Building a digital systems microscope

Benno Schwikowski
Systems Biology Lab
Institut Pasteur, Paris

Research questions

Why is person A severely affected by dengue virus infection, and person B not?

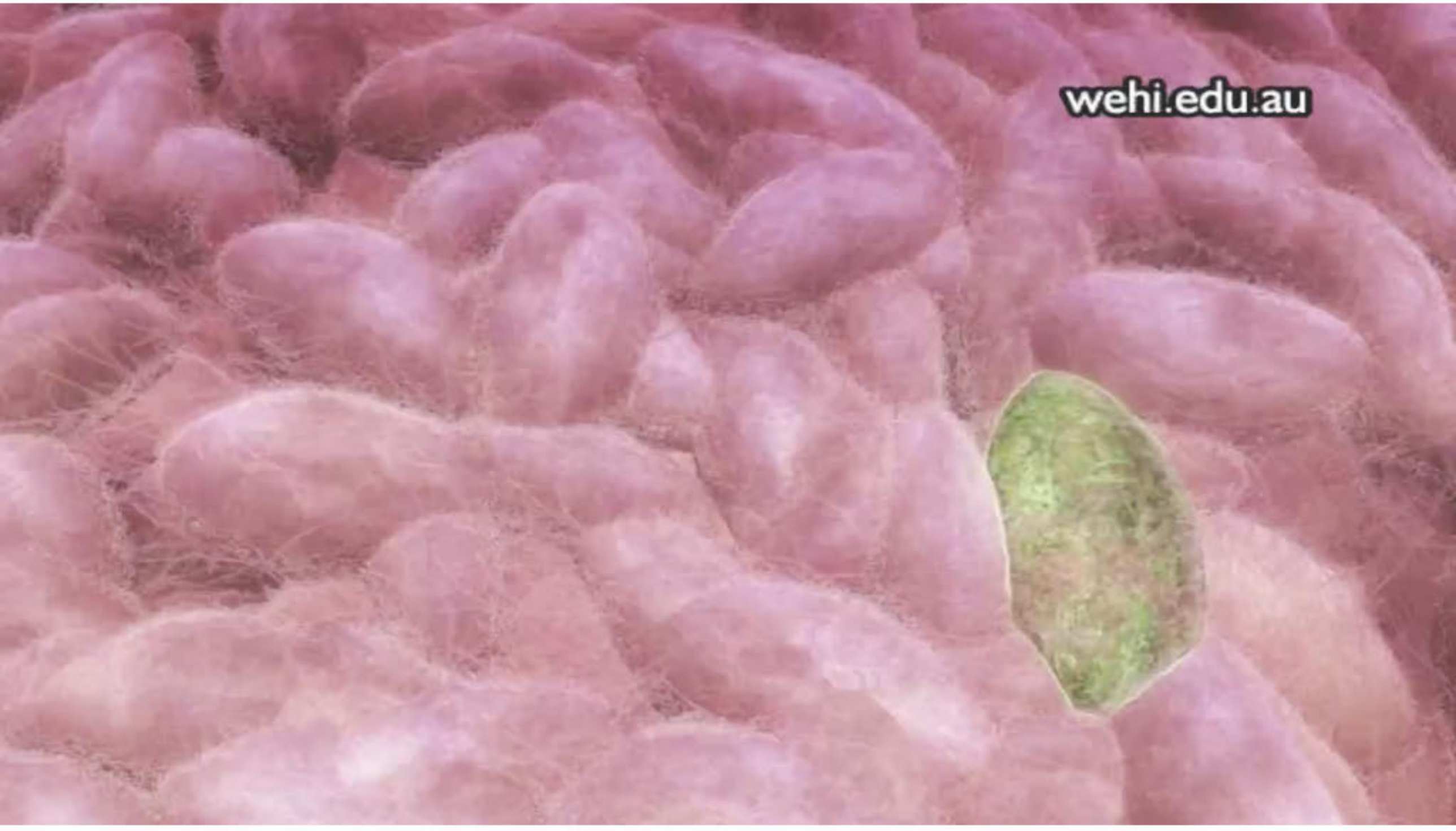
Can we *predict* whether a drug will work for A? For B?

How about drug *combinations*?

Why do more and more people become allergic to their environment?

