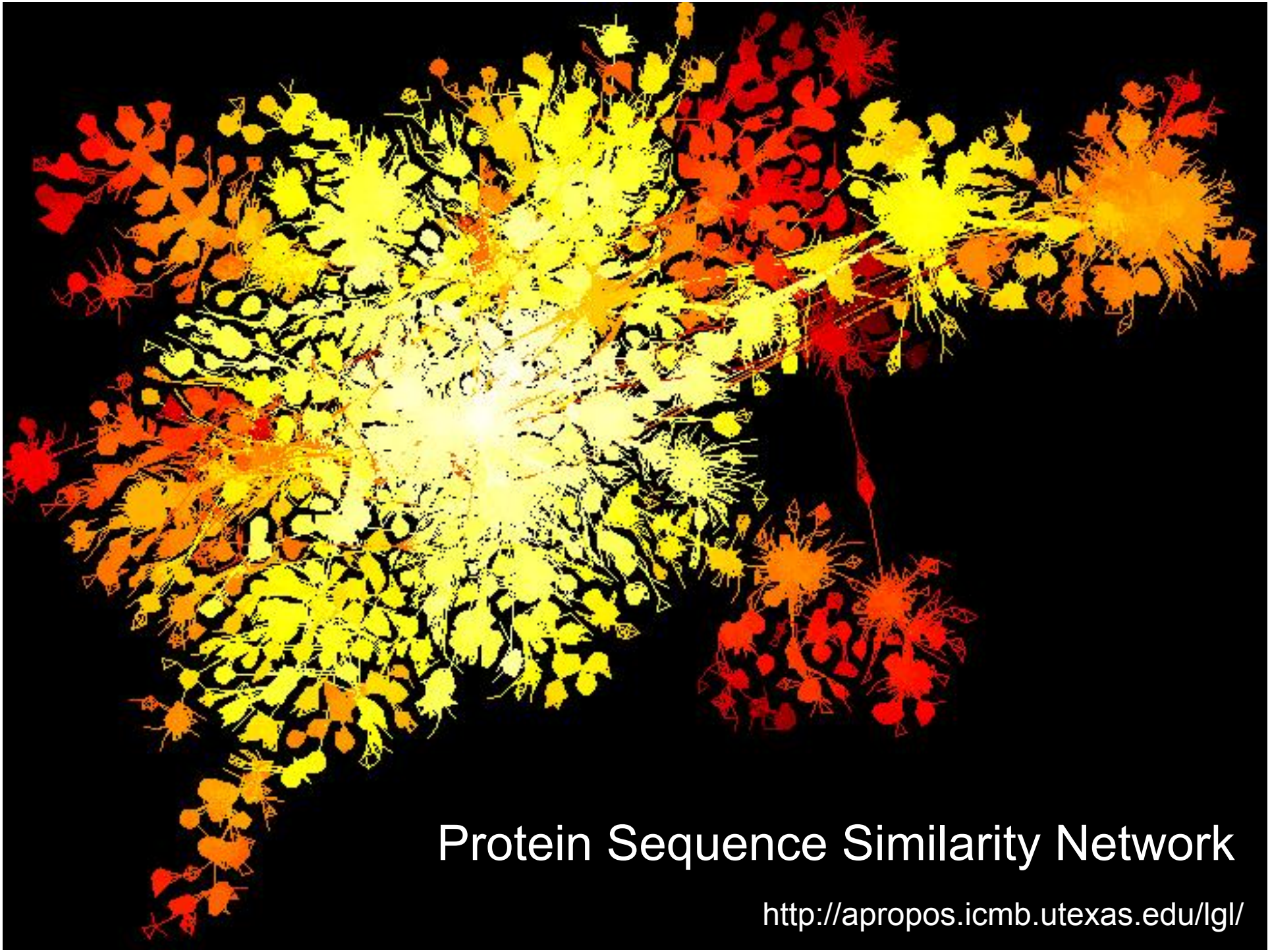




We map molecular interaction networks to graphs

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

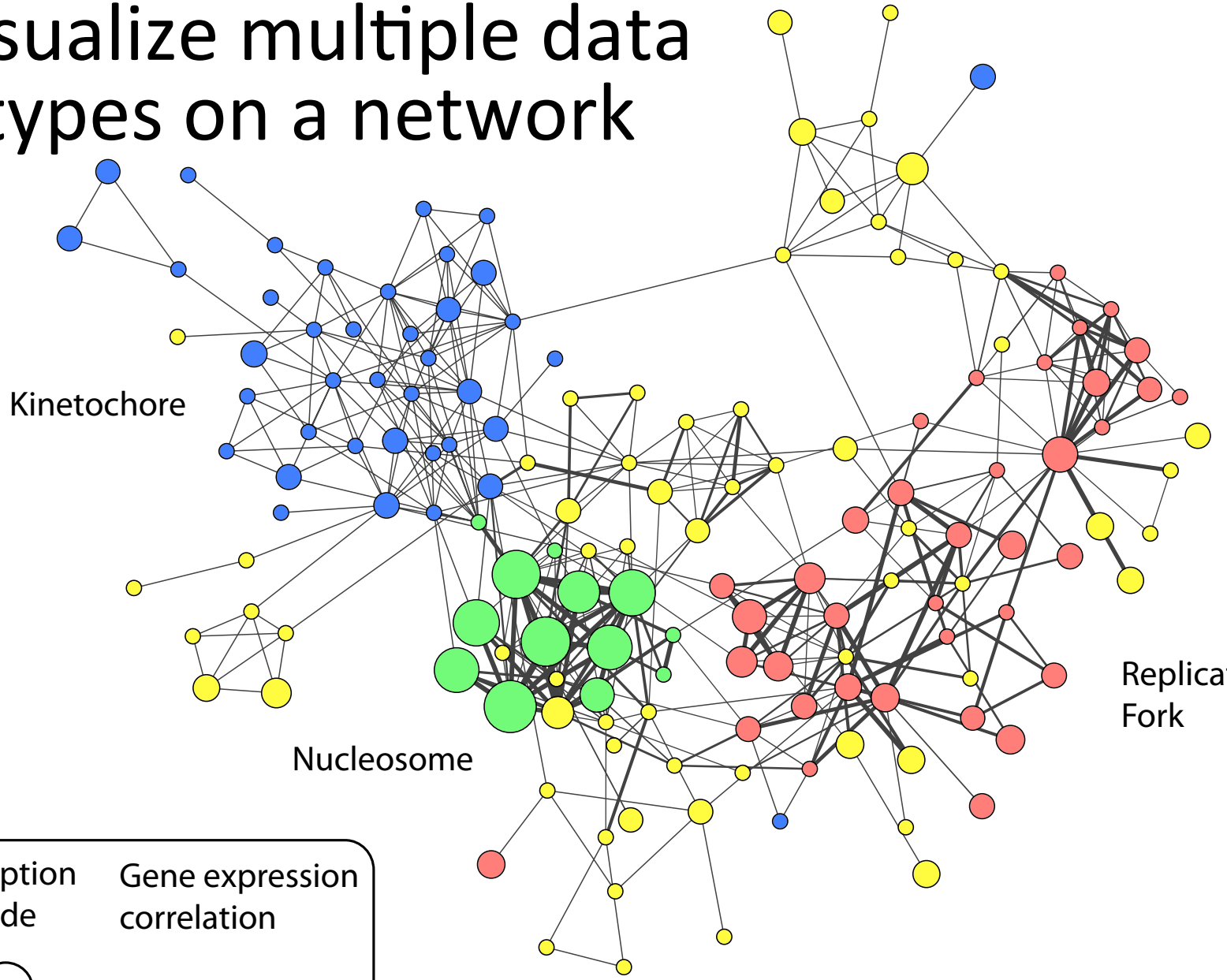

Protein Sequence Similarity Network
http://apropos.icmb.utexas.edu/lgl/

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



Transcription amplitude

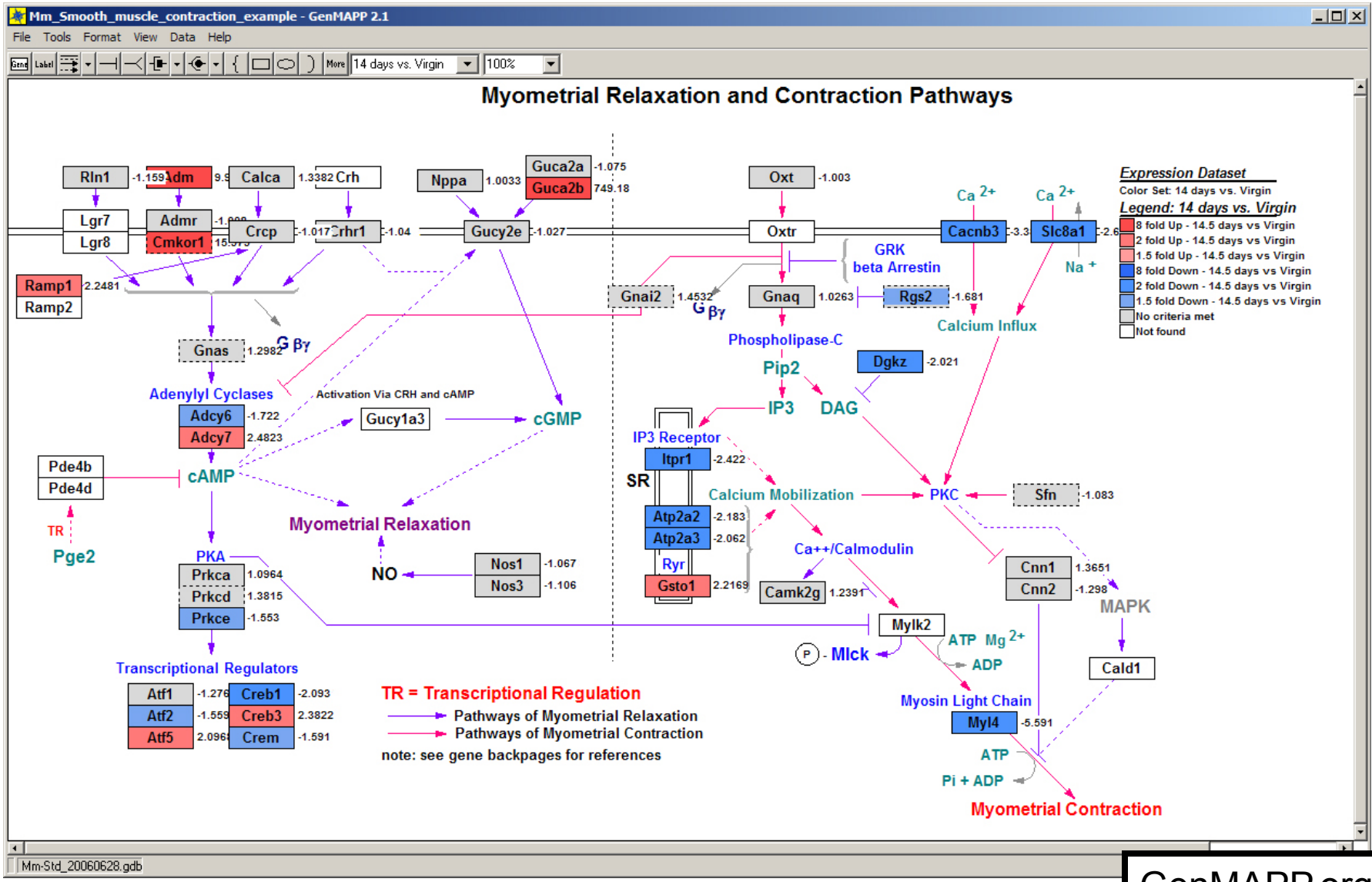
low high

Gene expression correlation

low high

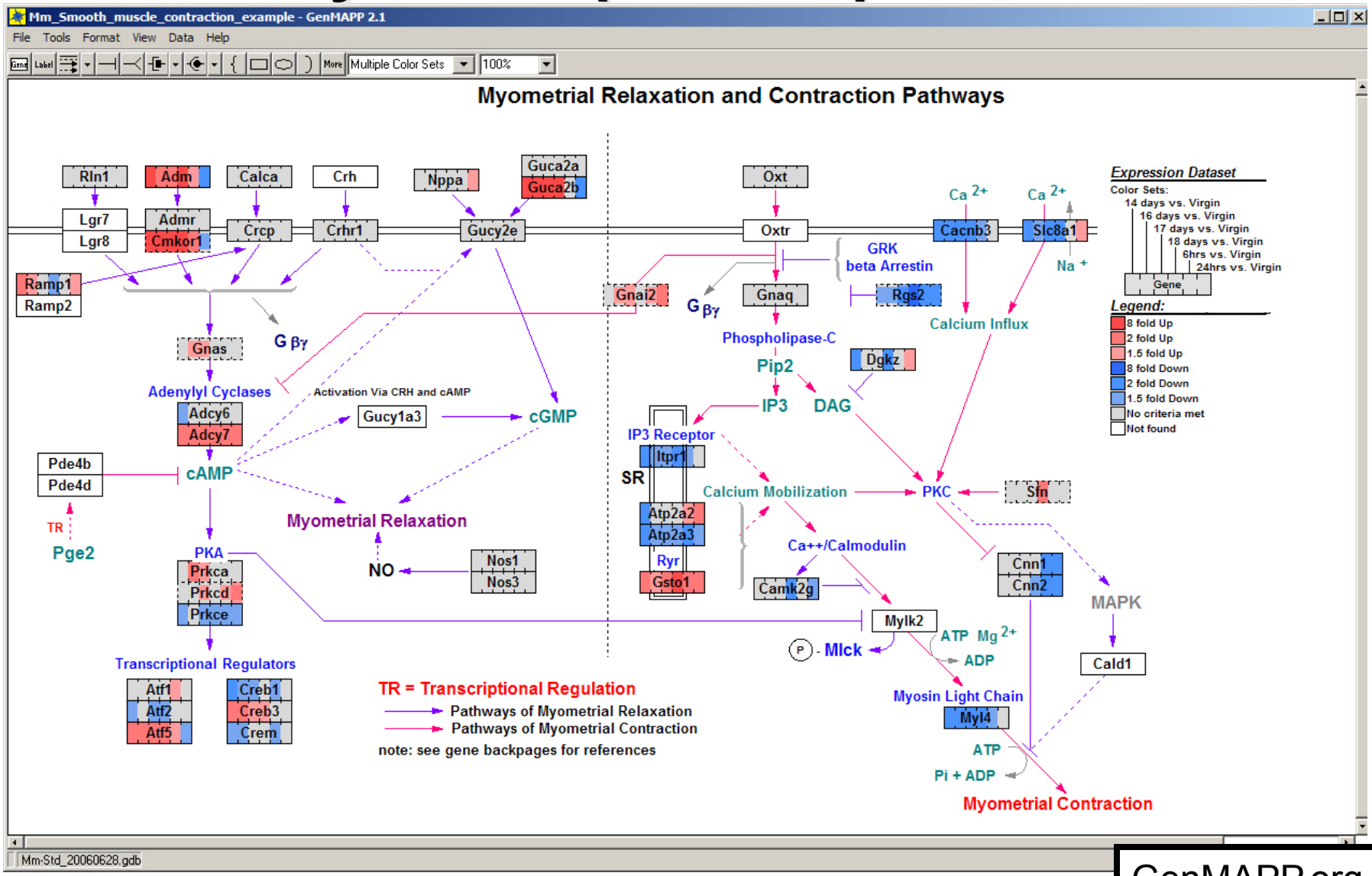
Control: node/edge size, shape, color...

Visualizing Time Course Data on Pathways: Single Comparison View



GenMAPP.org

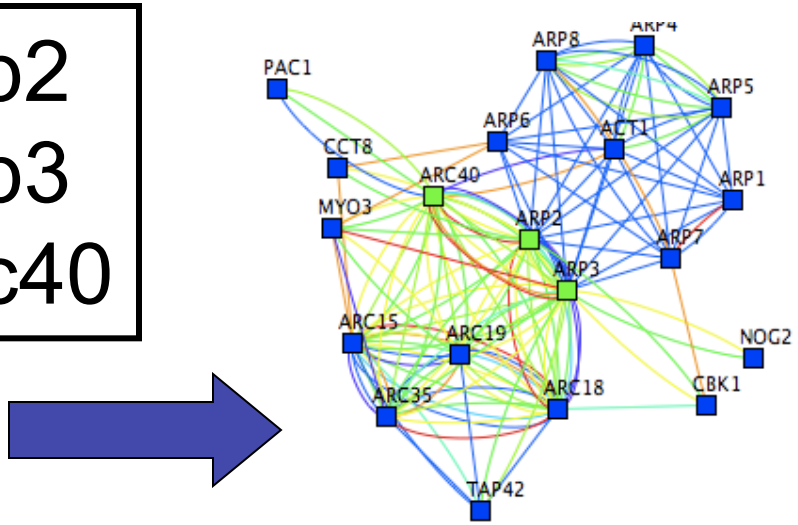
Visualizing Time Course Data on Pathways: Multiple Comparison View



GenMAPP.org

Predicting Gene Function

arp2
arp3
arc40



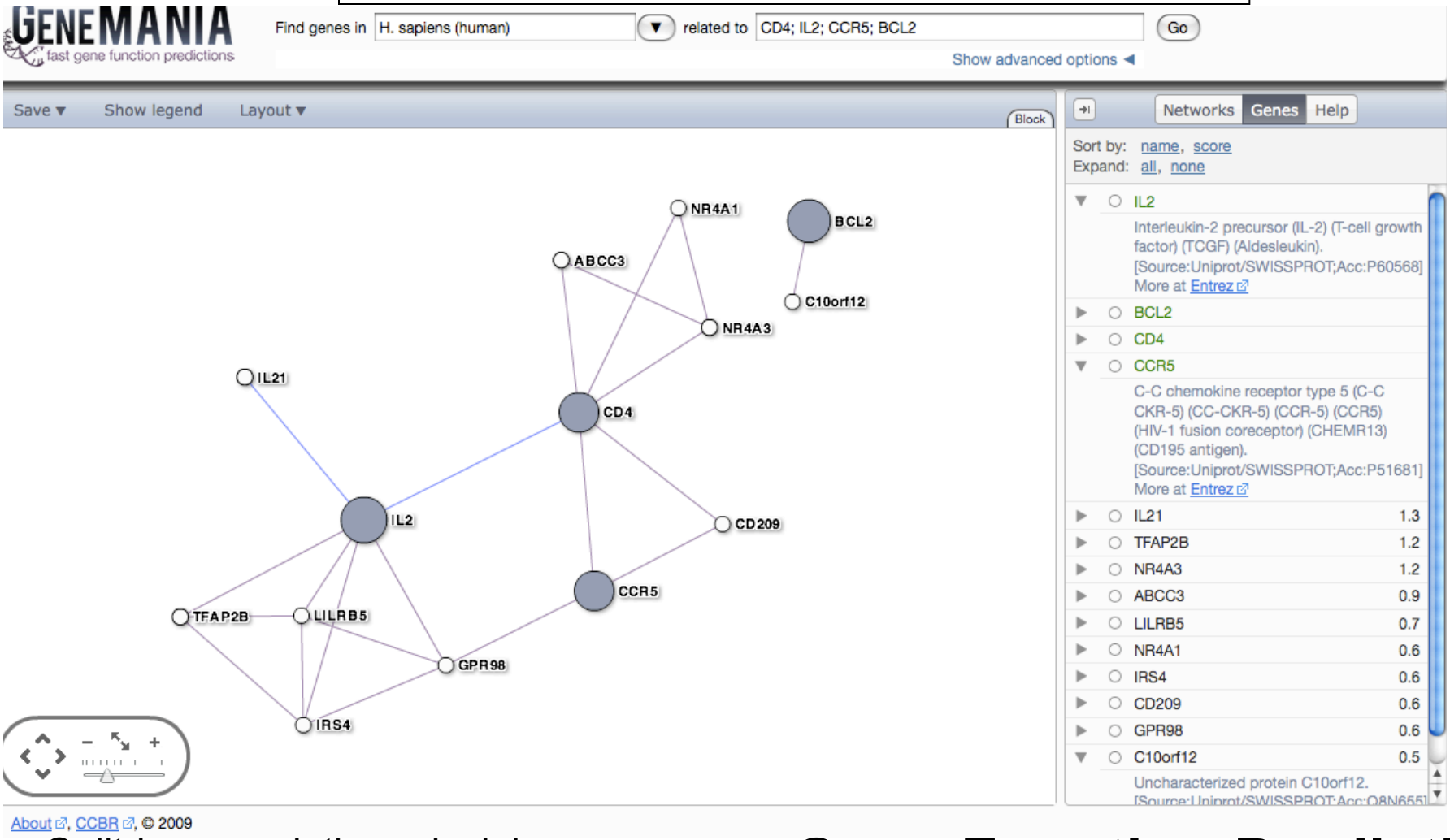
- STRING
  - http://string.embl.de/
- bioPIXIE
  - http://pixie.princeton.edu/pixie/
- GeneMania
  - http://www.genemania.org