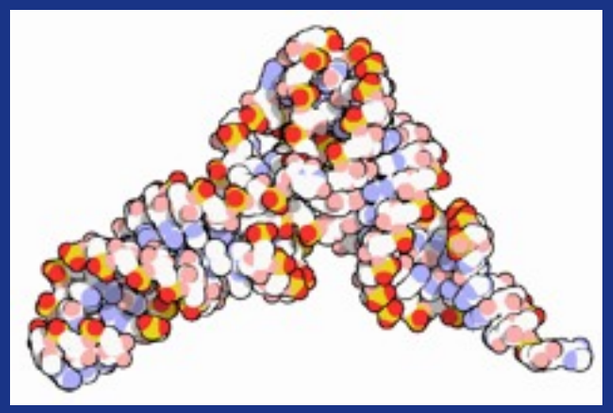
How to diagnose breast cancer *before* it starts?

Can we find the better drug targets against malaria/HIV/ tuberculosis/Alzheimer's/colon cancer/...

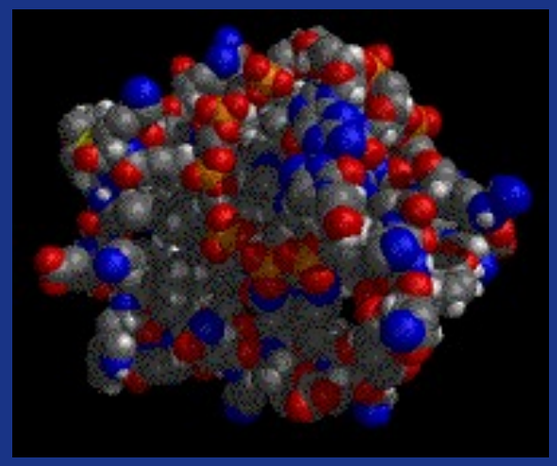

wehi.edu.au

Systems Biology:
From molecules to systems

DNA / genes



RNA

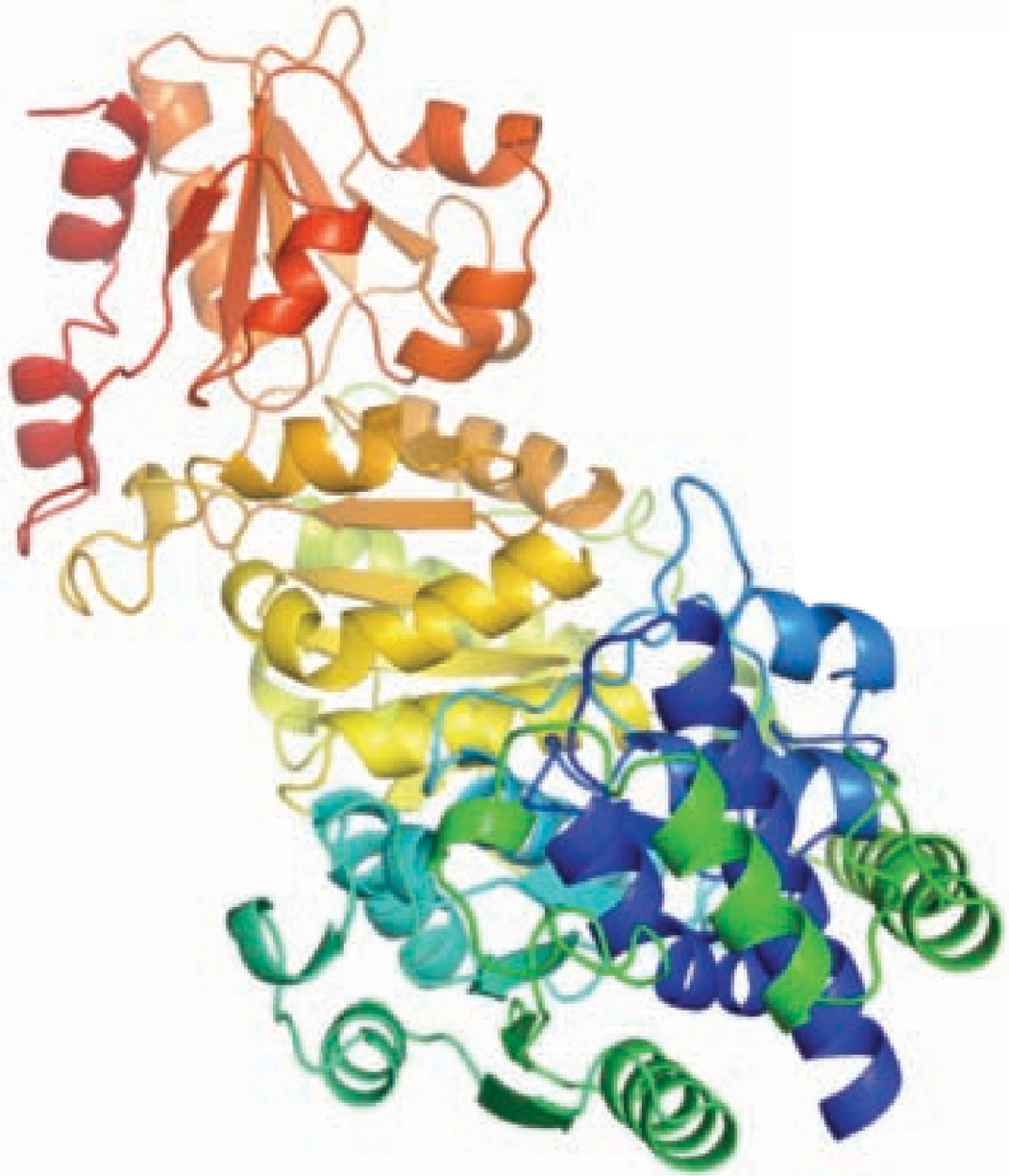


Protein

– Experimental technologies & data
– Computational models

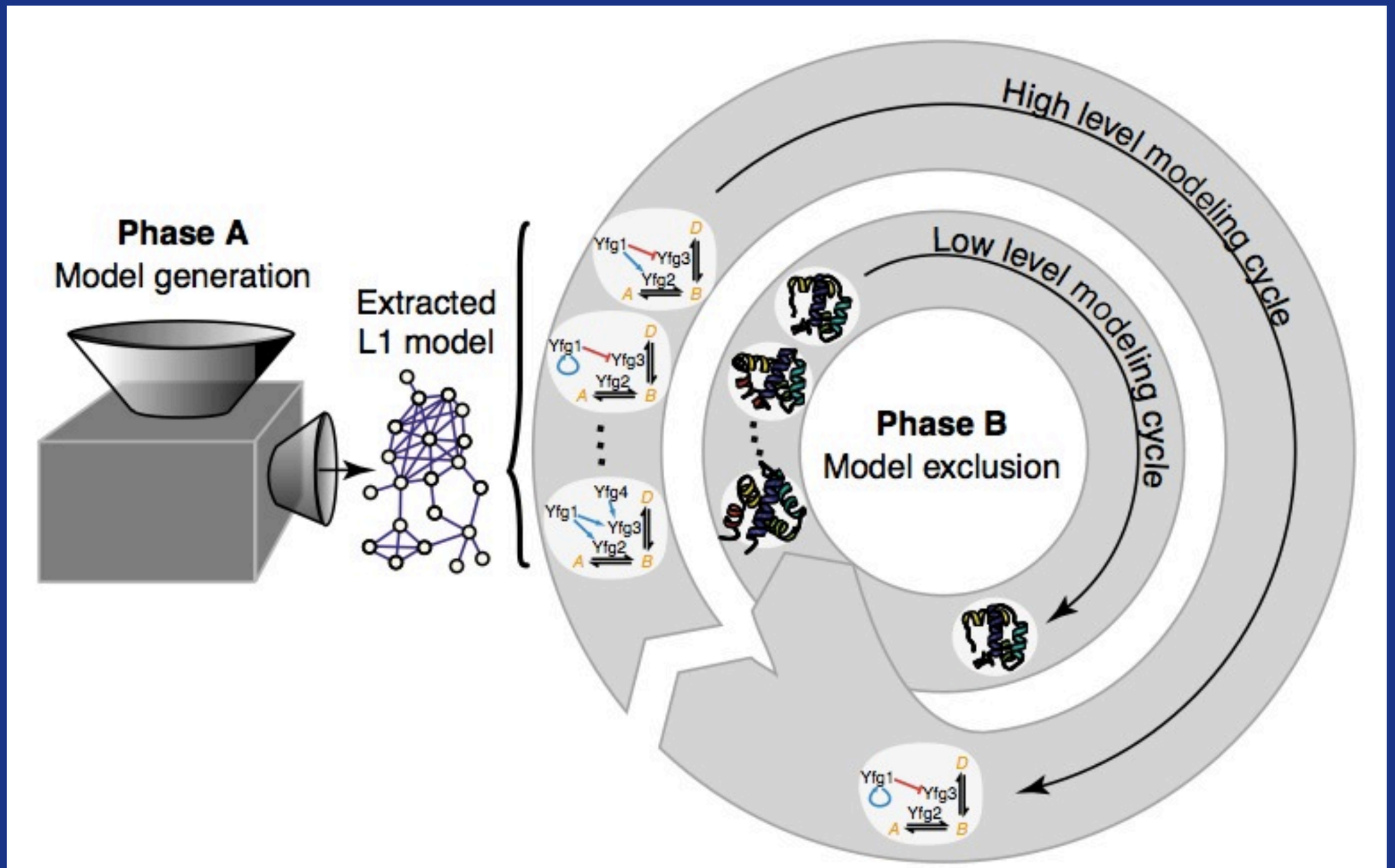
INSTITUT PASTEUR