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



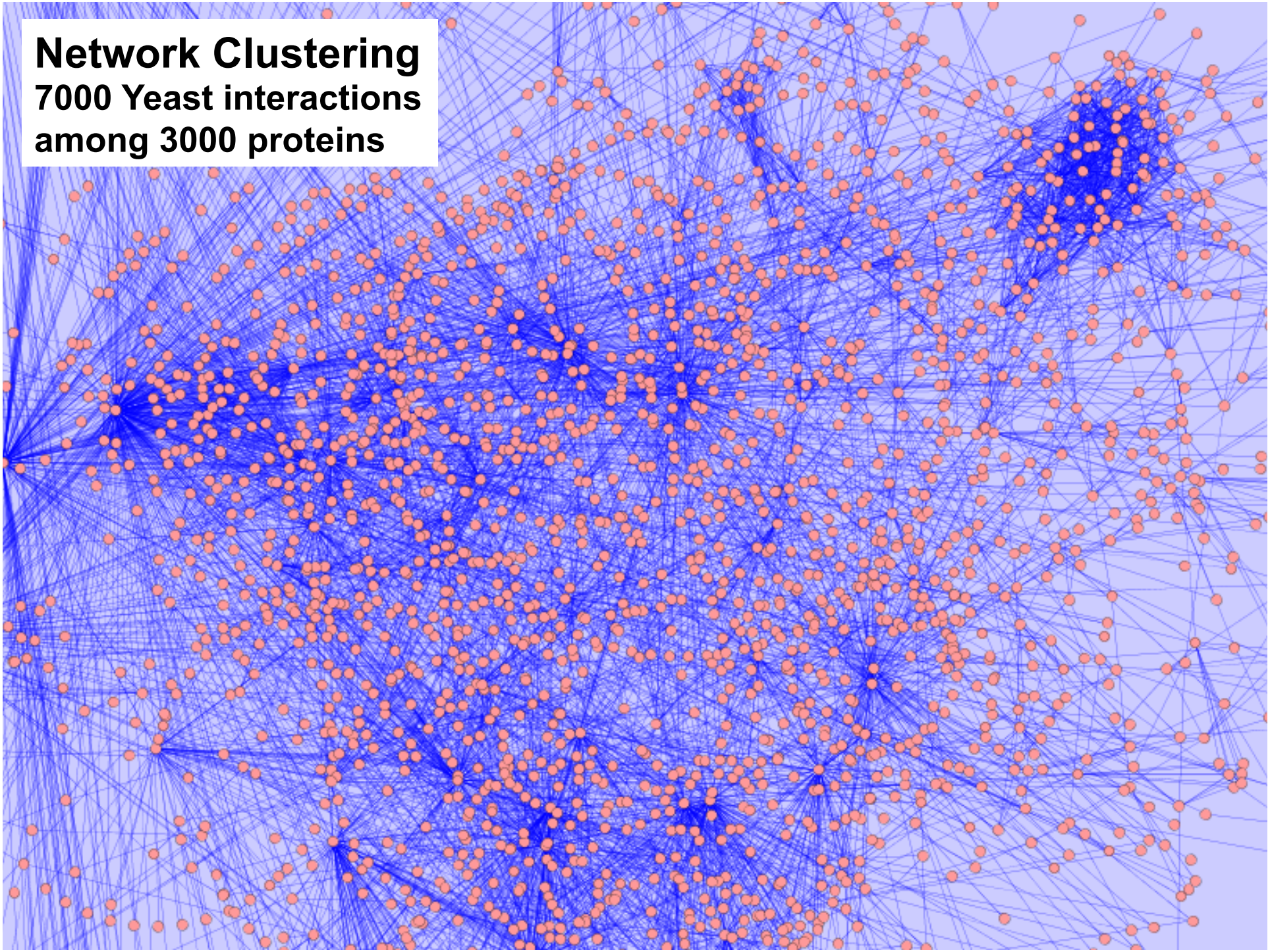
Gene Function Prediction

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

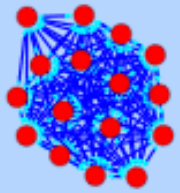
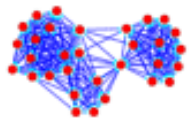
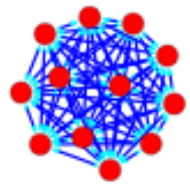
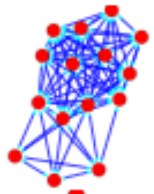
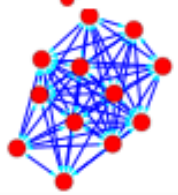
Quaid Morris (CCBR)
Rashad Badrawi, Ovi Comes, Sylva Donaldson, Christian Lopes, Farzana Kazi, Jason Montojo, Harold Rodriguez, Khalid Zuberi

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

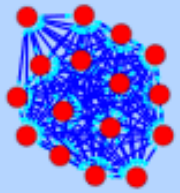
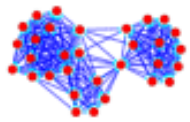
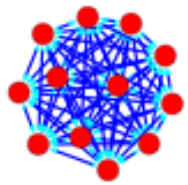
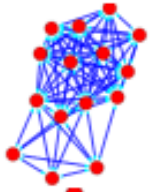
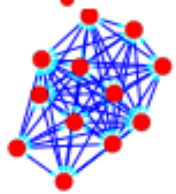
MCODE Results Summary

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

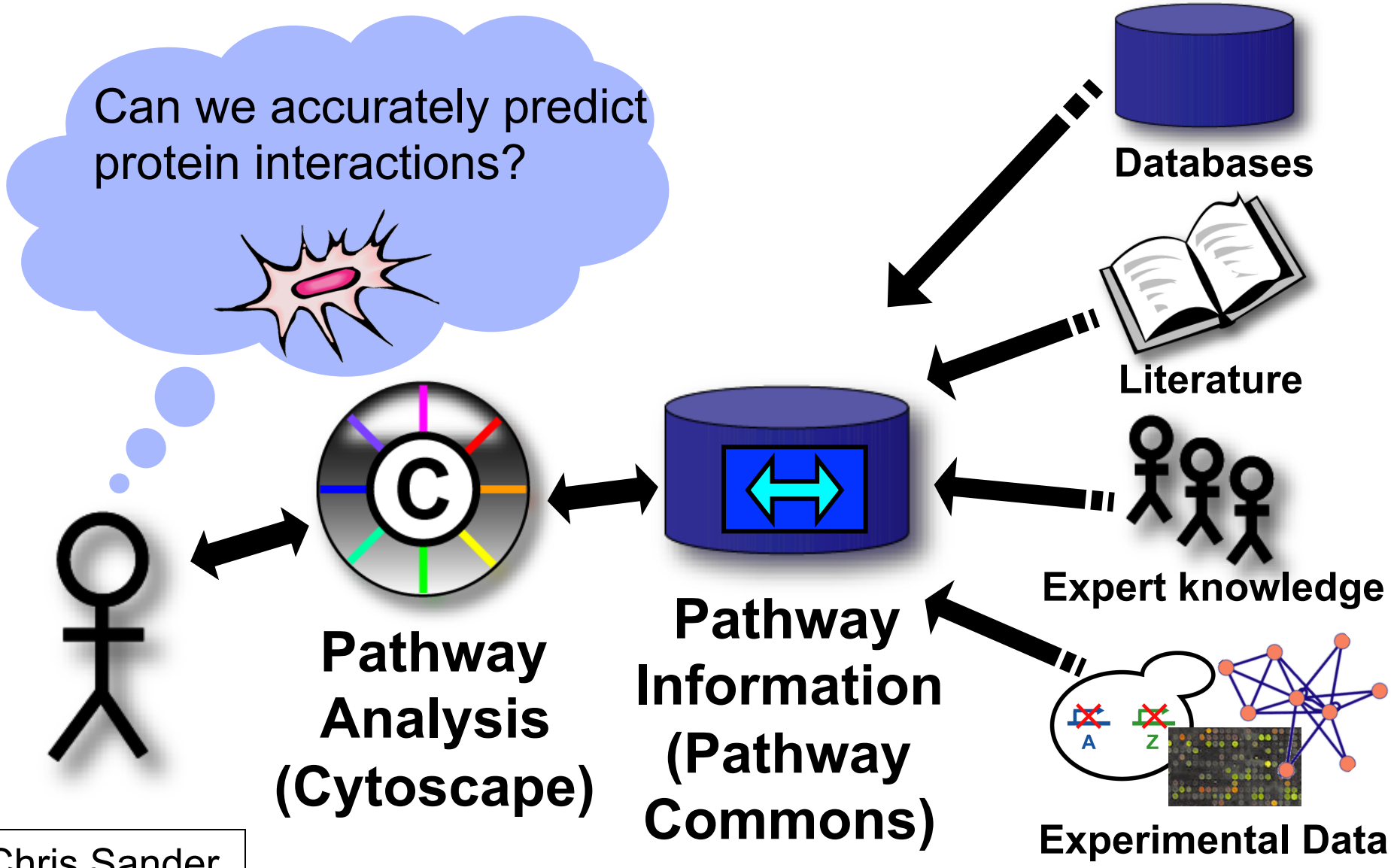
Rank	Score	Size	Names	Complex
1	7.25	16,116	YGR232W, YDL007W, YKL145W, YFR052W, YFR004W, YLR421C, YOR261C, YDL147W, YDR427W, YHR200W, YER021W, YOR117W, YDL097C, YOR259C, YPR108W, YDR394W	
Proteasome 26S				
2	6.387	31,198	YPL093W, YBL004W, YOR272W, YNL110C, YKL009W, YFL002C, YOL077C, YPL126W, YIL035C,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

Create a new child network. Save Done

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

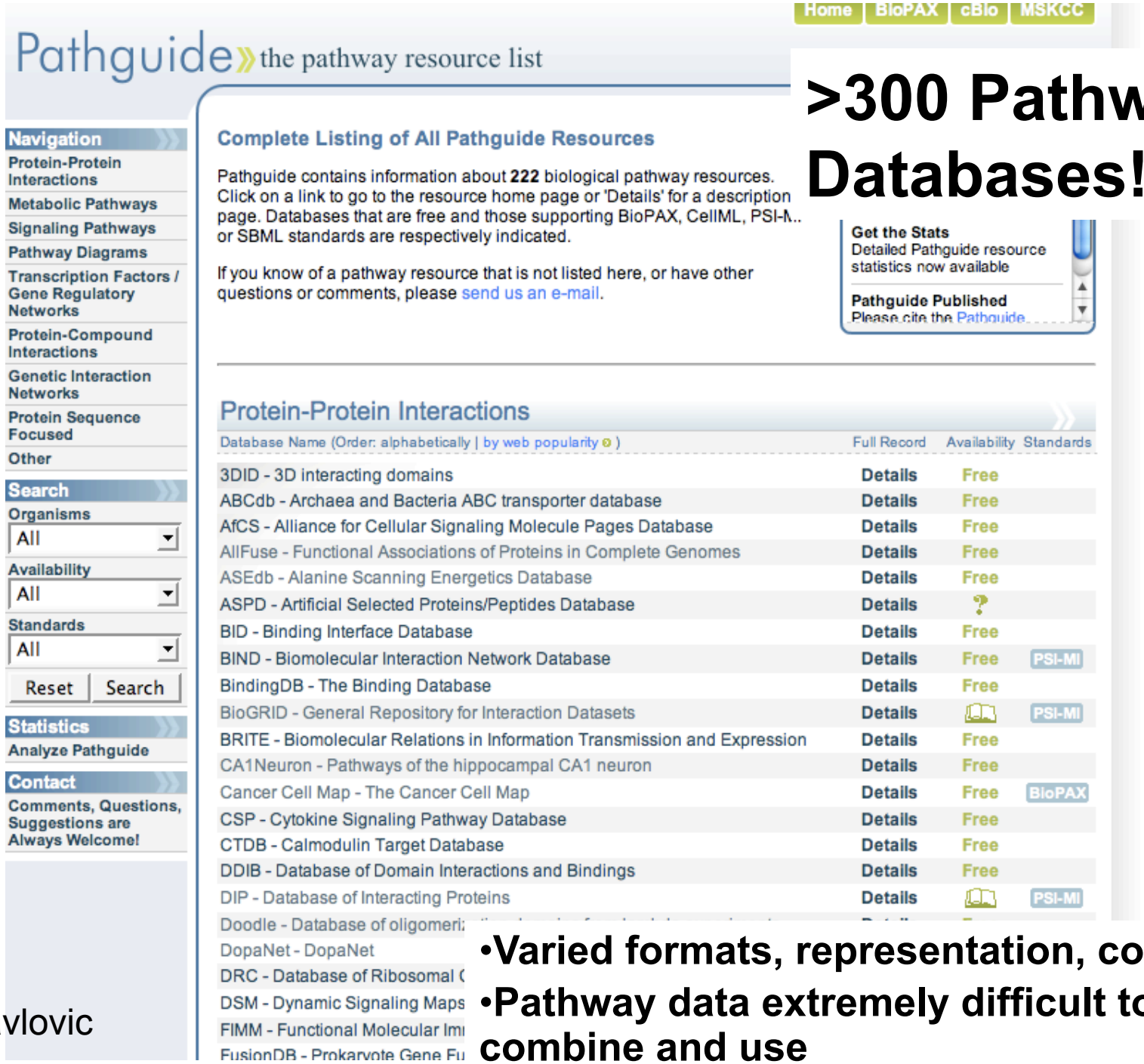
Network Data

Cell map exploration and analysis



Chris Sander

http://pathguide.org

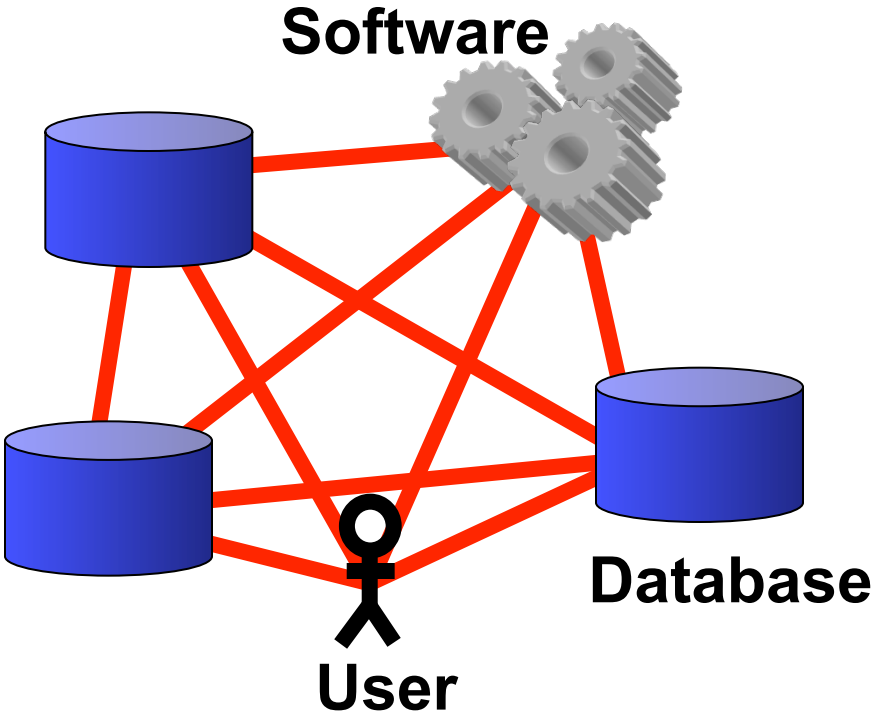


>300 Pathway Databases!

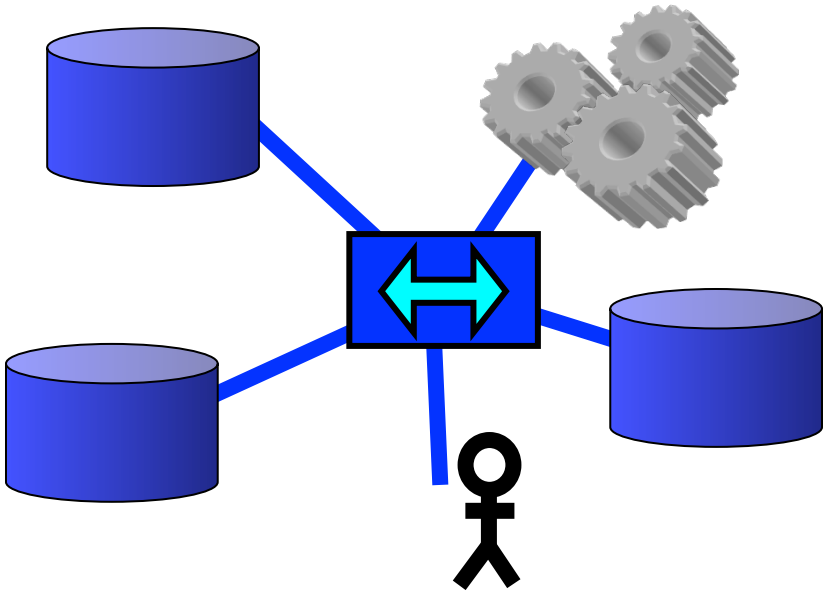
- **Varied formats, representation, coverage**
- **Pathway data extremely difficult to combine and use**

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

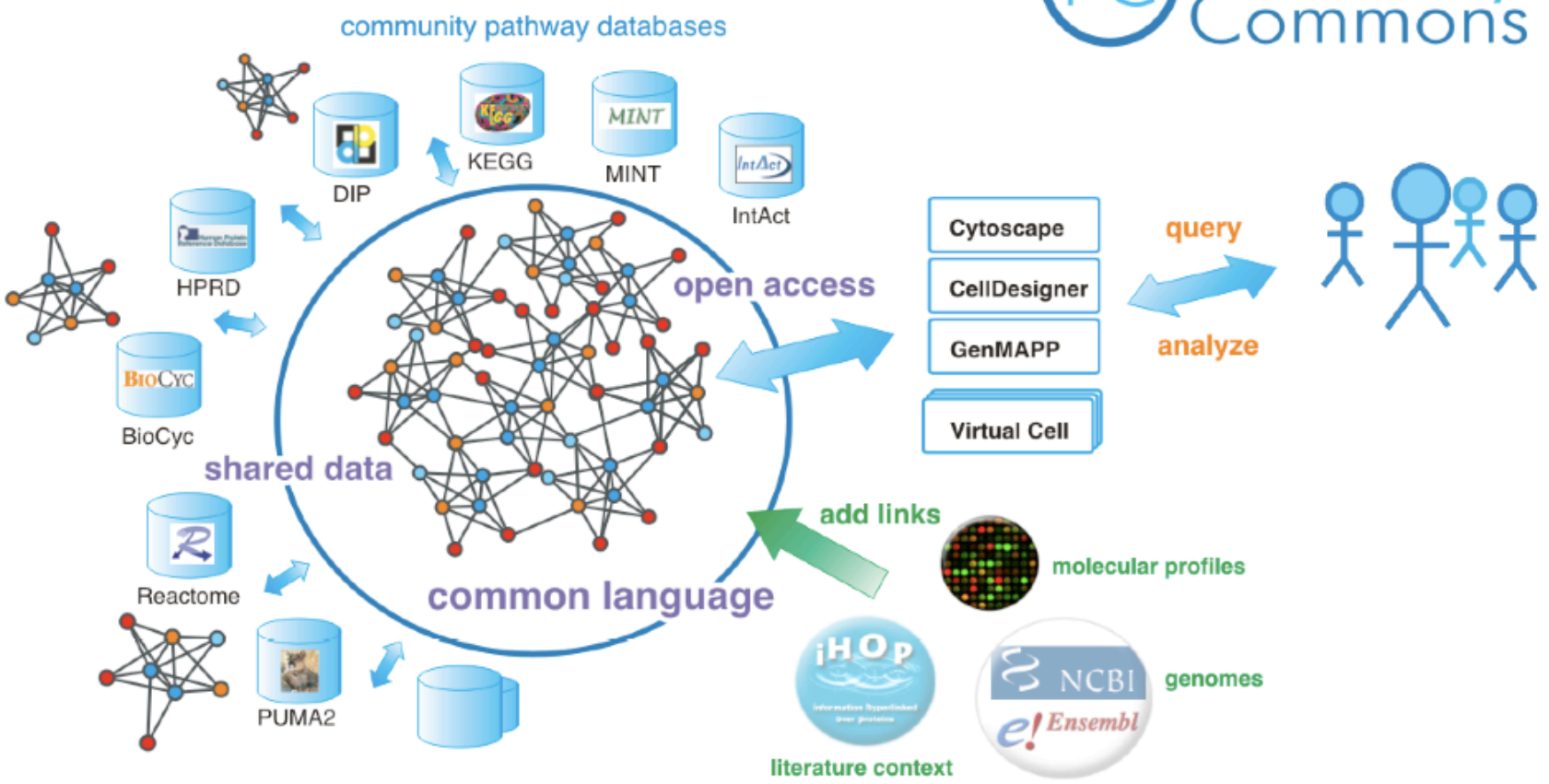
- PSI-MI http://psidev.sourceforge.net
  - Molecular interactions - protein-protein interaction focus
  - Peer reviewed, HUPO community standard
- BioPAX http://www.biopax.org
  - Biological pathways
  - Community ontology in OWL, Protégé
- SBML http://www.sbml.org
  - Widely adopted for representing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

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

Aim: Convenient Access to Pathway Information

http://www.pathwaycommons.org



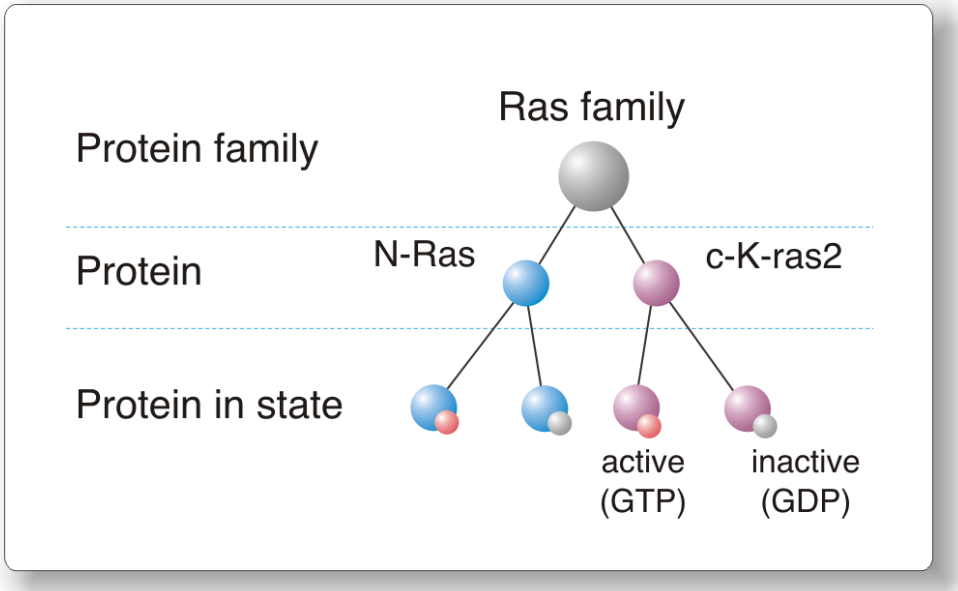
Facilitate creation and communication of pathway data
Aggregate pathway data in the public domain
Provide easy access for pathway analysis