Benno Schwikowski



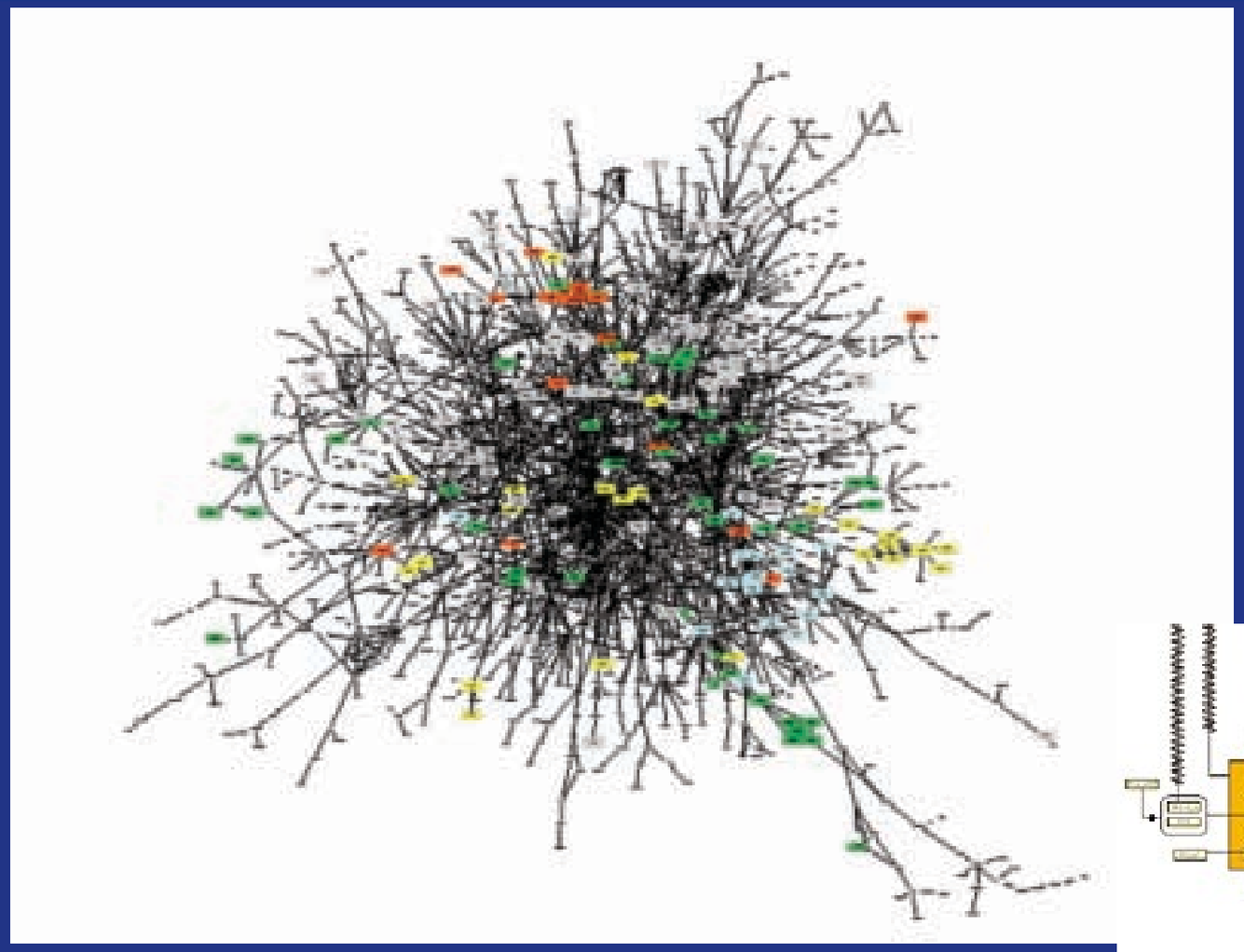
Benno Schwikowski

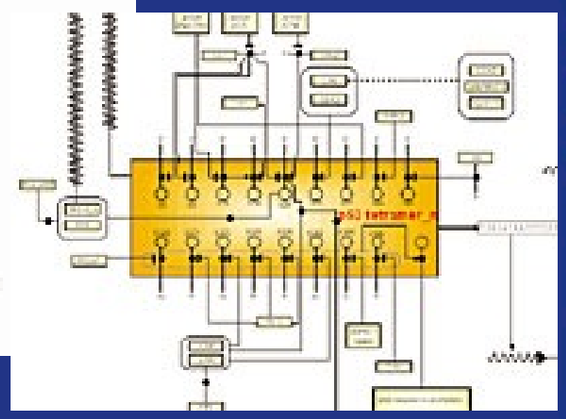
High-level and low-level modeling



Ideker and Lauffenburger (2003), TiBS

What is complex about biological systems?