Long term: Converge to integrated cell map

http://pathwaycommons.org

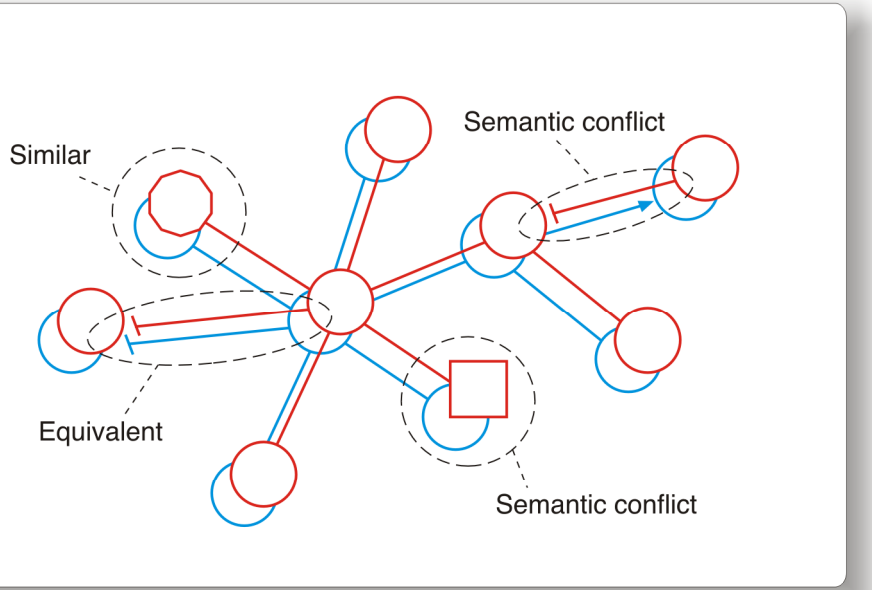
Towards an Integrated Cell Map

- Semantic pathway integration is difficult



Physical entities

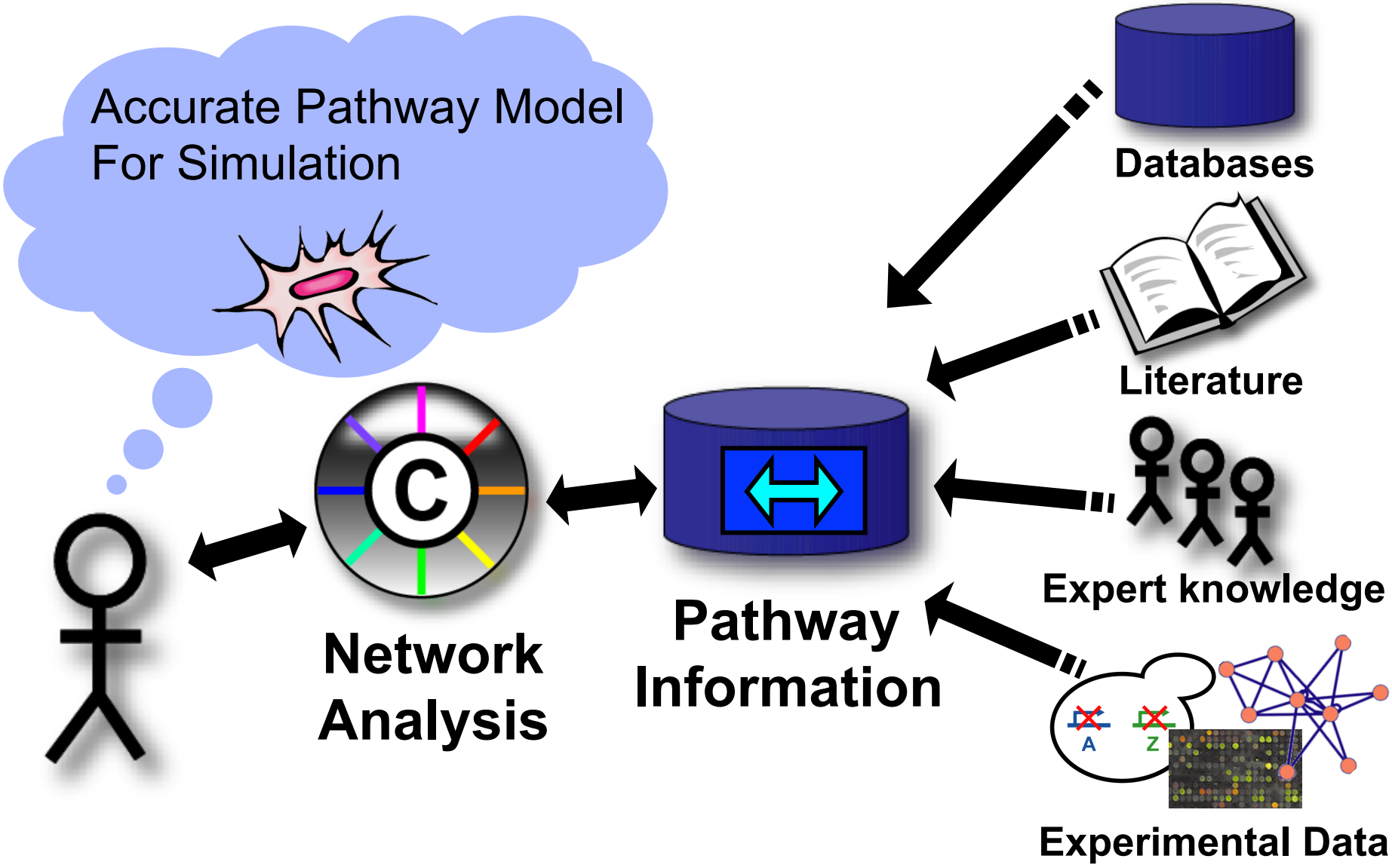
Relationships



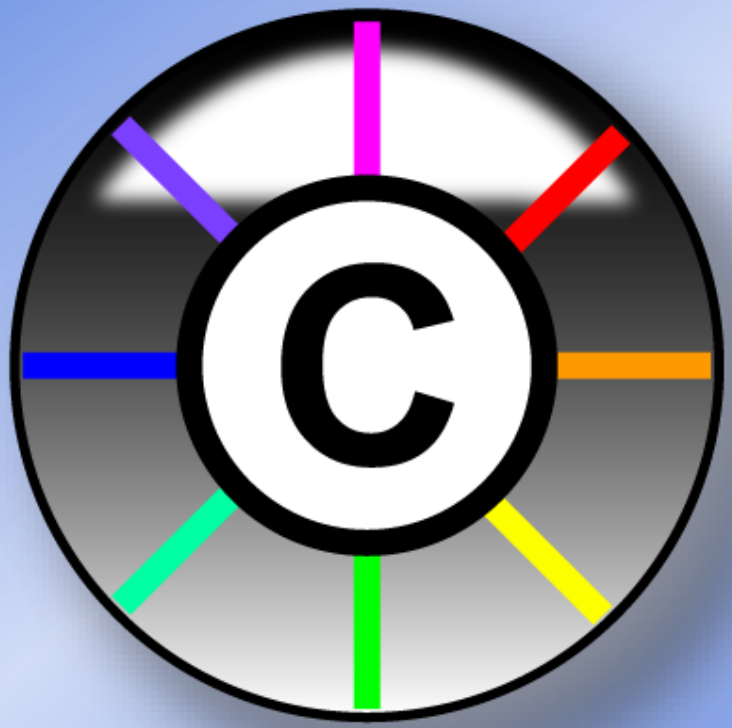
Determining equivalent entities is critical

Network Visualization and Analysis using Cytoscape

Using Pathway Information



Cytoscape



UNIVERSITY OF TORONTO

UCSF

NCIBI
National Center for Integrative Biomedical Informatics

Unilever

MEMORIAL SLOAN-KETTERING CANCER CENTER
ESTABLISHED 1884

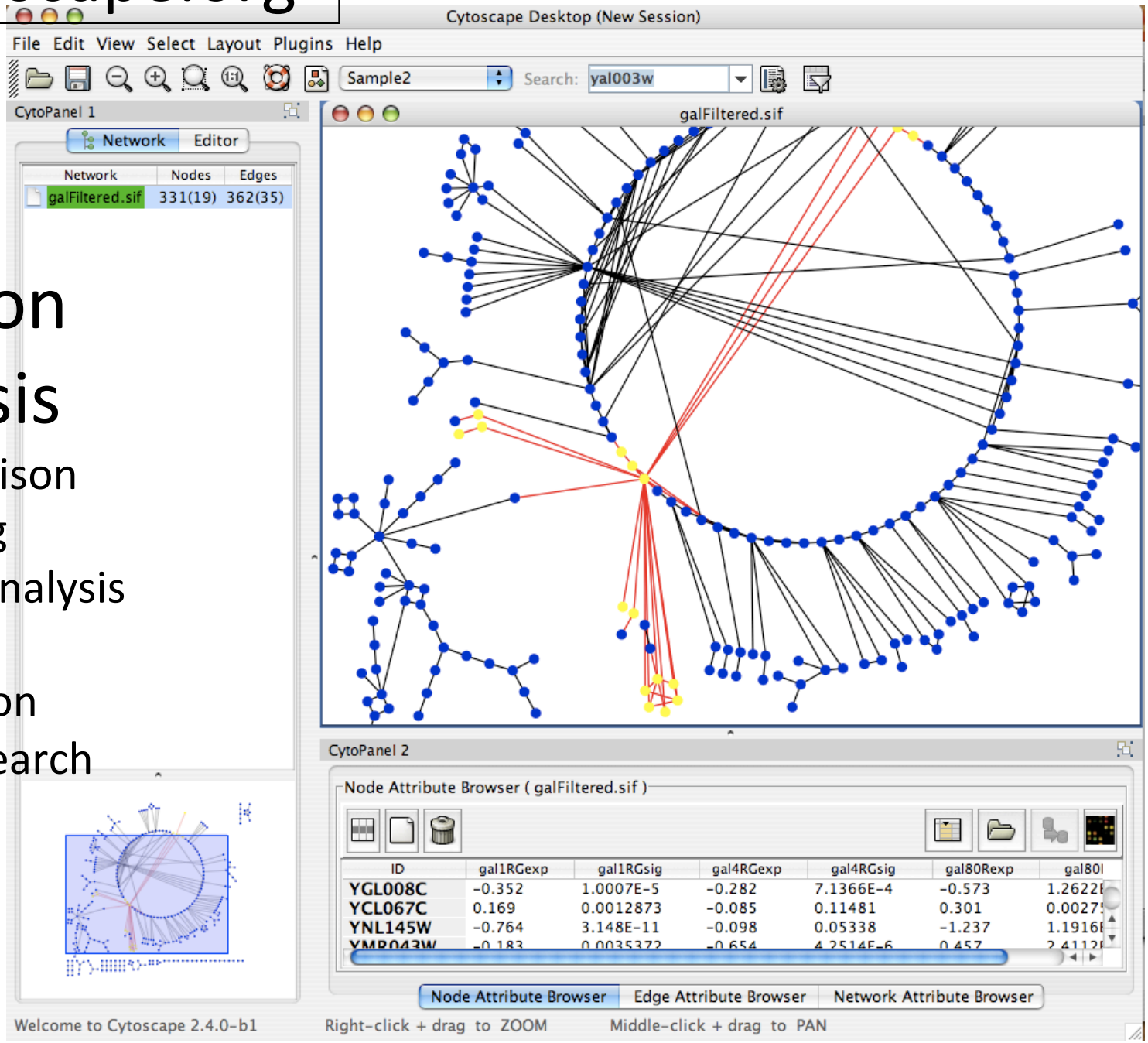
Agilent Technologies

http://cytoscape.org

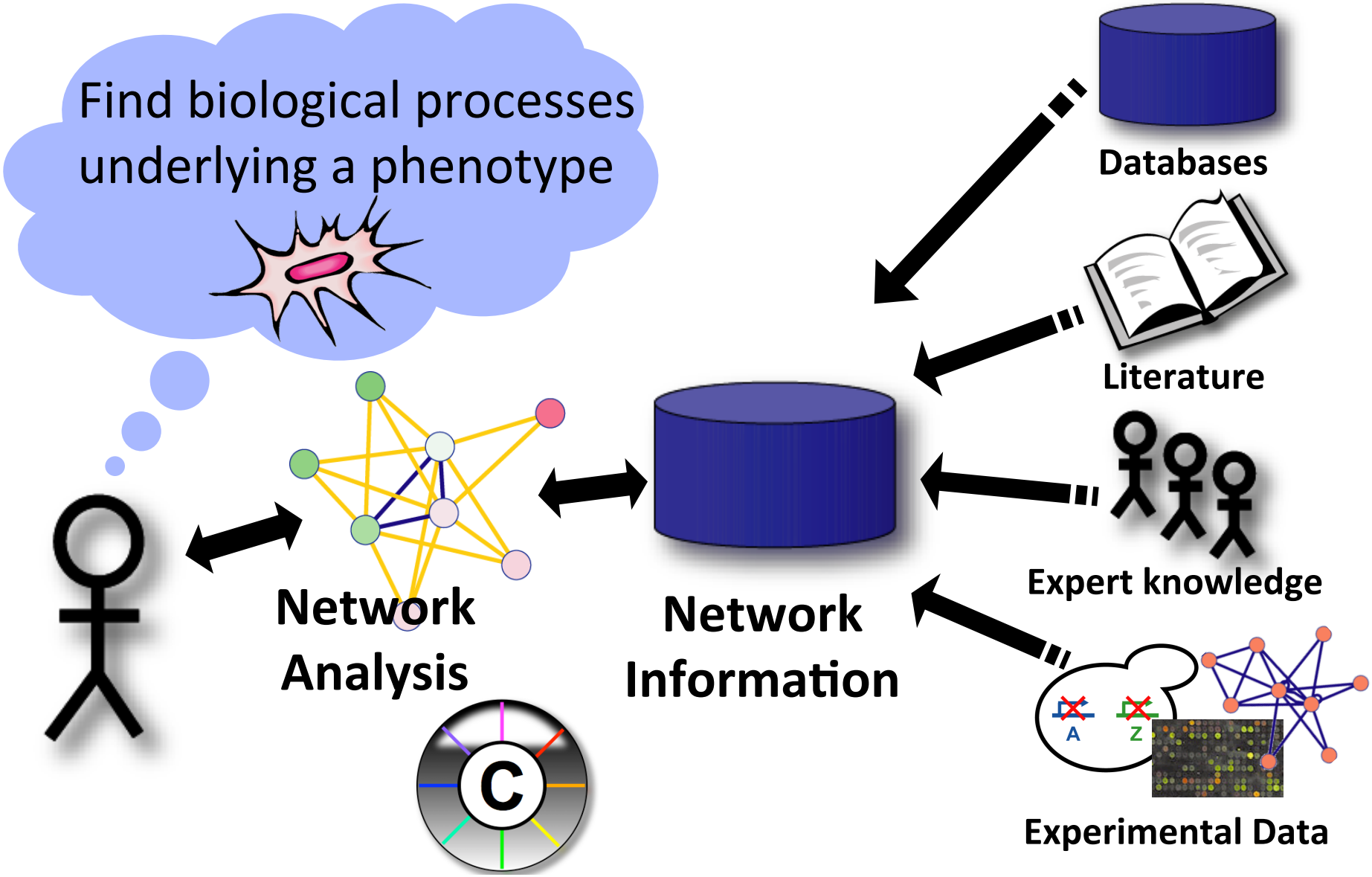
Network visualization and analysis

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

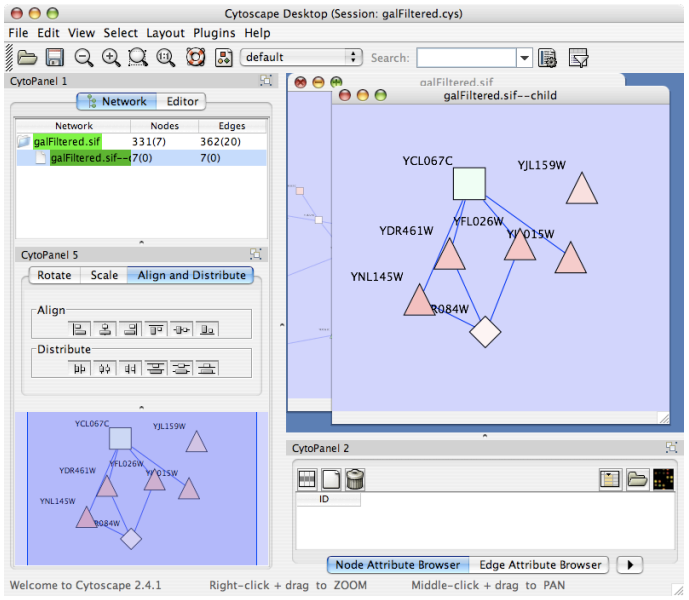
UCSD, ISB, Agilent, MSKCC, Pasteur, UCSF, Unilever, UToronto, U Michigan



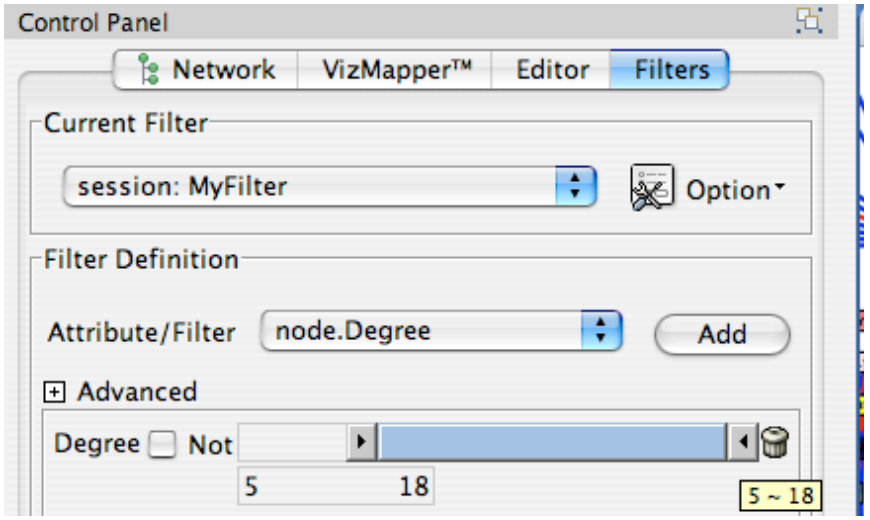
Network Analysis using Cytoscape



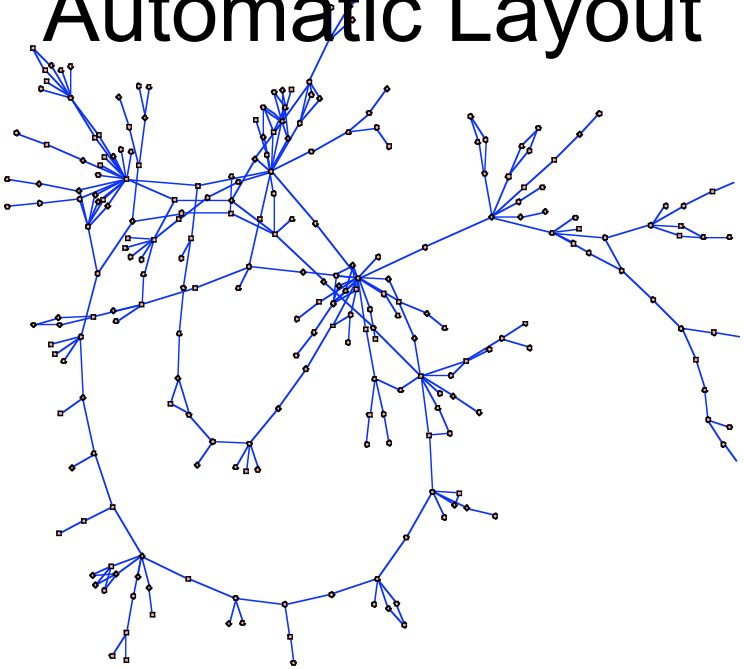
Manipulate Networks



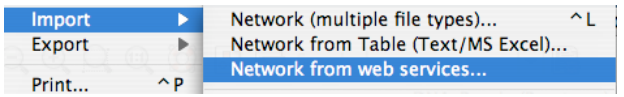
Filter/Query

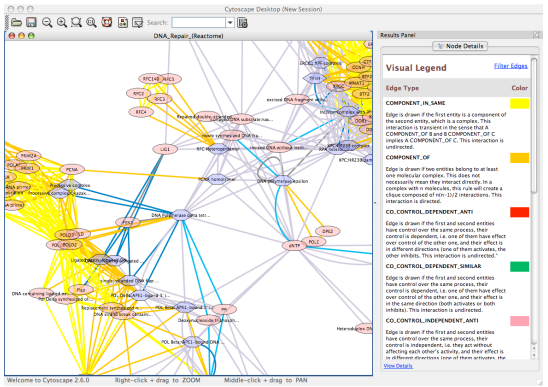


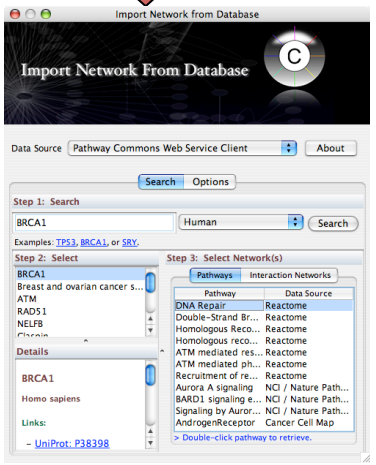
Automatic Layout



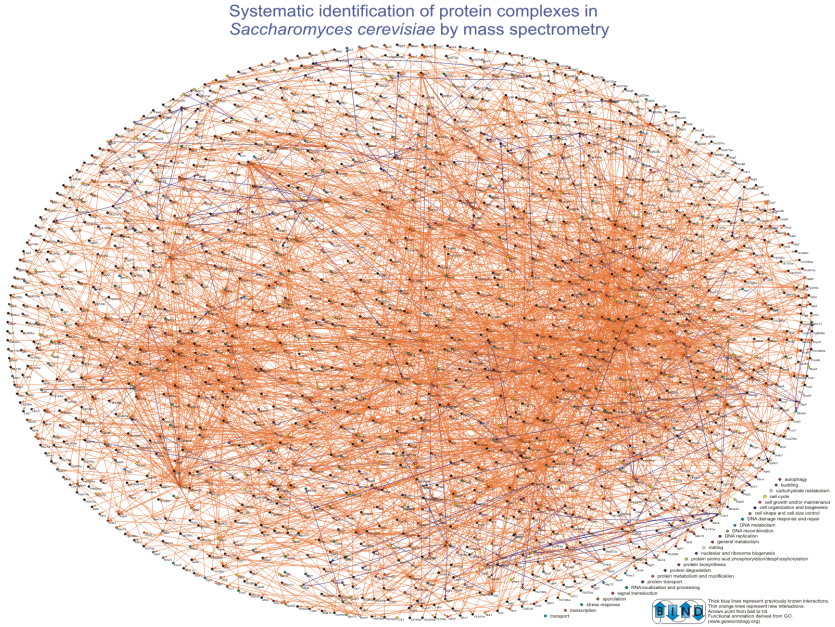
Interaction Database Search







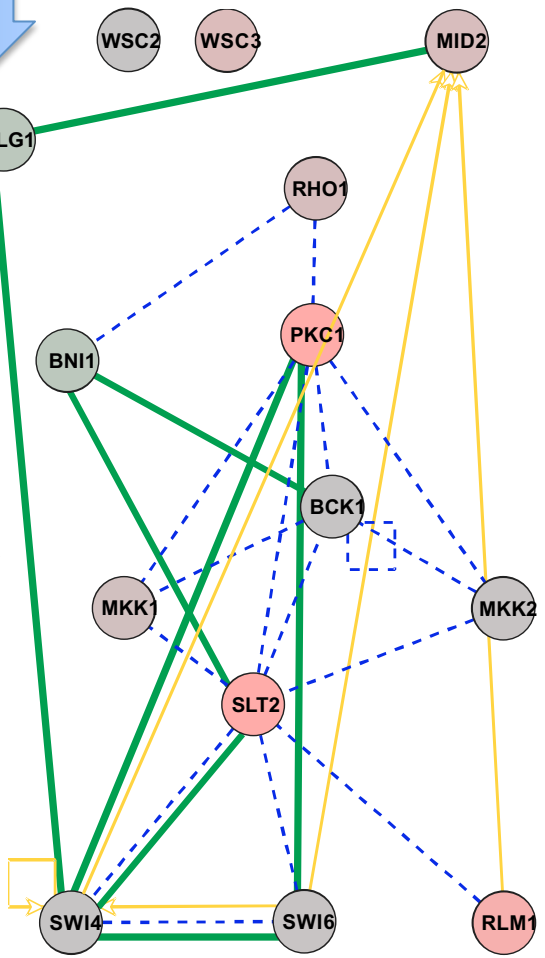
Overview



Zoom

Focus

PKC Cell Wall Integrity



Synthetic Lethal

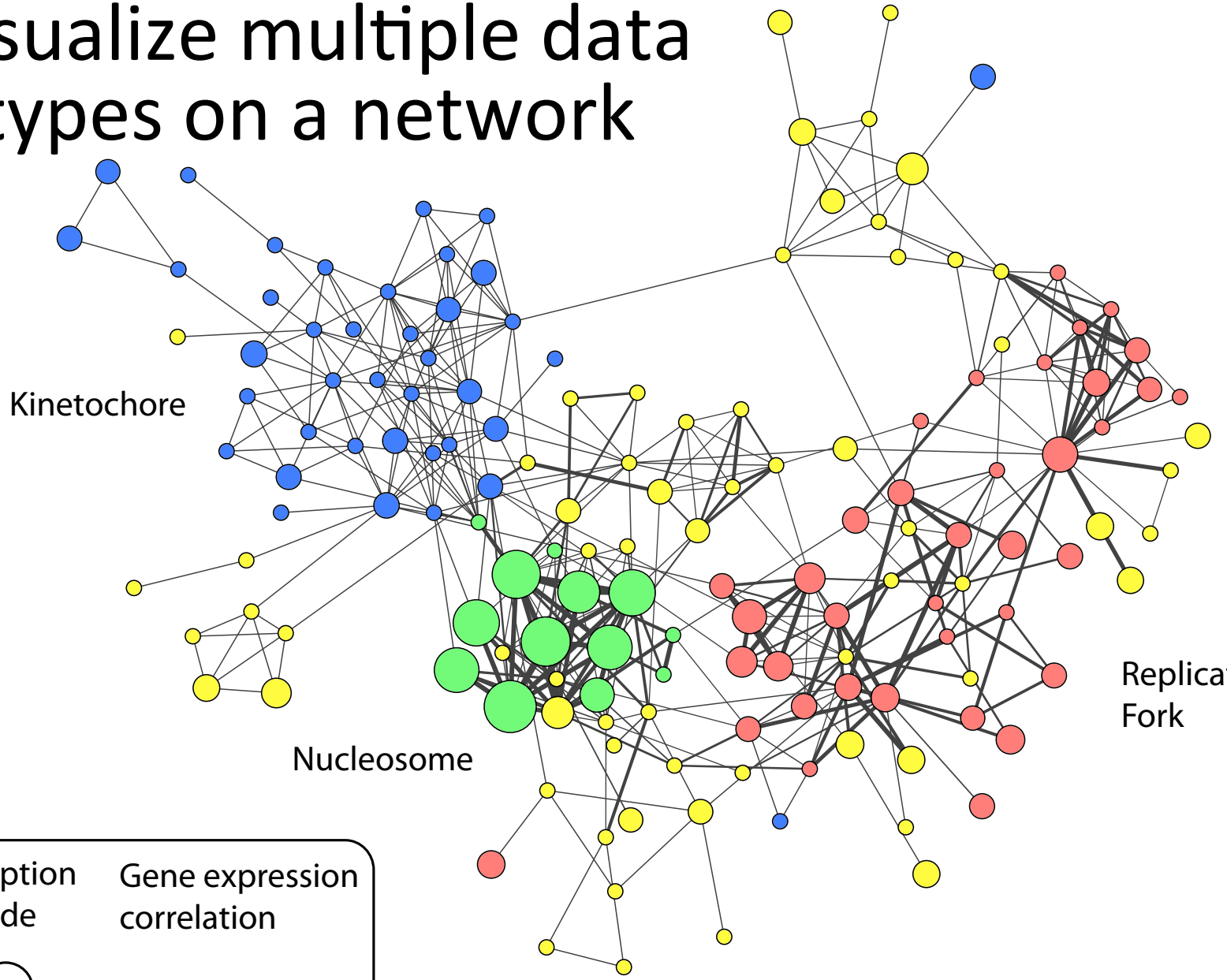
Transcription Factor Regulation

Protein-Protein Interaction

Up Regulated Gene Expression

Down Regulated Gene Expression

Visualize multiple data types on a network



Transcription amplitude

low high

Gene expression correlation

low high

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

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

- 10,000s users, 2500 downloads/month
- >40 Plugins Extend Functionality
  - Build your own, requires programming
    - e.g. Retina Workbench

Prediction of TF Binding Sites

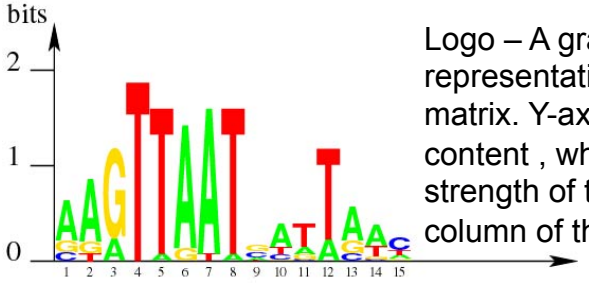
Teaching a computer to find TFBS…

TFBS section from Wyeth Wasserman – full set of slides from:
http://bioinformatics.ca/workshops/2011/course-content

Representing Binding Sites for a TF

- A single site
  - AAGTTAATGA
- A set of sites represented as a consensus
  - VDRTWRWWSHD (IUPAC degenerate DNA)
- A matrix describing a set of sites:

A	14	16	4	0	1	19	20	1	4	13	4	4	13	12	3
C	3	0	0	0	0	0	0	0	7	3	1	0	3	1	12
G	4	3	17	0	0	2	0	0	9	1	3	0	5	2	2
T	0	2	0	21	20	0	1	20	1	4	13	17	0	6	4



Logo – A graphical representation of frequency matrix. Y-axis is information content , which reflects the strength of the pattern in each column of the matrix

Set of binding sites

AAGTTAATGA
CAGTTAATAA
GAGTTAAACA
CAGTTAATTA
GAGTTAATAA
CAGTTATTCA
GAGTTAATAA
CAGTTAATCA
AGATTAAAGA
AAGTTAACGA
AGGTTAACGA
ATGTTGATGA
AAGTTAATGA
AAGTTAACGA
AAATTAATGA
GAGTTAATGA
AAGTTAATCA
AAGTTGATGA
AAATTAATGA
ATGTTAATGA
AAGTAAATGA
AAGTTAATGA
AAGTTAATGA
AAATTAATGA
AAGTTAATGA
AAGTTAATGA
AAGTTAATGA
AAGTTAATGA

Conversion of PFMs to Position Specific Scoring Matrices (PSSM)

Add the following features to the matrix profile:

1. Correct for nucleotide frequencies in genome
2. Weight for the confidence (depth) in the pattern
3. Convert to log-scale probability for easy arithmetic

pfm

A	5	0	1	0	0
C	0	2	2	4	0
G	0	3	1	0	4
T	0	0	1	1	1

$$\text{Log}\left(\frac{f(\text{b,i}) + s(\text{n})}{p(\text{b})}\right)$$

pssm

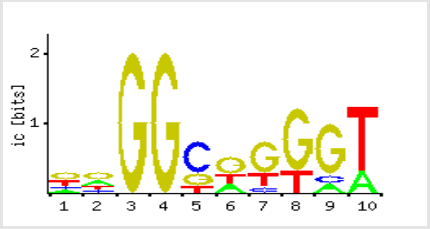
A	1.6	-1.7	-0.2	-1.7	-1.7
C	-1.7	0.5	0.5	1.3	-1.7
G	-1.7	1.0	-0.2	-1.7	1.3
T	-1.7	-1.7	-0.2	-0.2	-0.2

TGCTG = 0.9

Detecting binding sites in a single sequence

Raw Scores

Sp1



ACCCTCCCCAGGGGCGGGGGGCGGTGGCCAGGACGGTAGCTCC

```
A  [-0.2284  0.4368   -1.5    -1.5    -1.5  0.4368   -1.5   -1.5 -0.2284  0.4368 ]
C  [-0.2284 -0.2284   -1.5    -1.5  1.5128   -1.5 -0.2284   -1.5 -0.2284   -1.5 ]
G  [ 1.2348  1.2348  2.1222  2.1222  0.4368  1.2348  1.5128  1.7457  1.7457   -1.5 ]
T  [ 0.4368 -0.2284   -1.5    -1.5 -0.2284  0.4368  0.4368  0.4368   -1.5  1.7457 ]
```

Abs_score = 13.4 (sum of column scores)

Relative Scores

```
A  [-0.2284  0.4368   -1.5    -1.5    -1.5  0.4368   -1.5   -1.5 -0.2284  0.4368 ]
C  [-0.2284 -0.2284   -1.5    -1.5  1.5128   -1.5 -0.2284   -1.5 -0.2284   -1.5 ]
G  [ 1.2348  1.2348  2.1222  2.1222  0.4368  1.2348  1.5128  1.7457  1.7457   -1.5 ]
T  [ 0.4368 -0.2284   -1.5    -1.5 -0.2284  0.4368  0.4368  0.4368   -1.5  1.7457 ]
```

Max_score = 15.2 (sum of highest column scores)

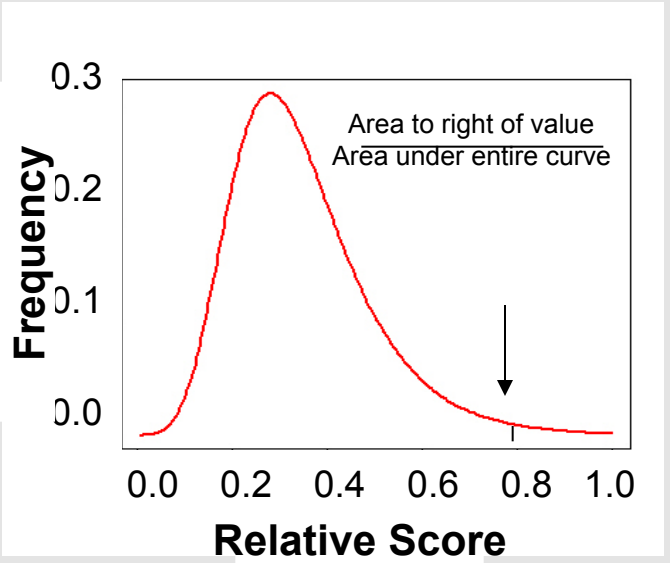
```
A  [-0.2284  0.4368   -1.5    -1.5    -1.5  0.4368   -1.5   -1.5 -0.2284  0.4368 ]
C  [-0.2284 -0.2284   -1.5    -1.5  1.5128   -1.5 -0.2284   -1.5 -0.2284   -1.5 ]
G  [ 1.2348  1.2348  2.1222  2.1222  0.4368  1.2348  1.5128  1.7457  1.7457   -1.5 ]
T  [ 0.4368 -0.2284   -1.5    -1.5 -0.2284  0.4368  0.4368  0.4368   -1.5  1.7457 ]
```

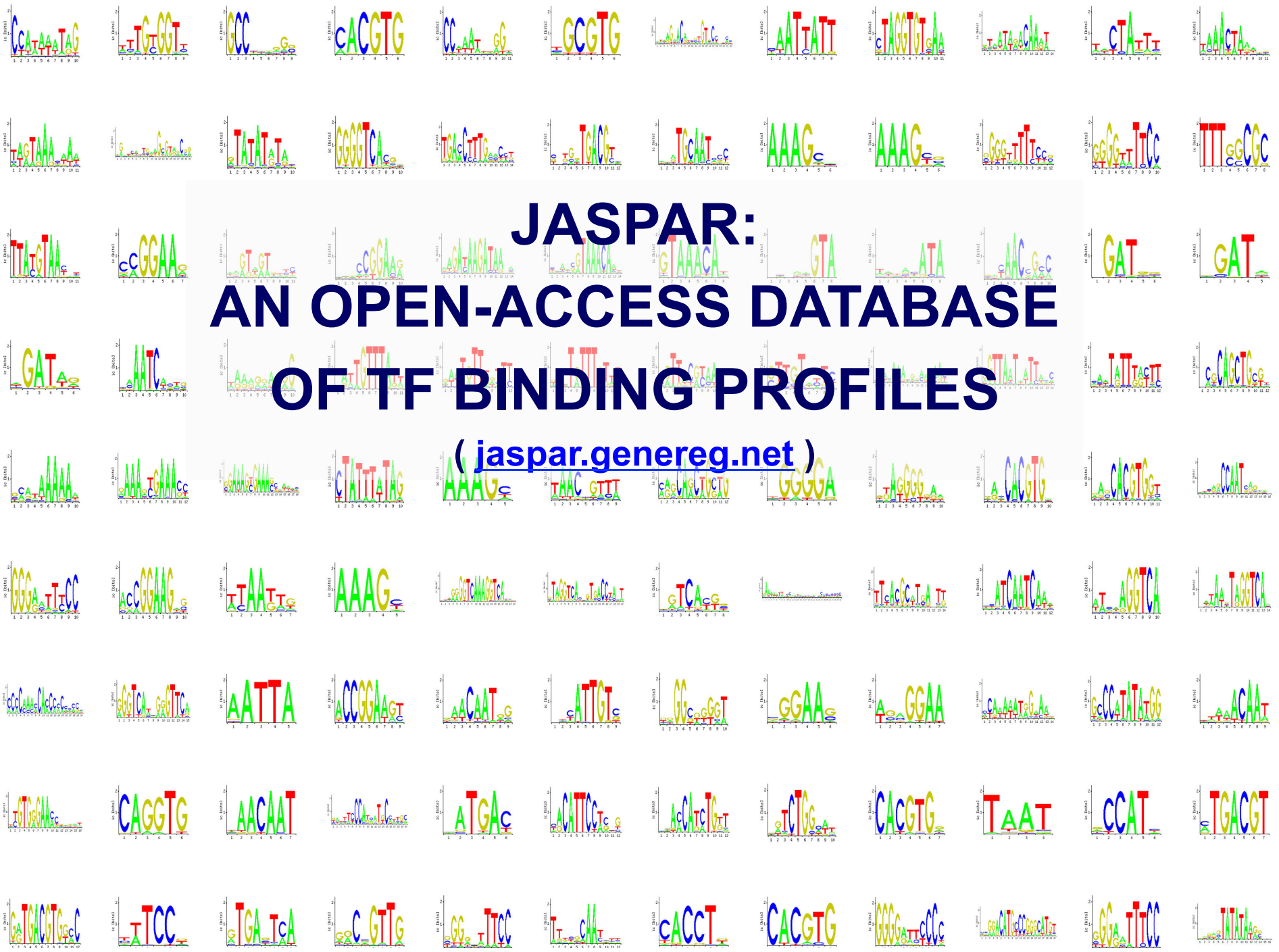
Min_score = -10.3 (sum of lowest column scores)

$$\text{Rel\_score} = \frac{\text{Abs\_score} - \text{Min\_score}}{\text{Max\_score} - \text{Min\_score}} \cdot 100\ \%$$

$$= \frac{13.4 - (-10.3)}{15.2 - (-10.3)} \cdot 100\% = \mathbf{93\%}$$

Empirical p-value Scores





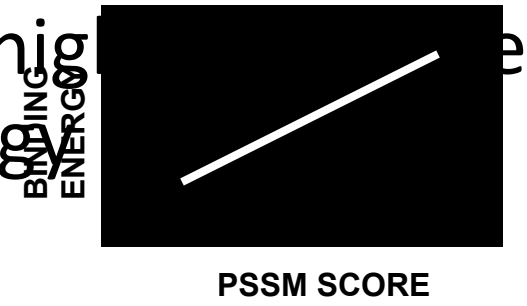
JASPAR:
AN OPEN-ACCESS DATABASE OF TF BINDING PROFILES

(jaspar.genereg.net)

The Good…

- Tronche (1997) tested 50 predicted HNF1 TFBS using an in vitro binding test and found that 96% of the predicted sites were bound!

- Stormo and Fields (1998) found in detailed biochemical studies that the best weight matrices produce scores hig ed with in vitro binding energy



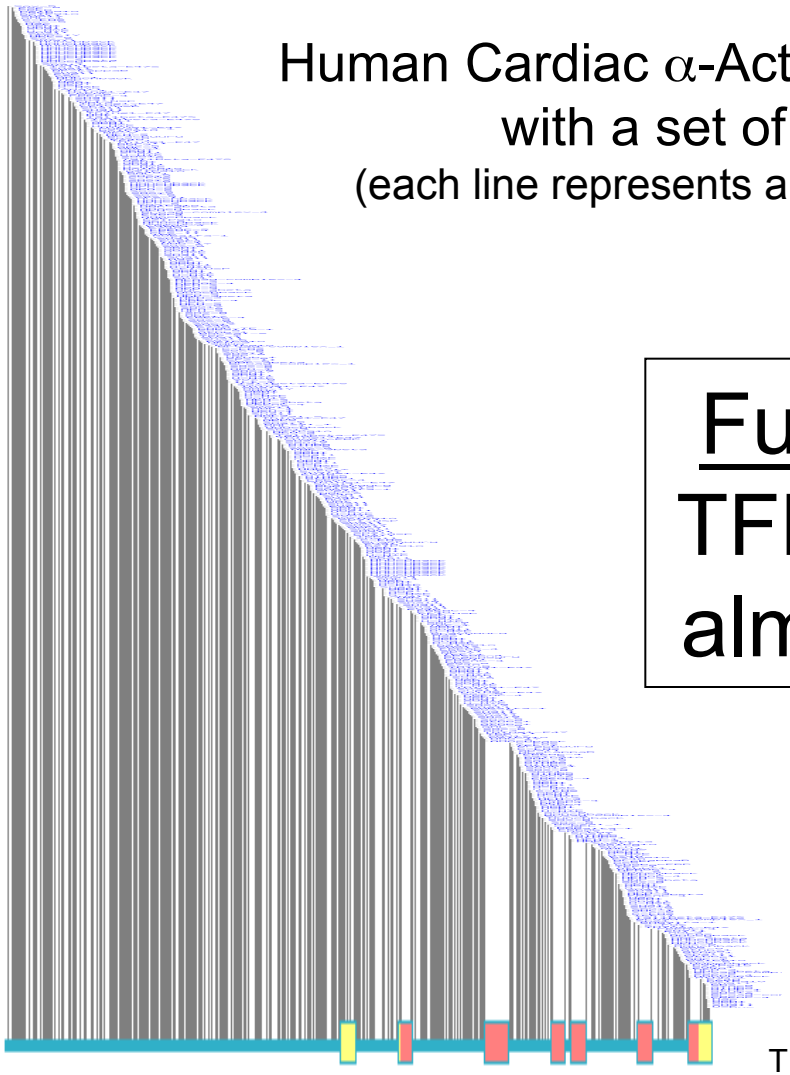
…the Bad…

- Fickett (1995) found that a profile for the myoD TF made predictions at a rate of 1 per ~500bp of human DNA sequence

  – This corresponds to an average of 20 sites / gene (assuming 10,000 bp as average gene size)

…and the Ugly!