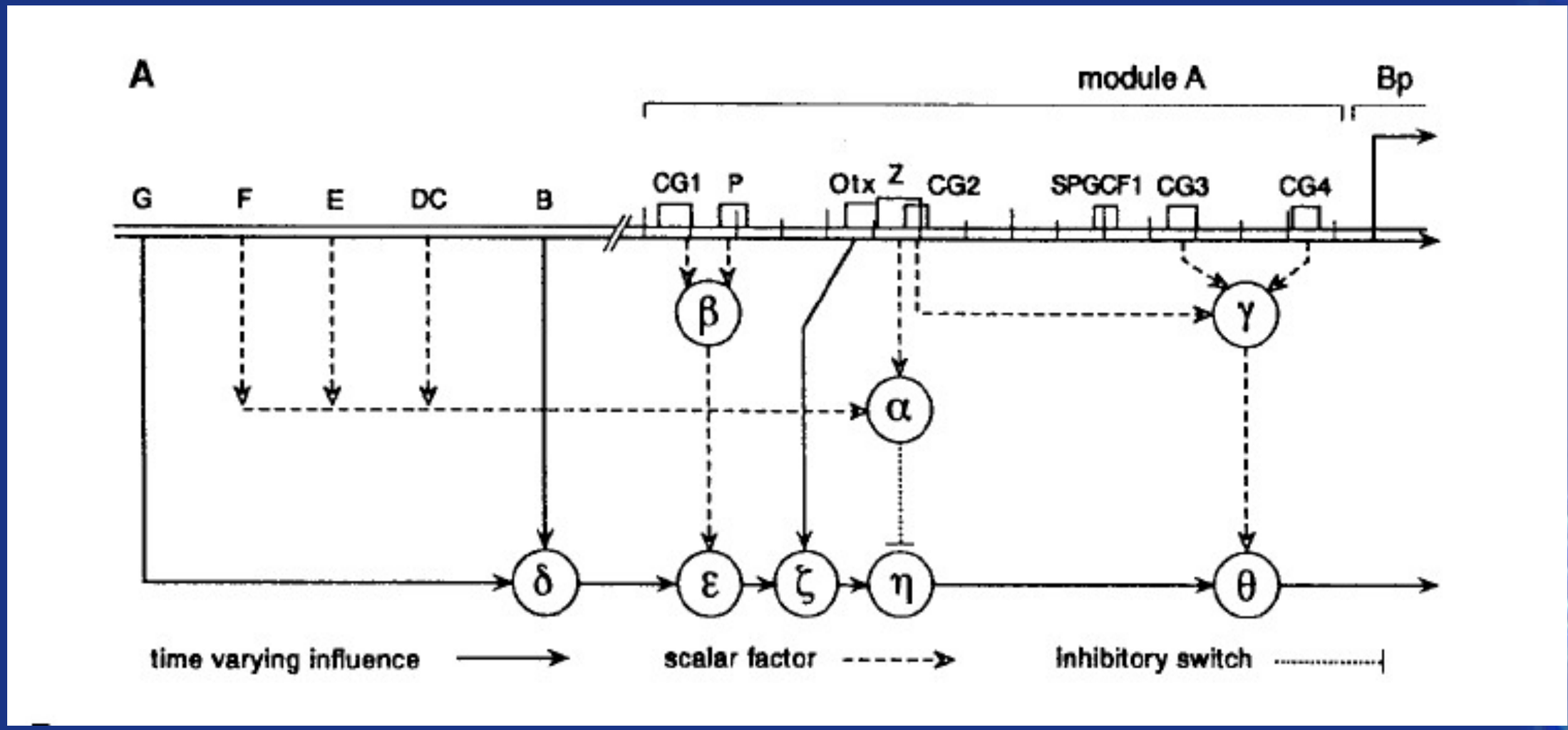




1. Biological parts interact in large networks.

INSTITUT PASTEUR
Benno Schwikowski

What is complex about biological systems?



Yuh, Bolouri, Davidson (1998), Science

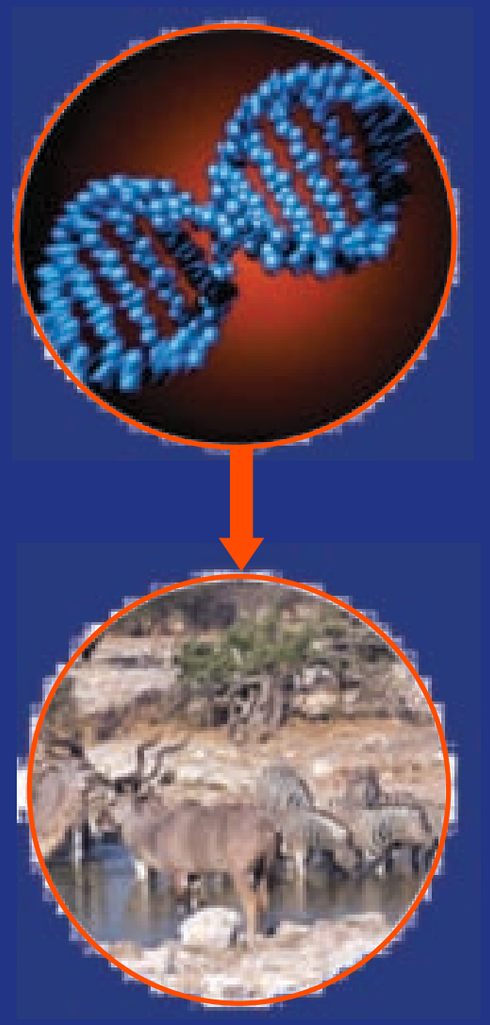
What is complex about biological systems?

B

if (F = 1 or E = 1 or CD = 1) and (Z = 1)	Repression functions of modules F, E, and DC mediated by Z site
$\alpha = 1$	
else $\alpha = 0$	
if (P = 1 and CG_1 = 1)	Both P and CG_1 needed for synergistic link with module B
$\beta = 2$	
else $\beta = 0$	
if (CG_2 = 1 and CG_3 = 1 and CG_4 = 1)	Final step up of system output
$\gamma = 2$	
else $\gamma = 1$	
$\delta(t) = B(t) + G(t)$	Positive input from modules B and G
$\varepsilon(t) = \beta^*\delta(t)$	Synergistic amplification of module B output by CG_1-P subsystem
if ($\varepsilon(t) = 0$)	Switch determining whether Otx site in module A, or upstream modules (i.e., mainly module B), will control level of activity
$\xi(t) = Otx(t)$	
else $\xi(t) = \varepsilon(t)$	
if ($\alpha = 1$)	Repression function inoperative in endoderm but blocks activity elsewhere
$\eta(t) = 0$	
else $\eta(t) = \xi(t)$	
$\Theta(t) = \gamma^*\eta(t)$	Final output communicated to BTA

What is complex about biological systems?