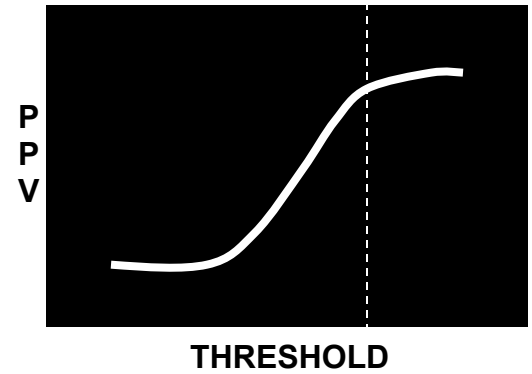
Human Cardiac α-Actin gene analyzed
with a set of profiles
(each line represents a TFBS prediction)



Futility Conjuncture:
TFBS predictions are almost always wrong

Red boxes are protein coding exons -
TFBS predictions excluded in this analysis

A Conundrum…

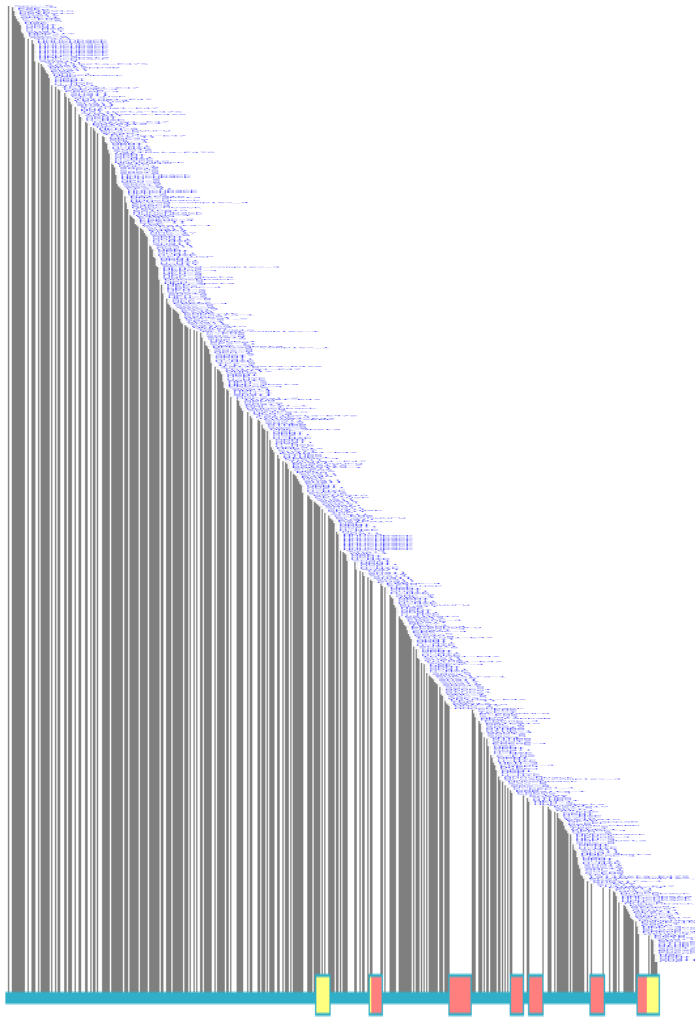


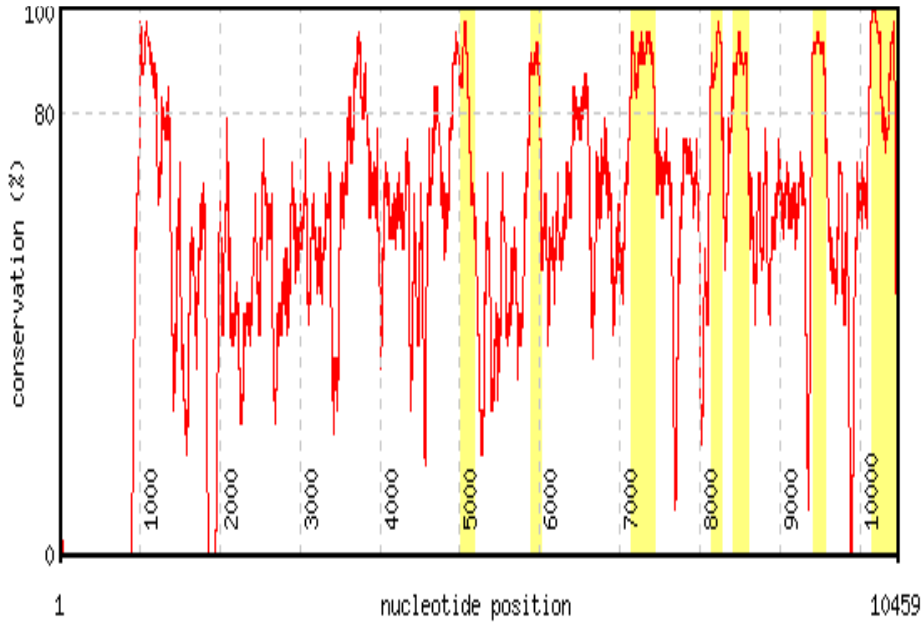
- Counter to intuition, the ratio of true positives to predictions fails to improve for "stringent" thresholds
    - For most predictive models this ratio would increase
- Why?
    - True binding sites are defined by properties not incorporated into the profile scores - above some threshold all sites *could* be bound if present in the

Using Phylogenetic Footprinting to Improve TFBS Discrimination

70,000,000 years of evolution can reveal regulatory regions

Phylogenetic Footprinting Dramatically Reduces Spurious Hits





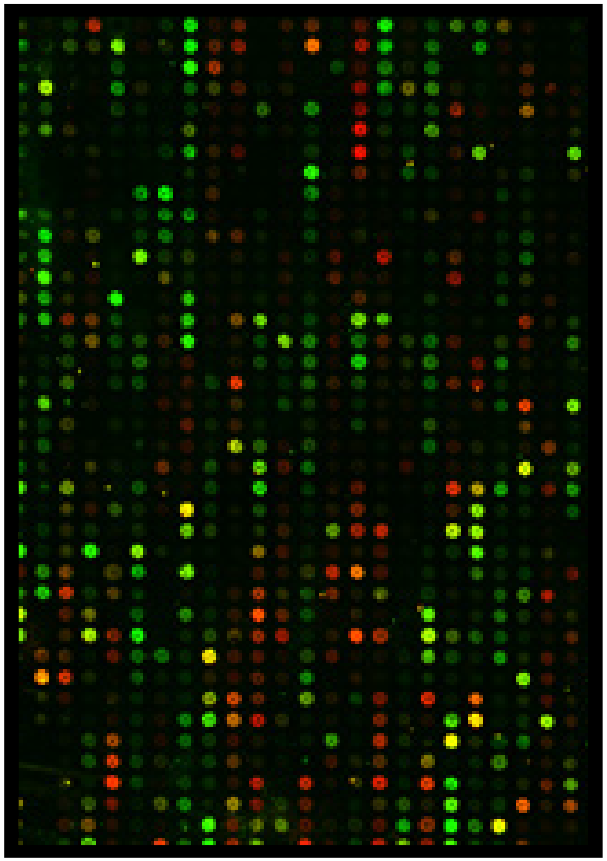
Actin, alpha cardiac

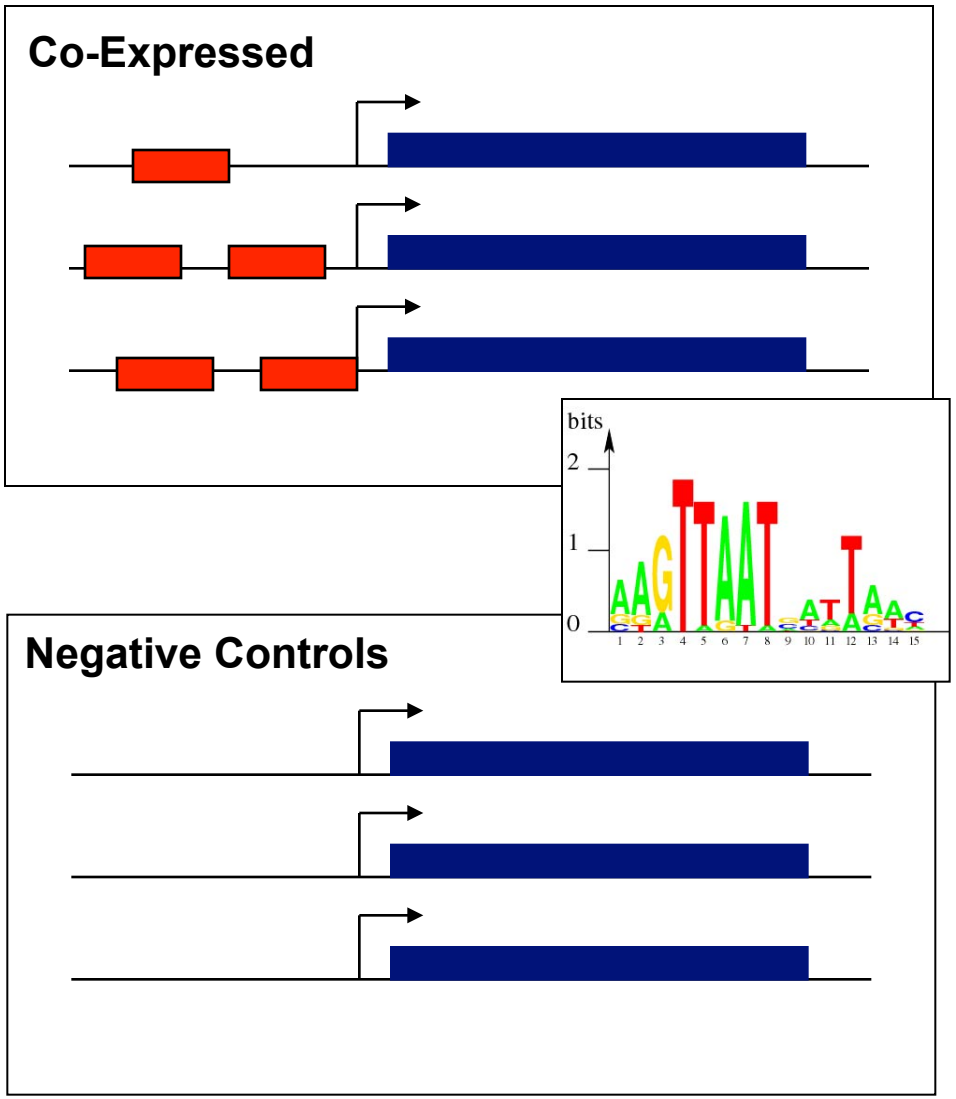
TFBS Discrimination Tools

- Phylogenetic Footprinting Servers
  - FOOTER http://biodev.hgen.pitt.edu/footer_php/Footerv2_0.php
  - CONSITE http://asp.ii.uib.no:8090/cgi-bin/CONSITE/consite/
  - rVISTA http://rvista.dcode.org/
- SNPs in TFBS Analysis
  - RAVEN http://burgundy.cmmt.ubc.ca/cgi-bin/RAVEN/a?rm=home
  - is-rSNP http://www.genomics.csse.unimelb.edu.au/is-rSNP/
- Prokaryotes
  - PRODORIC http://prodoric.tu-bs.de/
- Software Packages
  - TOUCAN http://homes.esat.kuleuven.be/~saerts/software/toucan.php
  - RSA Tools http://rsat.ulb.ac.be/
- Programming Tools
  - TFBS http://tfbs.genereg.net/
  - ORCAtk http://burgundy.cmmt.ubc.ca/cgi-bin/OrcaTK/orcatk

Inferring Regulating TFs for Sets of Co-Expressed Genes

Deciphering Regulation of Co-Expressed Genes

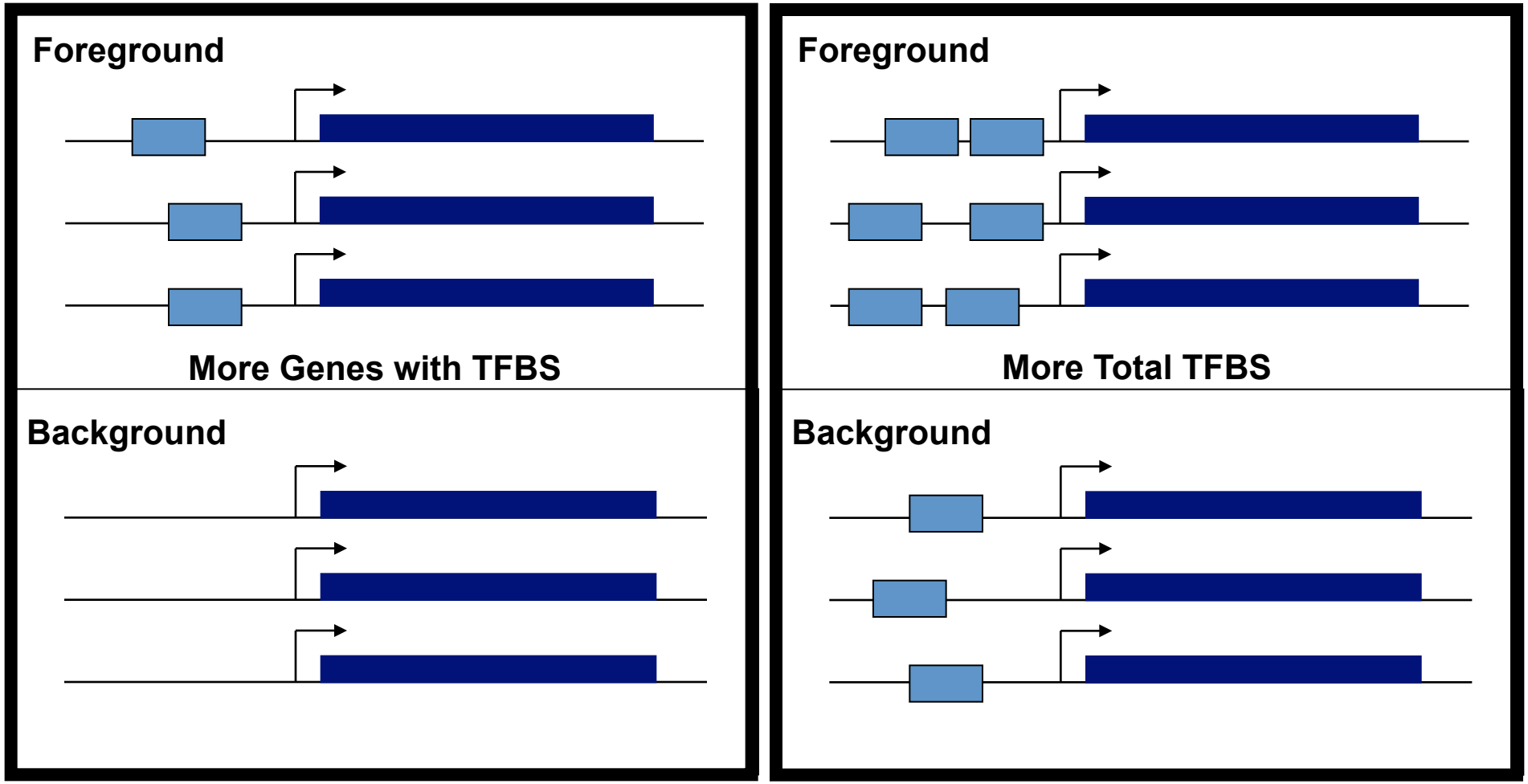




TFBS Over-representation

- Akin to methods for GO term over-representation analysis, we seek to determine if a set of co-expressed genes contains an over-abundance of predicted binding sites for a known TF
    - Phylogenetic footprinting to reduce false prediction rate for metazoan genomes

Two Examples of TFBS Over-Representation



Statistical Methods for Identifying Over-represented TFBS

- Binomial test (Z scores)
  - Based on the ***number of occurrences*** of the TFBS relative to background
  - Normalized for sequence length
  - Simple binomial distribution model

- Fisher exact probability scores
  - Based on the ***number of genes*** containing the TFBS relative to background
  - Hypergeometric probability distribution

oPOSSUM Version 3.0

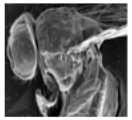
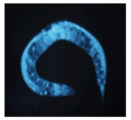
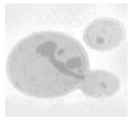
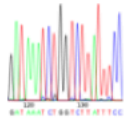
About | Help | FAQ | Download | Contact

Welcome to oPOSSUM

oPOSSUM is a web-based system for the detection of over-represented conserved transcription factor binding sites and binding site combinations in sets of genes or sequences.

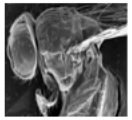
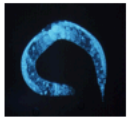
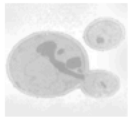
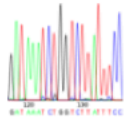
Single Site Analysis (SSA)

Detect over-represented conserved transcription factor binding sites in a set of genes or sequences.

 Human
 Mouse
 Fly
 Worm
 Yeast
 Sequence-based

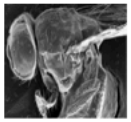
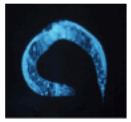
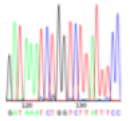
Anchored Combination Site Analysis (aCSA)

Detect over-represented **combinations** of conserved transcription factor binding sites in a set of genes or sequences.

 Human
 Mouse
 Fly
 Worm
 Yeast
 Sequence-based

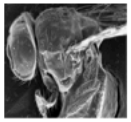
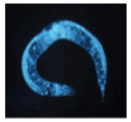
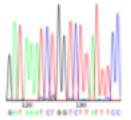
TFBS Cluster Analysis (TCA)

Detect over-represented conserved transcription factor binding site clusters in a set of sequences.

 Human
 Mouse
 Fly
 Worm
 Sequence-based

Anchored Combination TFBS Cluster Analysis (aCTCA)

Detect over-represented combinatons of conserved transcription factor binding site clusters in a set of sequences.

 Human
 Mouse
 Fly
 Worm
 Sequence-based

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

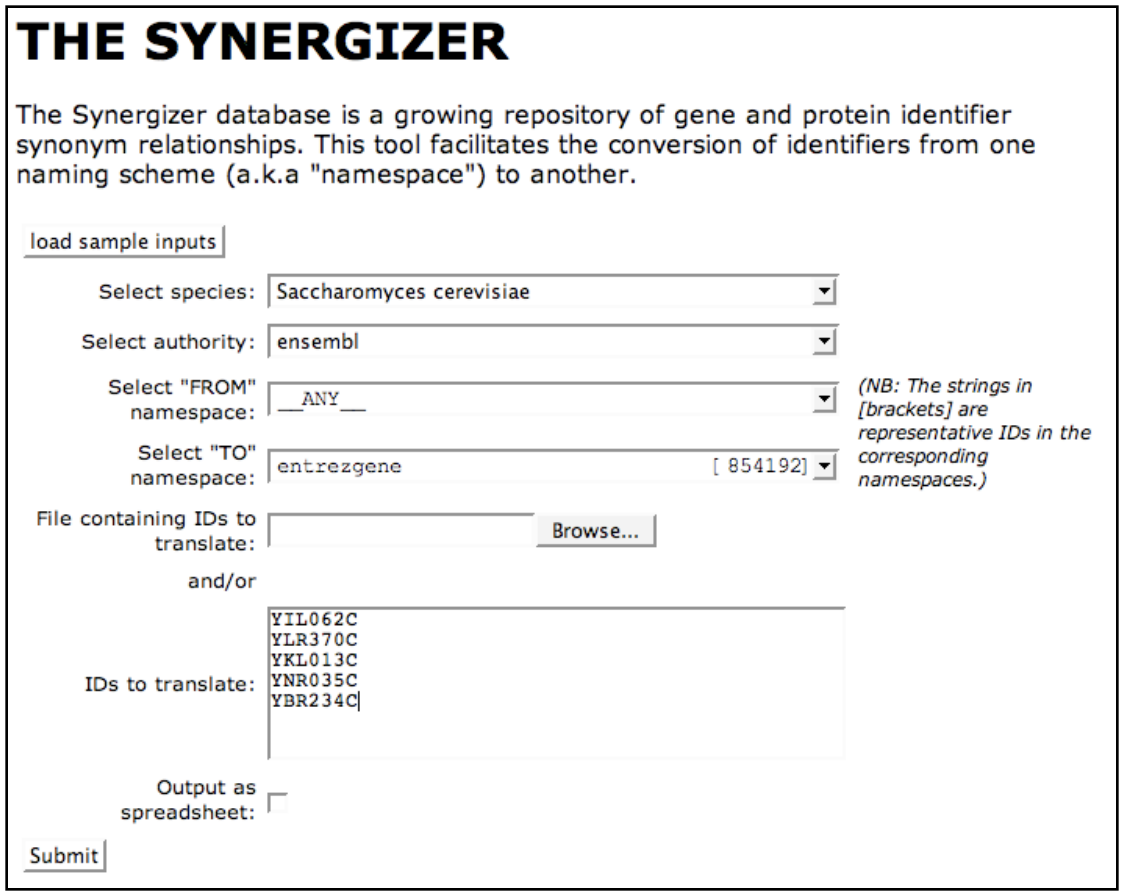
Species-specific
HUGO HGNC BRCA2
MGI MGI:109337
RGD 2219
ZFIN ZDB-GENE-060510-3
FlyBase CG9097
WormBase WBGene00002299 or ZK1067.1
SGD S000002187 or YDL029W

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



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

- 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

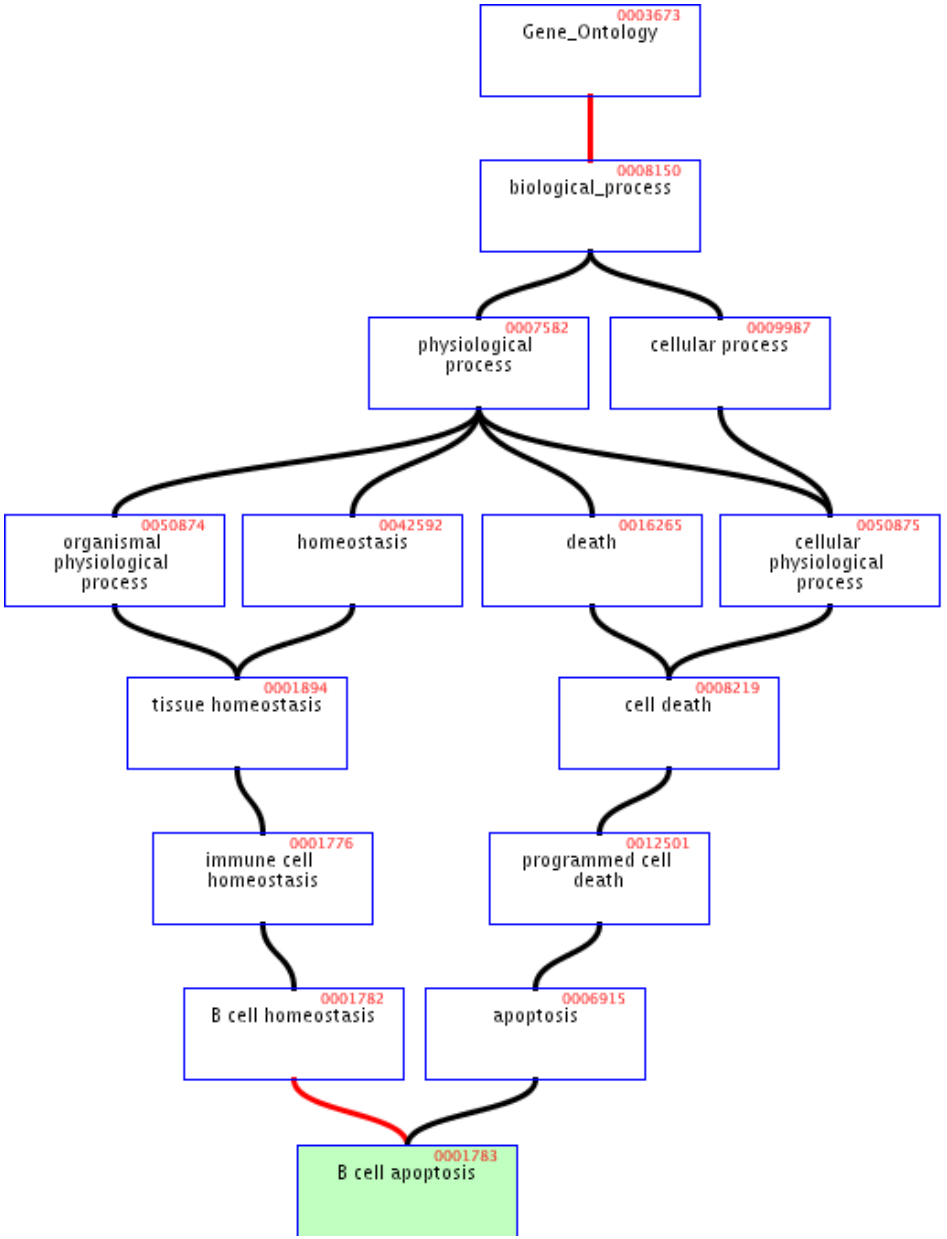
Additional Plugins

- 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

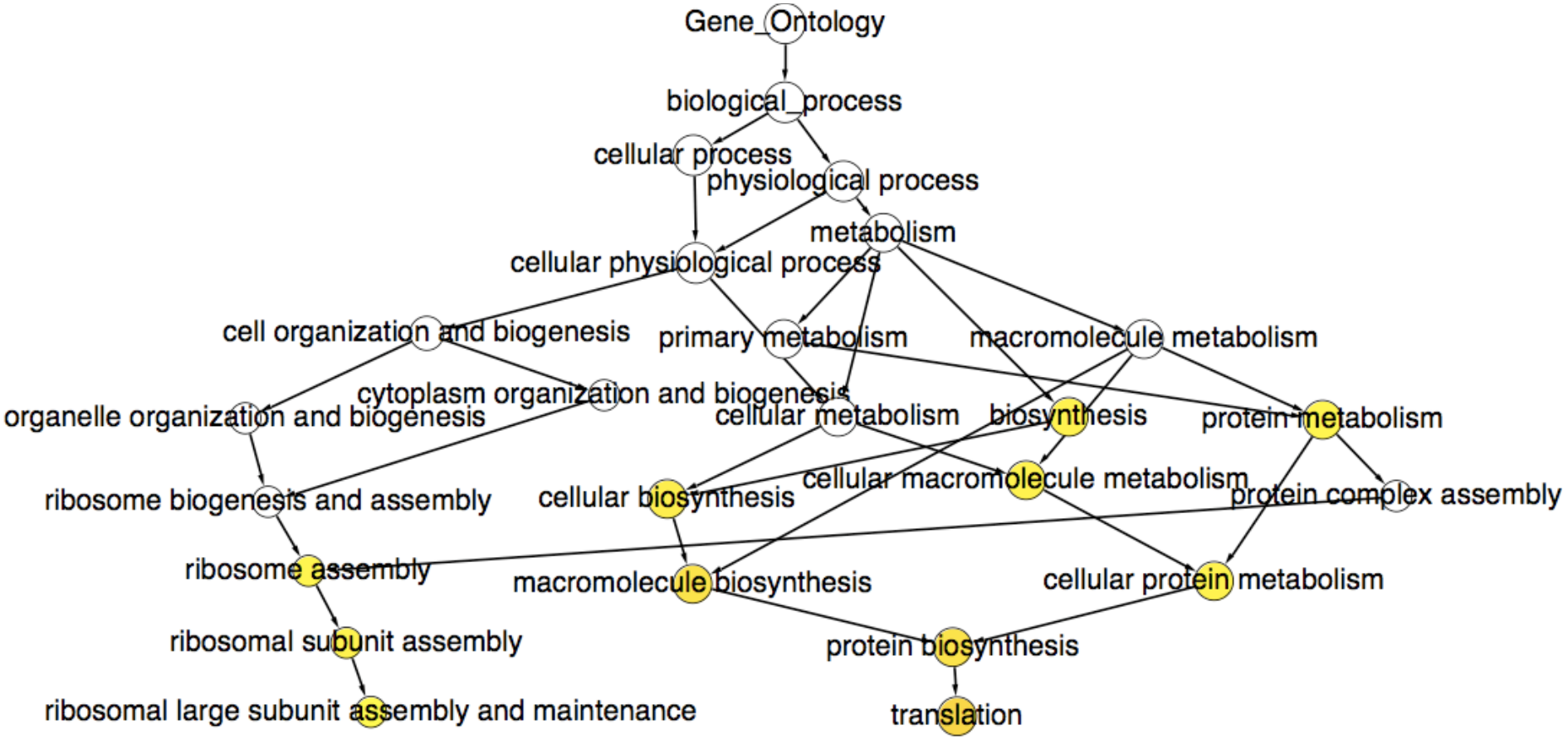
1. Agreed upon terms (controlled vocabulary)
   - Biological process
   - Cellular component
   - Molecular function
2. Genome annotation



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

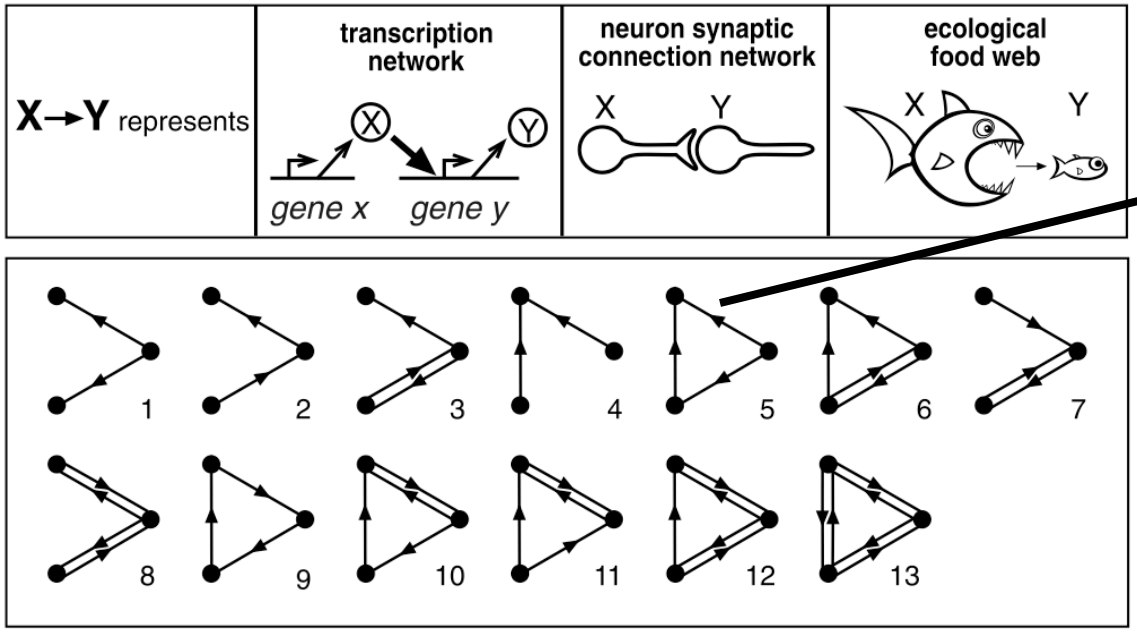
- Query a network for topological matches
- Input: query and target networks, optional node/edge labels
- Output: Topological query matches as subgraphs of target network
- Supports: subgraph matching, node/edge labels, label wildcards, approximate paths
- http://alpha.dmi.unict.it/~ctnyu/netmatch.html

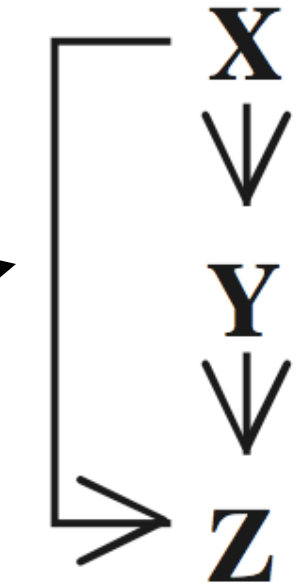
Ferro A, Giugno R, Pigola G, Pulvirenti A, Skripin D, Bader GD, Shasha D Bioinformatics 2007 Feb 3

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



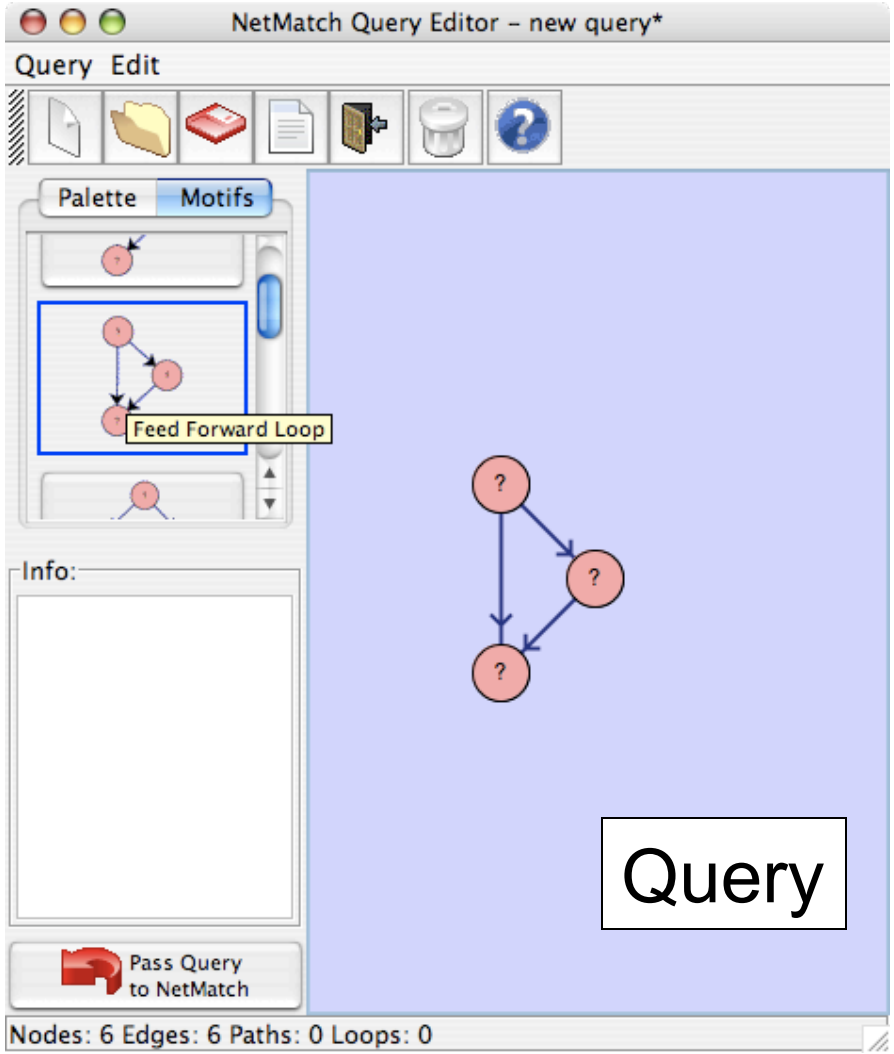


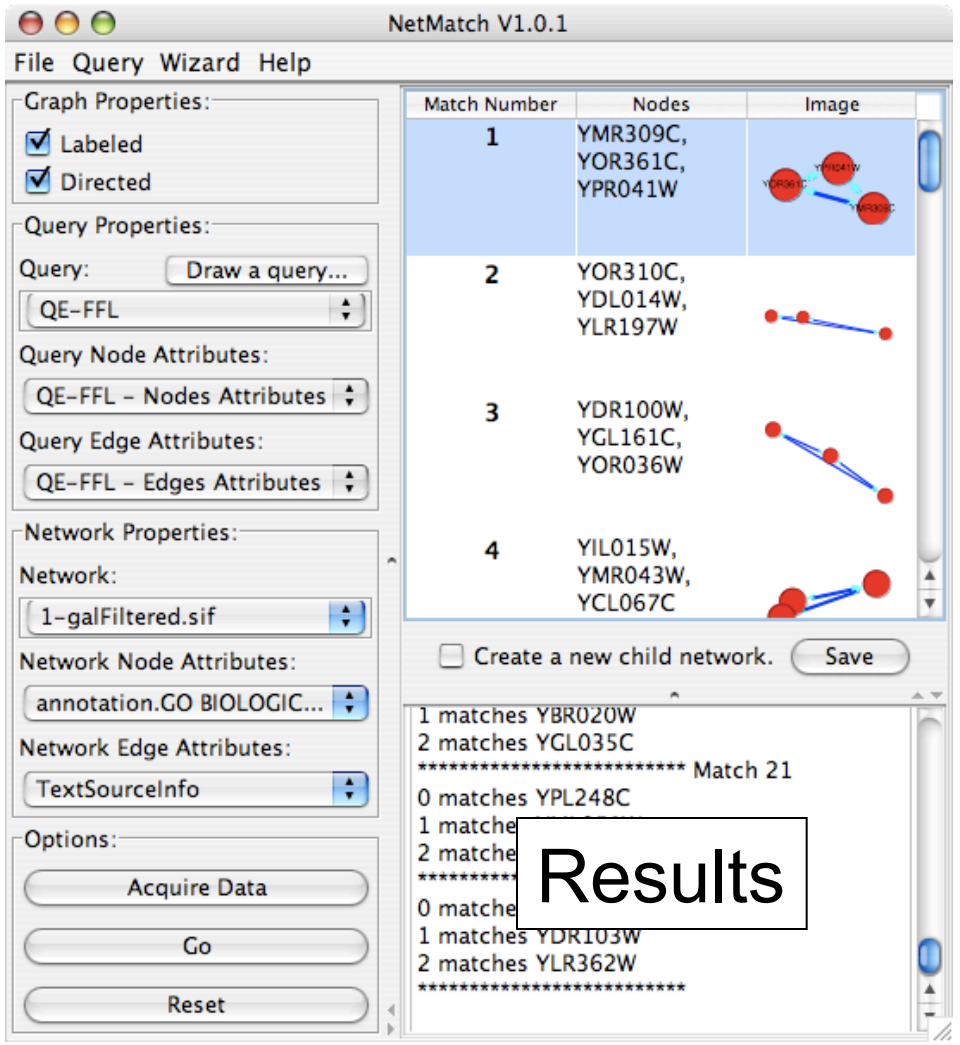
Feed-forward loop

Gene regulation
Neurons
Electronic circuits

Milo et al. Science 298, 2002

Find Feed-Forward Motifs

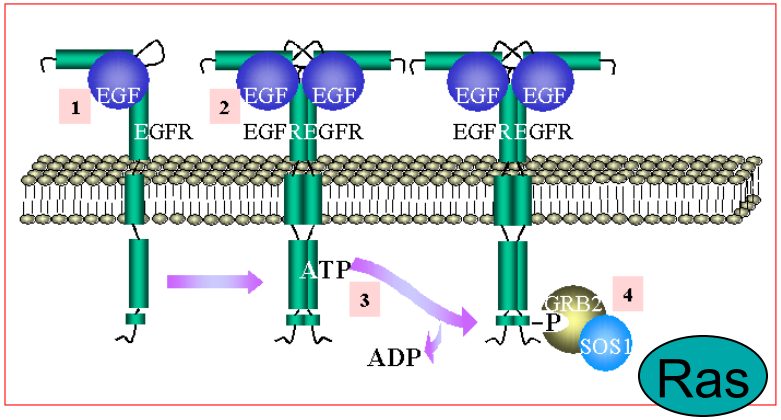




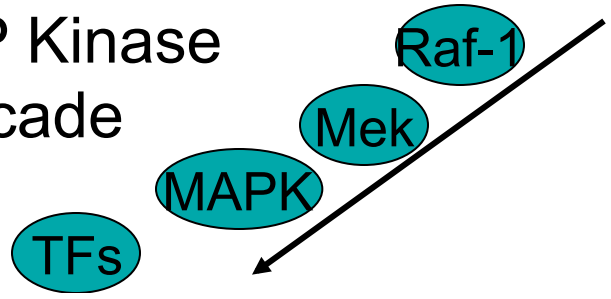
Find Signaling Pathways

- Potential signaling pathways from plasma membrane to nucleus via cytoplasm

Signaling pathway example

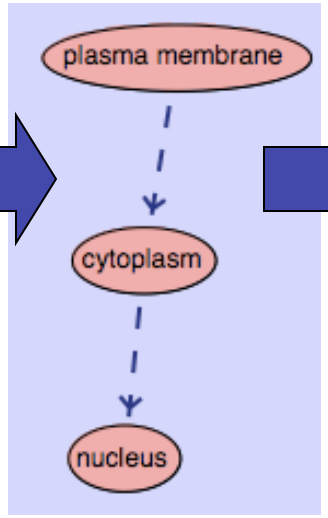


MAP Kinase Cascade



Nucleus - Growth Control Mitogenesis

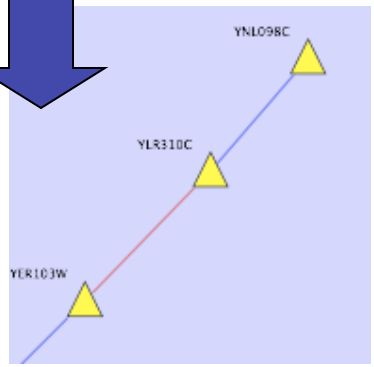
NetMatch query



NetMatch Results

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

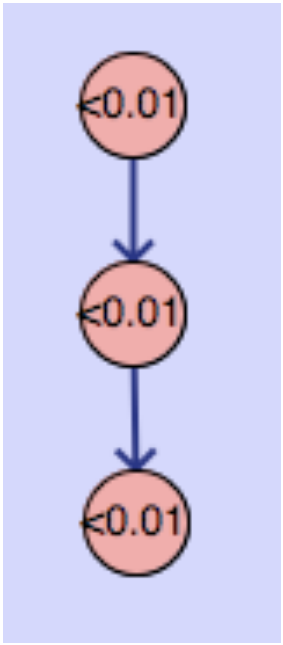
Shortest path between subgraph matches



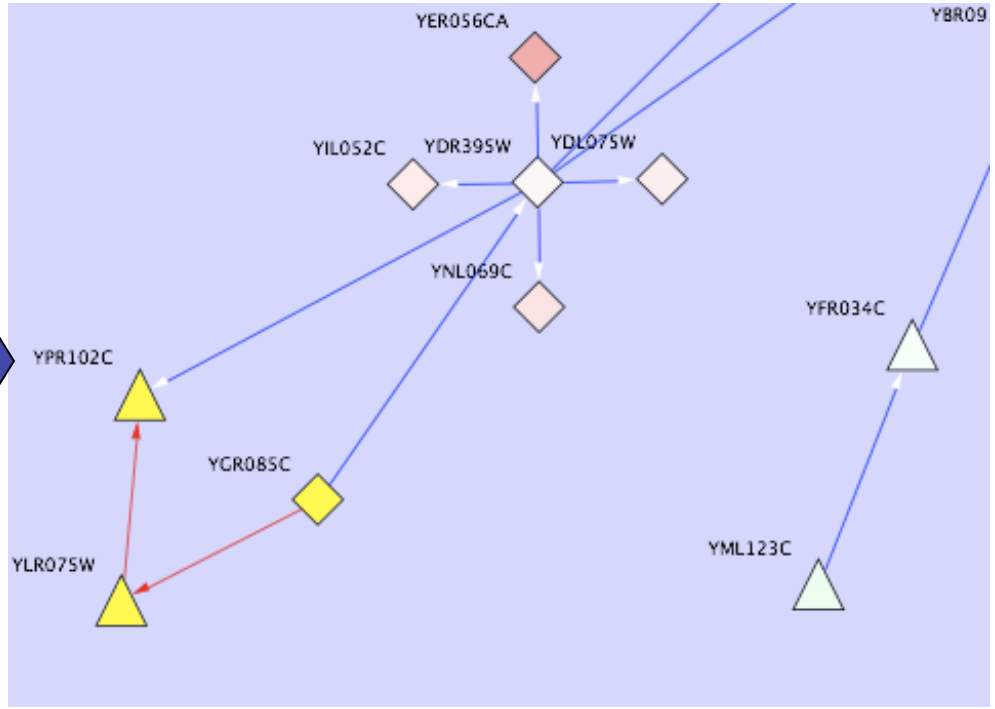
Find Expressed Motifs

Find specific subgraphs where certain nodes are significantly differentially expressed

NetMatch query



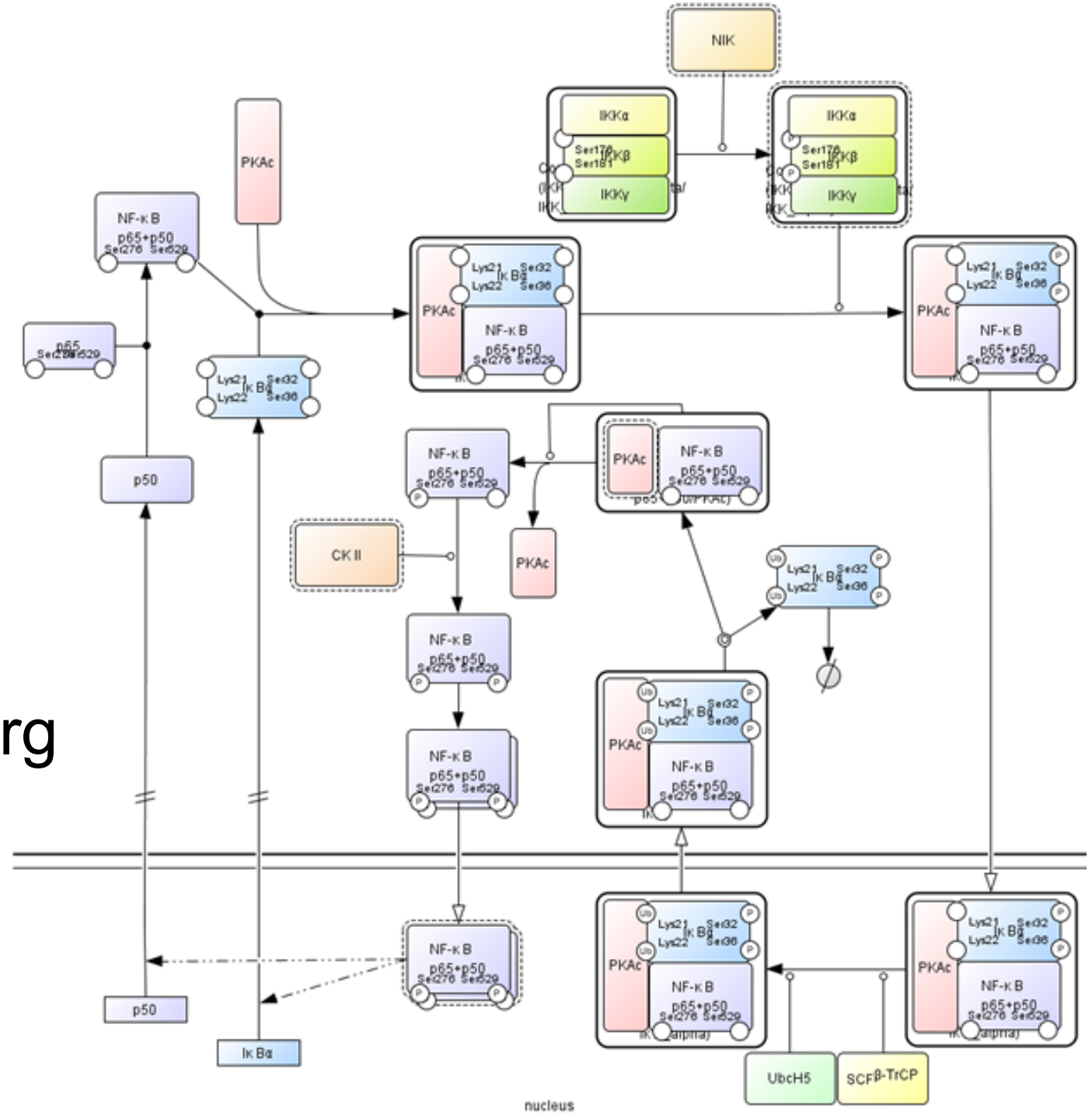
NetMatch Results



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

Systems Biology Graphical Notation

http://sbgn.org

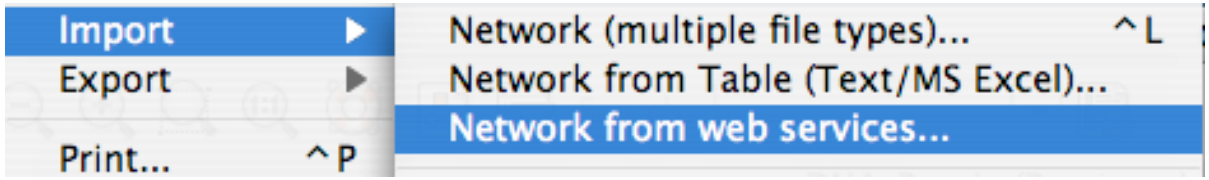


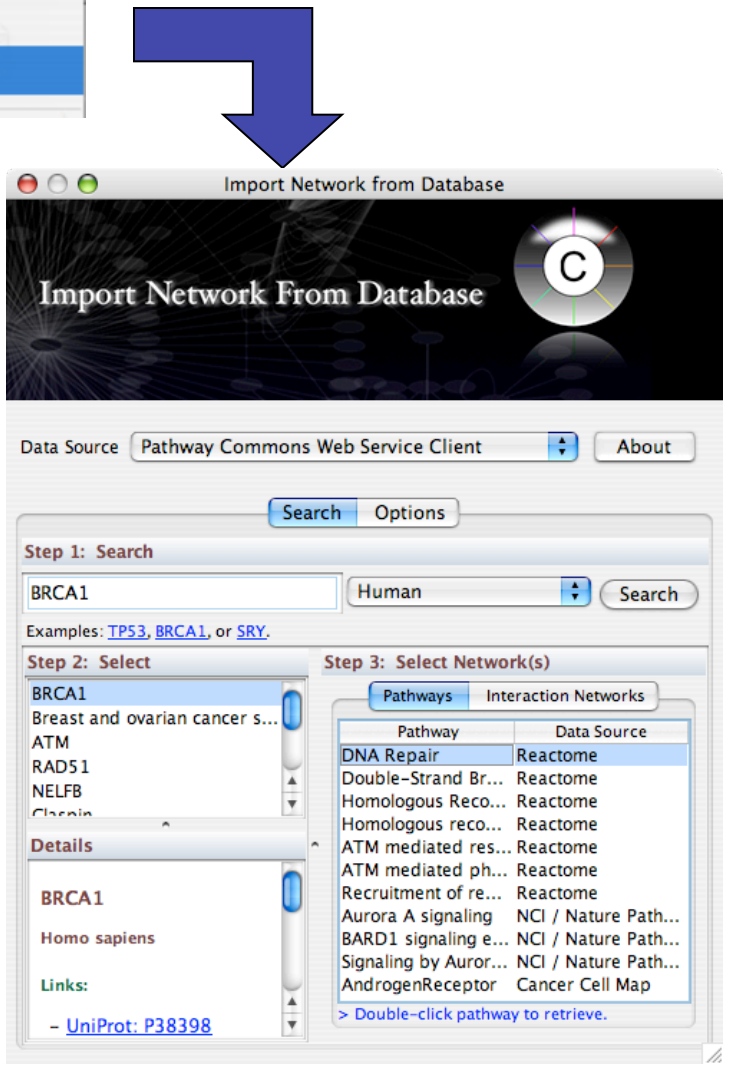
Analyzing Molecular Profiles

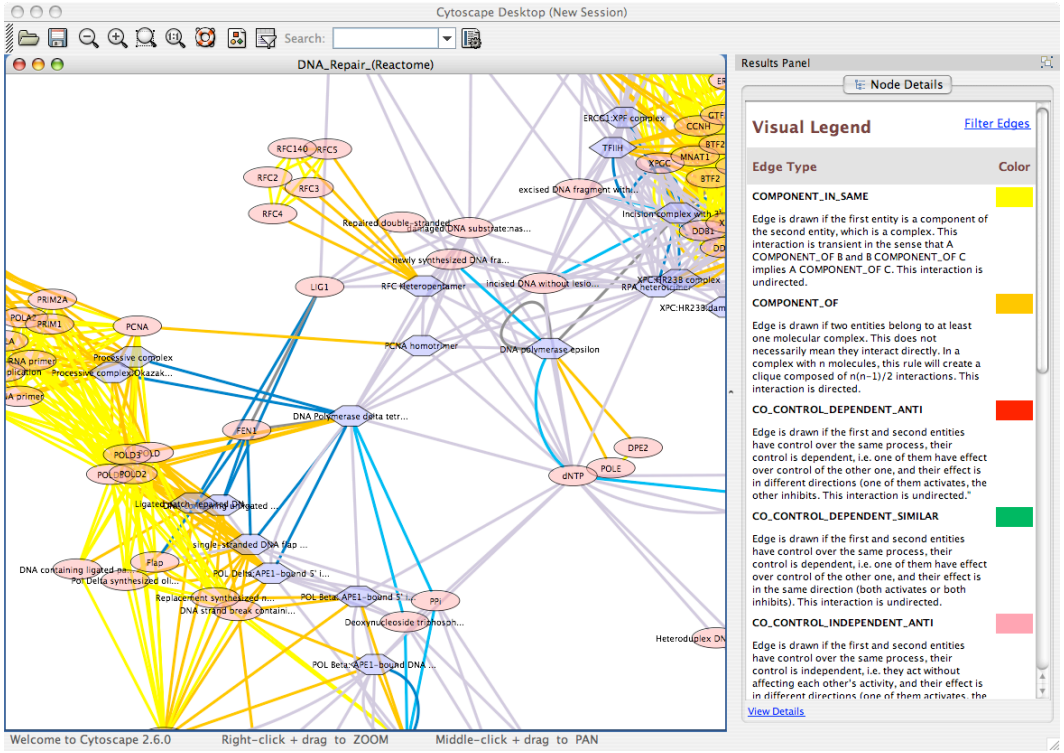
Analyzing gene expression data in a network context

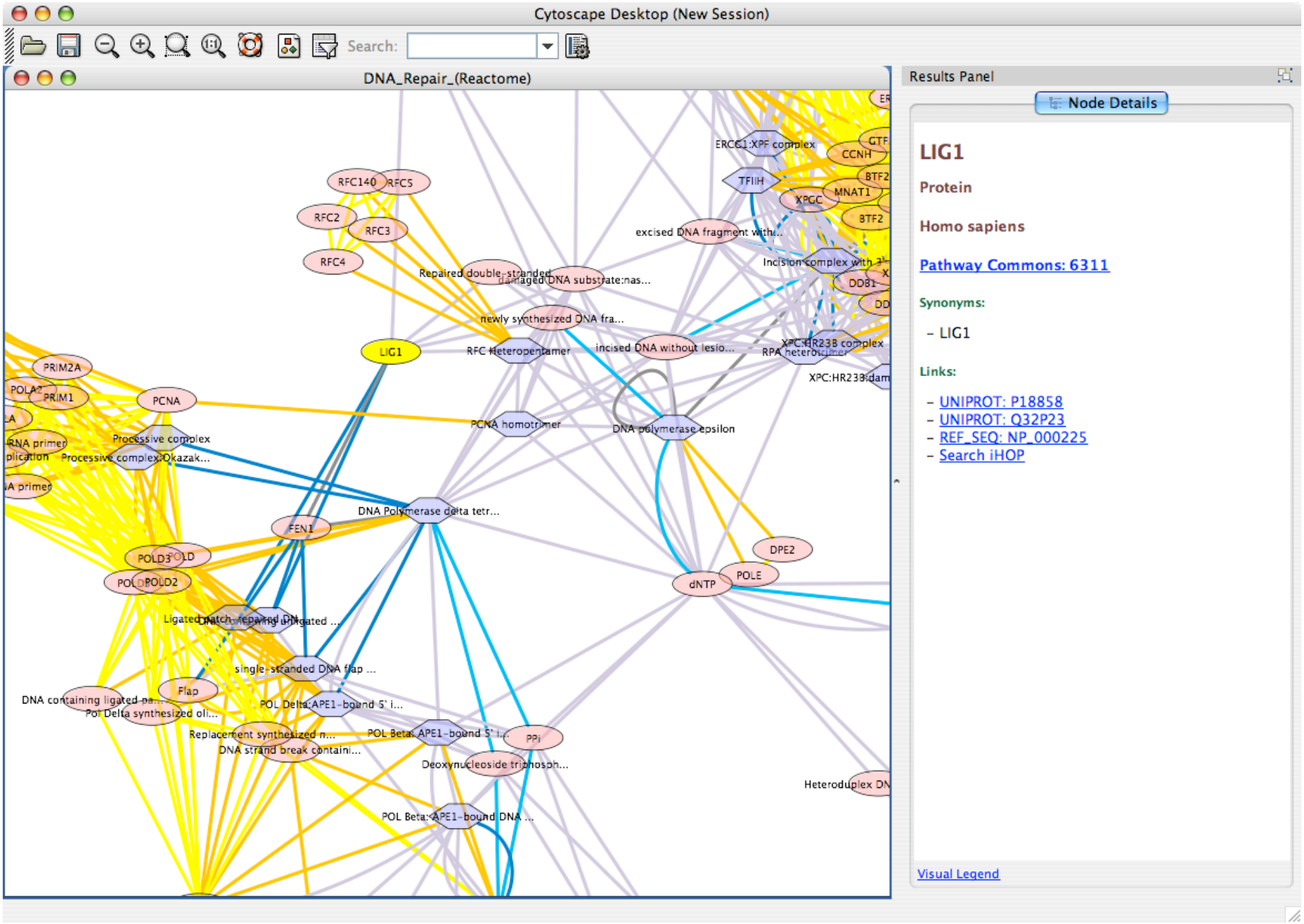
- Input
  - Gene expression data
  - Network data
- Output
  - Visual diagram of expression data on network
  - Active network regions
- Outline
  - Where to find network data?
    - Interaction database (cPath)
    - Literature associations via text mining
  - Load expression data
  - Identify active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

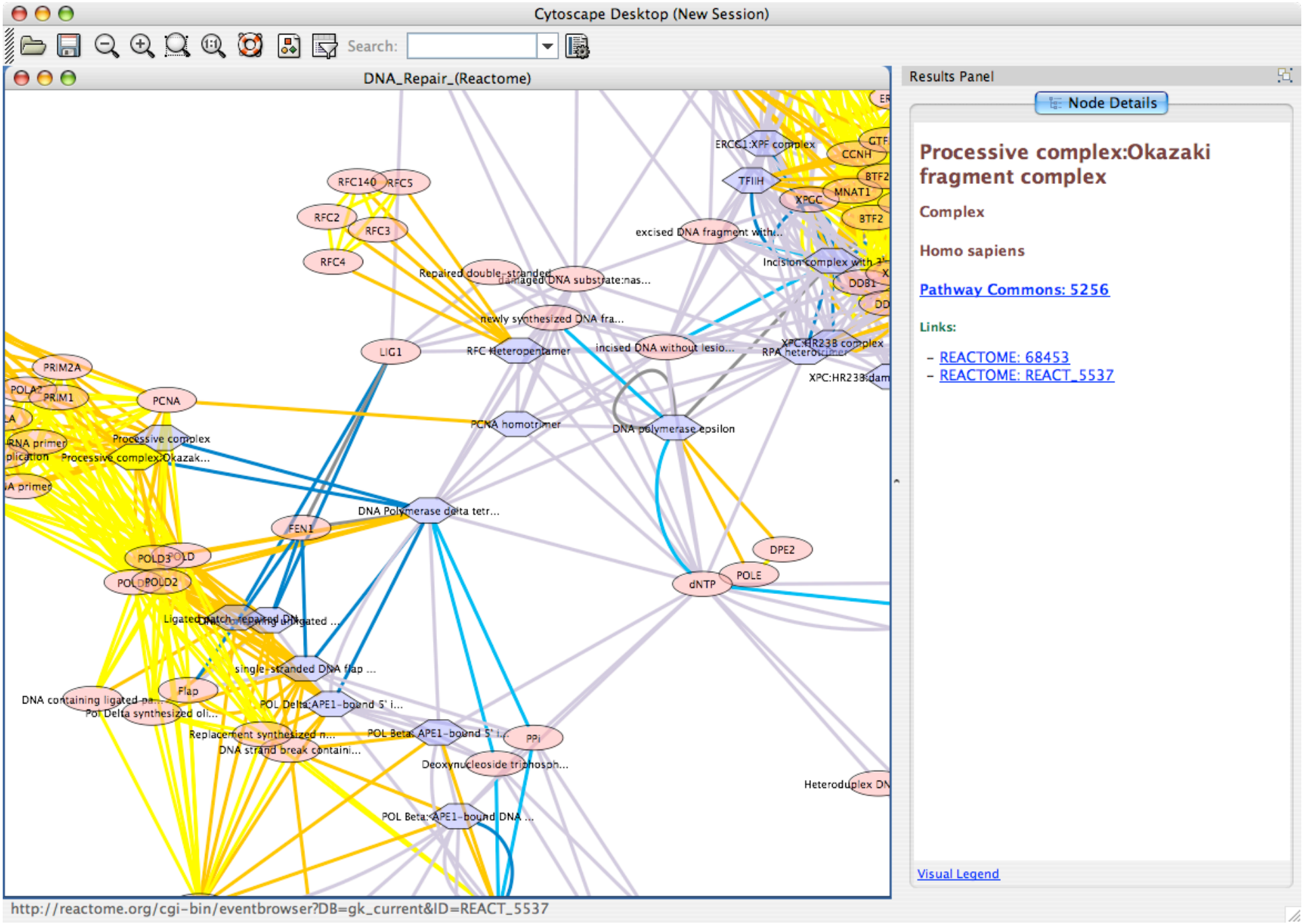
Interaction Database Search





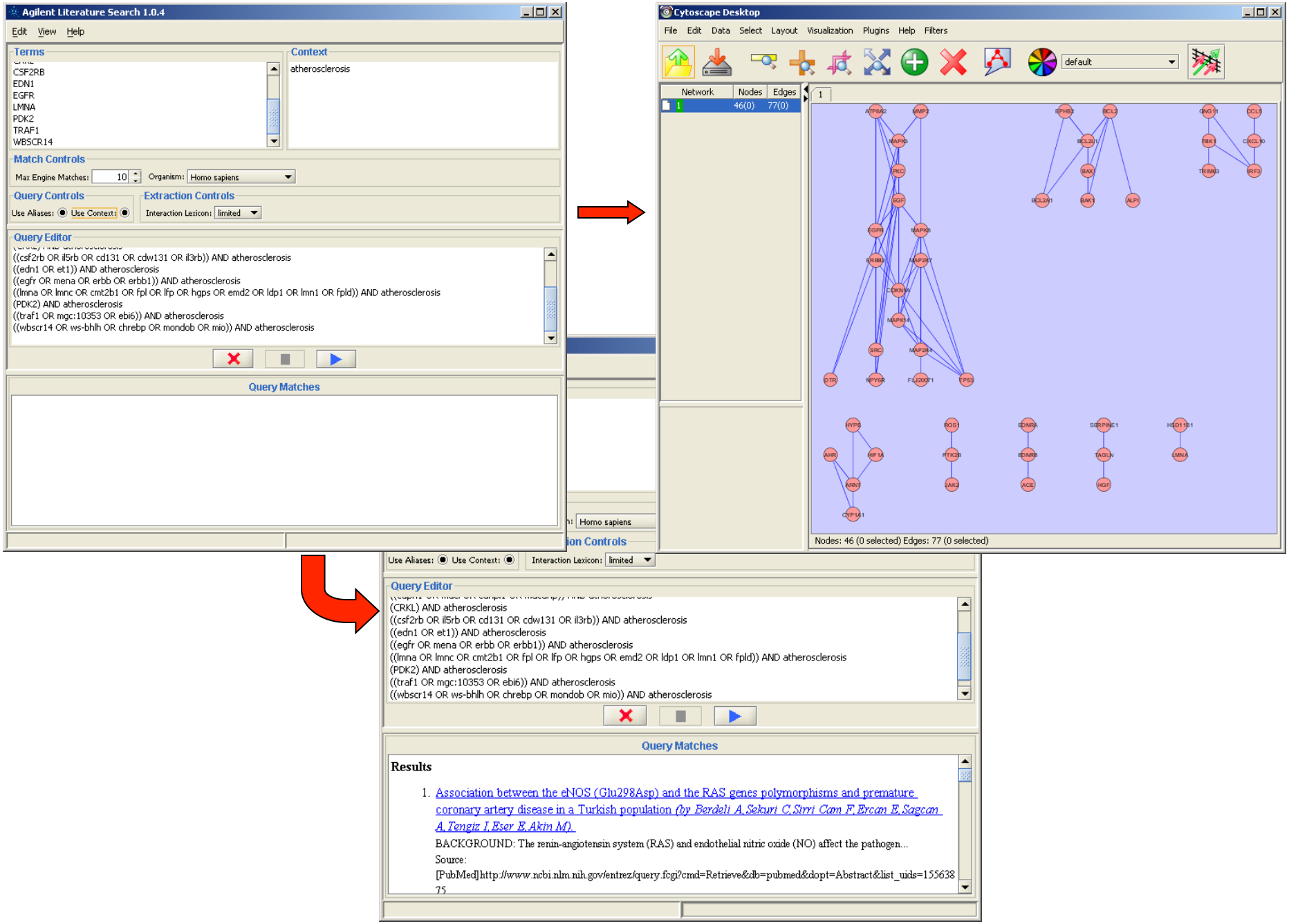



Cytoscape Desktop (New Session)
Search:
DNA_Repair_(Reactome)
Results Panel
Node Details
LIG1
Protein
Homo sapiens
Pathway Commons: 6311
Synonyms:
- LIG1
Links:
- UNIPROT: P18858
- UNIPROT: Q32P23
- REF_SEQ: NP_000225
- Search iHOP
Visual Legend
RFC140
RFC5
RFC2
RFC3
RFC4
LIG1
PRIM2A
PRIM1
PCNA
Processive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

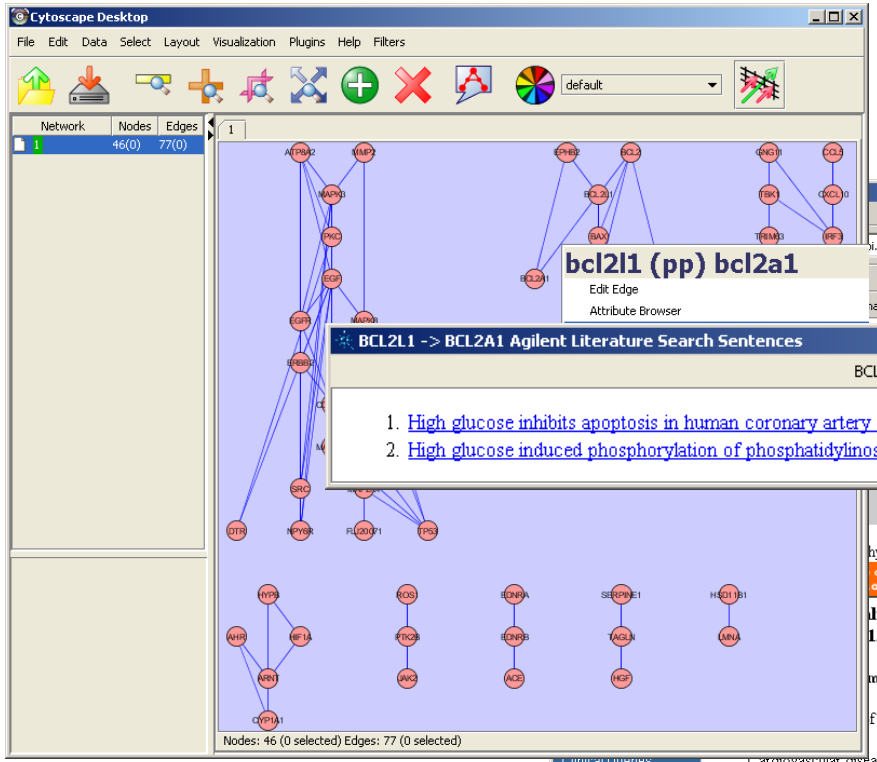

Cytoscape Desktop (New Session)
Search:
DNA_Repair_(Reactome)
Results Panel
Node Details
Processive complex:Okazaki fragment complex
Complex
Homo sapiens
Pathway Commons: 5256
Links:
– REACTOME: 68453
– REACTOME: REACT_5537
Visual Legend
http://reactome.org/cgi-bin/eventbrowser?DB=gk_current&ID=REACT_5537
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
XPCC
CCNH
MNAT1
BTF2
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/

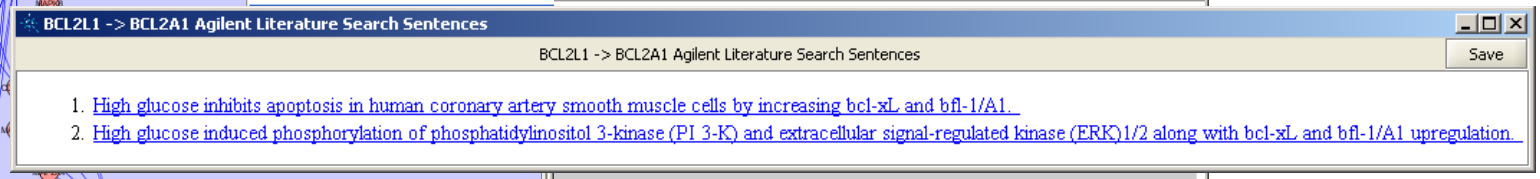

Agilent Literature Search 1.0.4
Edit View Help
Terms
CSF2RB
EDN1
EGFR
LMNA
PDK2
TRAF1
WBSCR14
Context
atherosclerosis
Match Controls
Max Engine Matches: 10
Organism: Homo sapiens
Query Controls
Use Aliases: Use Context:
Extraction Controls
Interaction Lexicon: limited
Query Editor
((csf2rb OR il5rb OR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



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.)

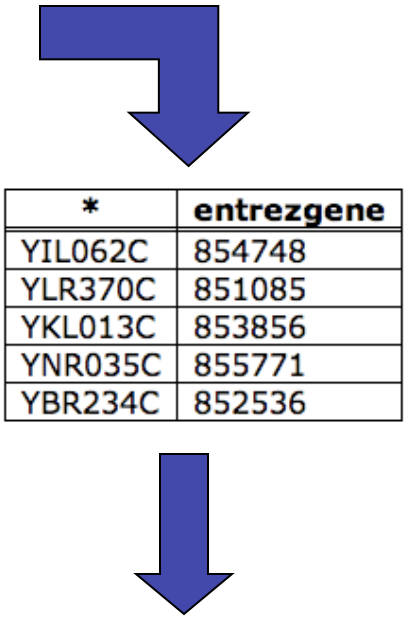
File containing IDs to translate: Browse...

and/or

IDs to translate:
YIL062C
YLR370C
YKL013C
YNR035C
YBR234C

Output as spreadsheet:

Submit

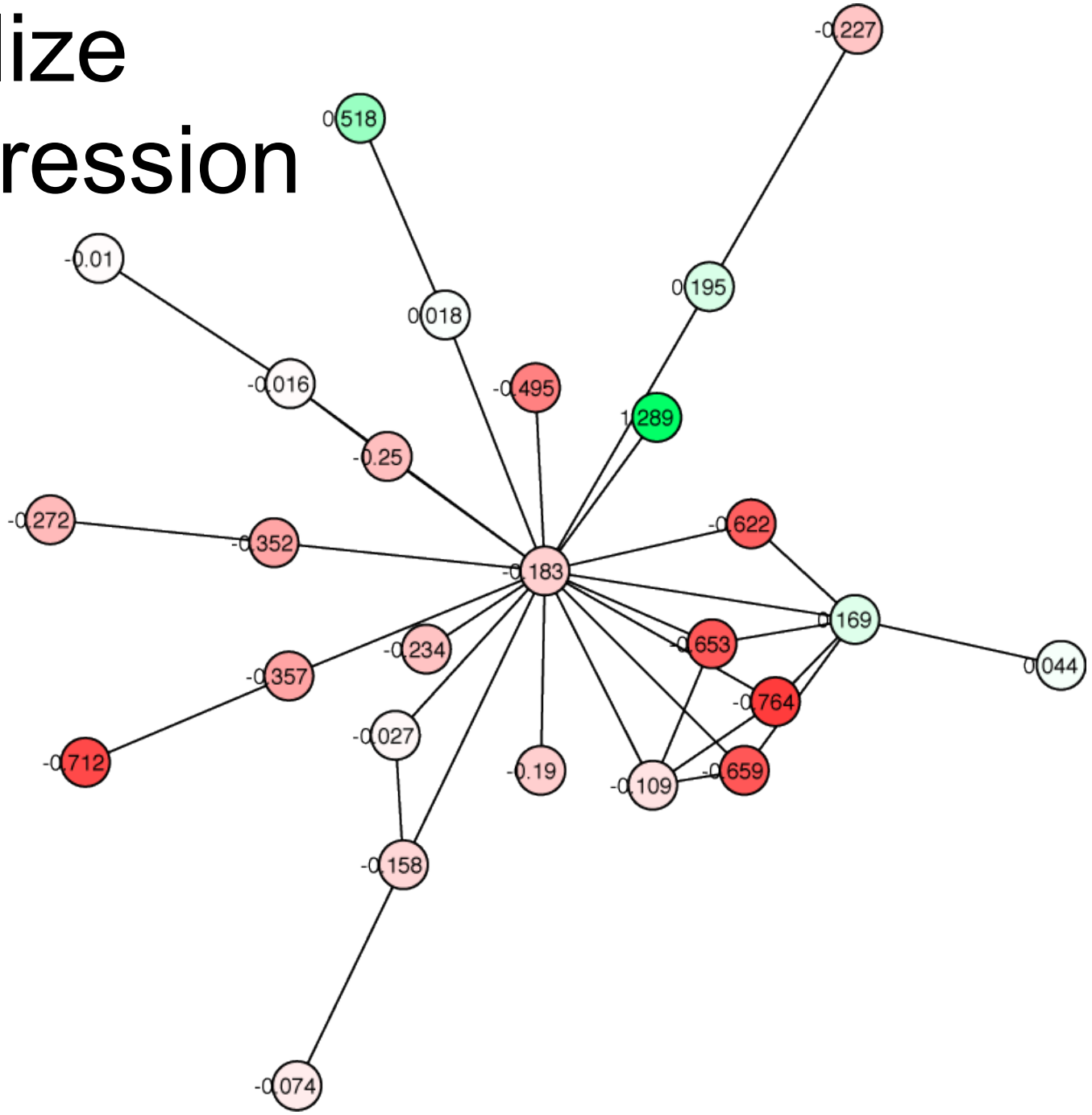


*	entrezgene
YIL062C	854748
YLR370C	851085
YKL013C	853856
YNR035C	855771
YBR234C	852536

1. Load as attributes in Cytoscape

2. Assign expression values to nodes using this attribute set

Visualize Gene Expression



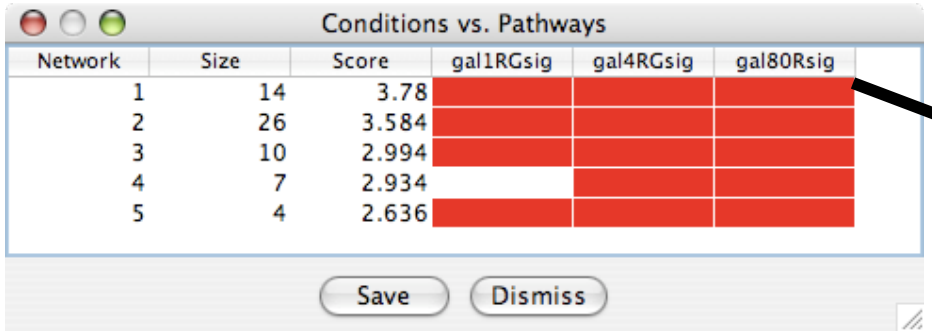
Find Active Subnetworks

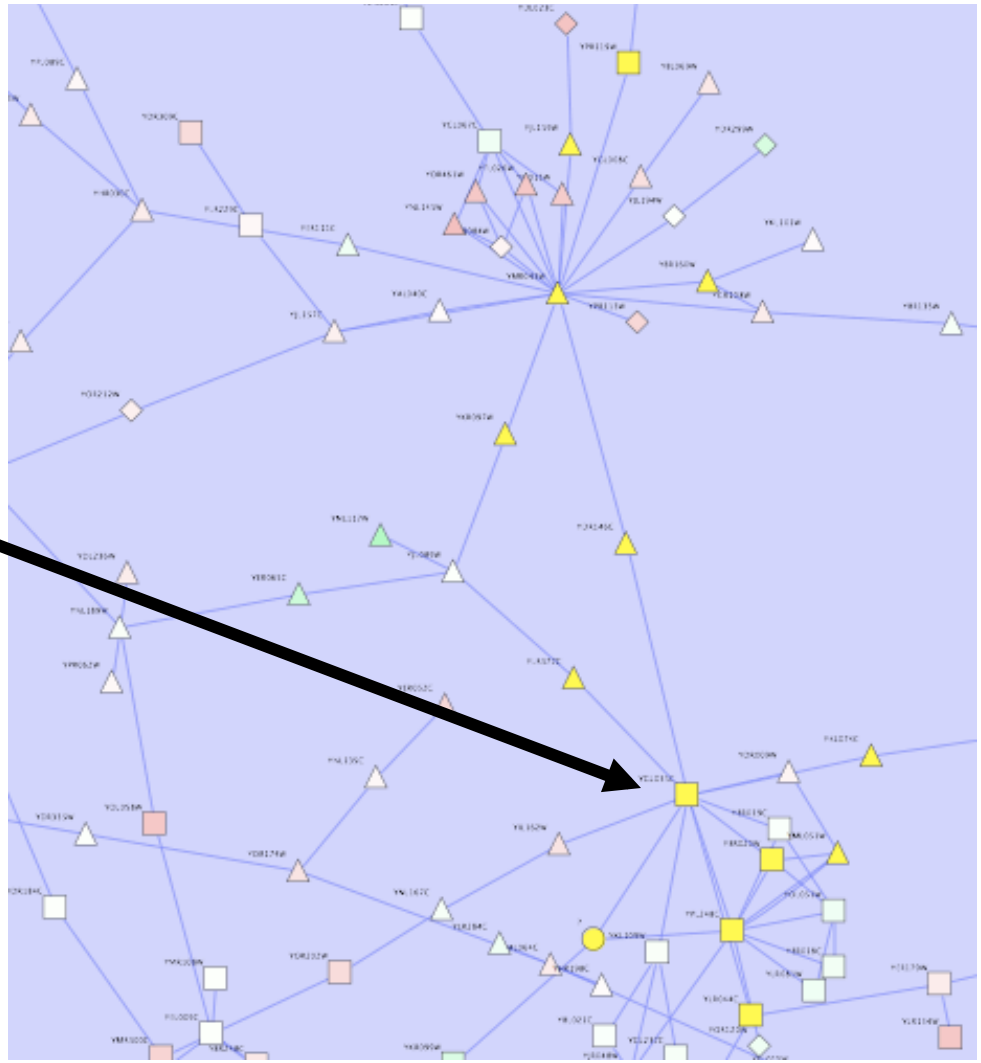
- Active modules
  - Input: network + p-values for gene expression values e.g. from GCRMA
  - Output: significantly differentially expressed subgraphs
- Method
  - Calculate z-score/node, Z_A score/subgraph, correct vs. random expression data sampling
  - Score over multiple experimental conditions
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



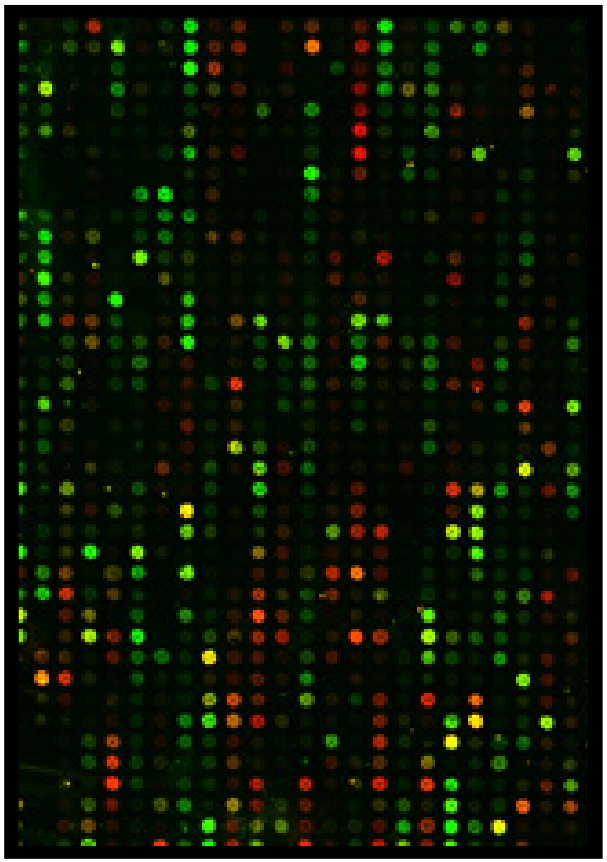


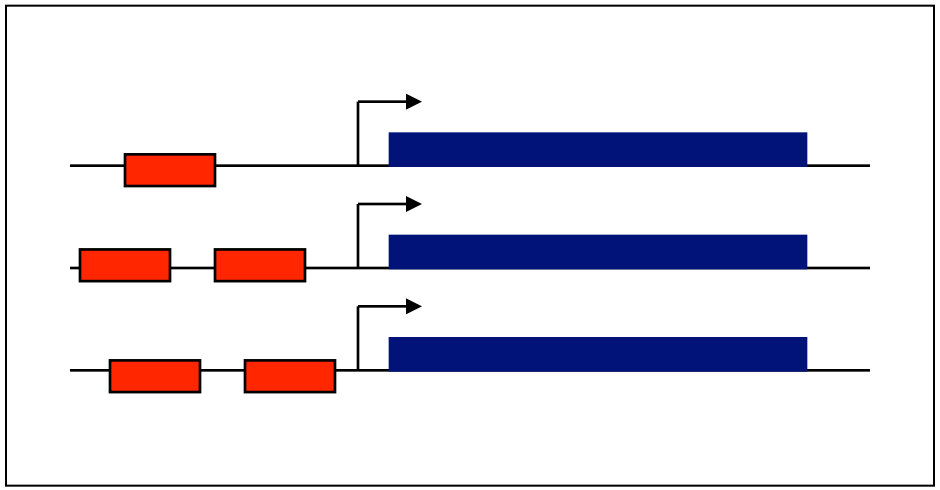
Note: non-deterministic, multiple runs required for confidence of result robustness

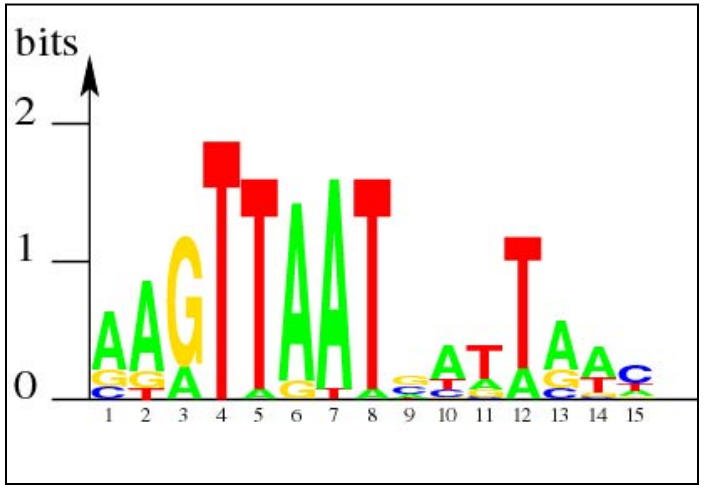
Ideker T et al. Science. 2001 May 4;292(5518):929-34.

de novo Discovery of TF Binding Sites

de novo Pattern Discovery







de novo Pattern Discovery

- Classic methods used string counting
  - String-based method has found renewed utility in the analysis of 3'UTRs for the presence of microRNA target sequences

- Most TF pattern studies now focusing on Profile-based Methods
  - e.g. MEME (Bailey & Elkin) or MEME-ChIP
  - Generalization: Identify strong patterns in "+" promoter collection vs. background model of expected sequence characteristics

Comments about String-based Methods

- While degeneracy codes can be used, TFBS are not words – we lose quantitation for variable positions with consensus sequences
    - Imagine a column within a PFM with 7 A's and 1 T --- in a consensus sequence we represent it as W or ignore the rare T
- In a benchmarking study of pattern discovery methods, some of the best performers were string-based

- A high-quality string-based method is Weeder
    - http://159.149.109.9:8080/weederweb2006/input.faces;jsessionid=10AD062F5E94860FA320631B56EBB672