DNA
mRNA
Proteins
Pathways/Modules
Cells
Tissues
Organs
Individuals
Populations
Ecosystems



2. Different levels interact.

INSTITUT PASTEUR

Benno Schwikowski

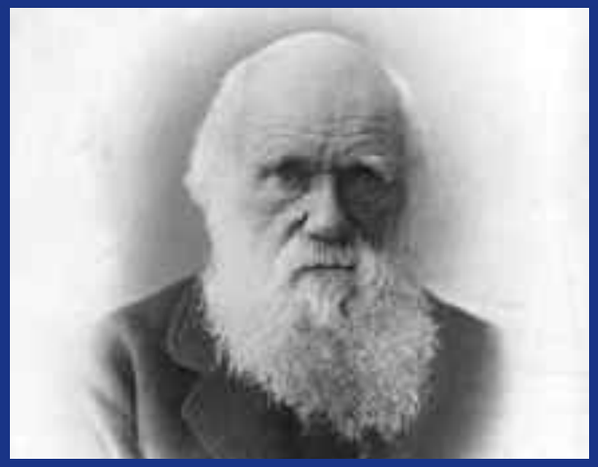
L. Hood

What is complex about biological systems?

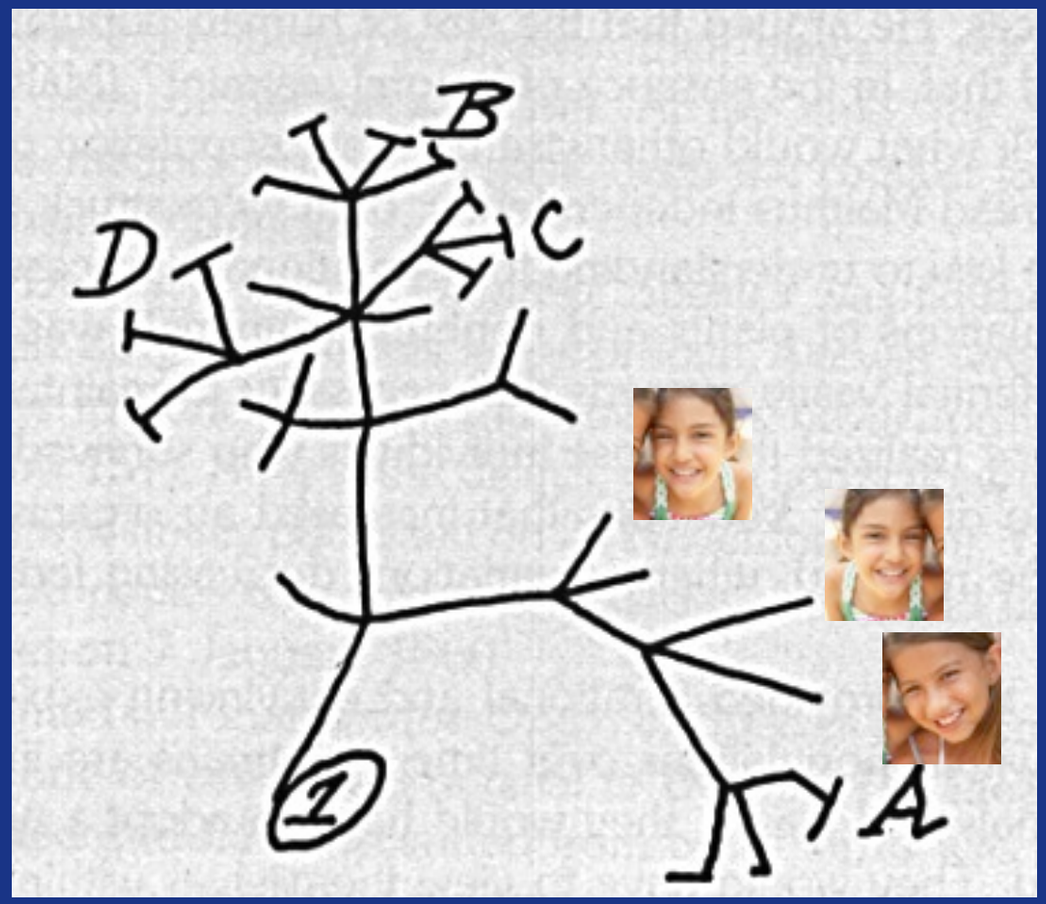
Benno Schwikowski

L. Hood

What is complex about biological systems?



Charles Darwin

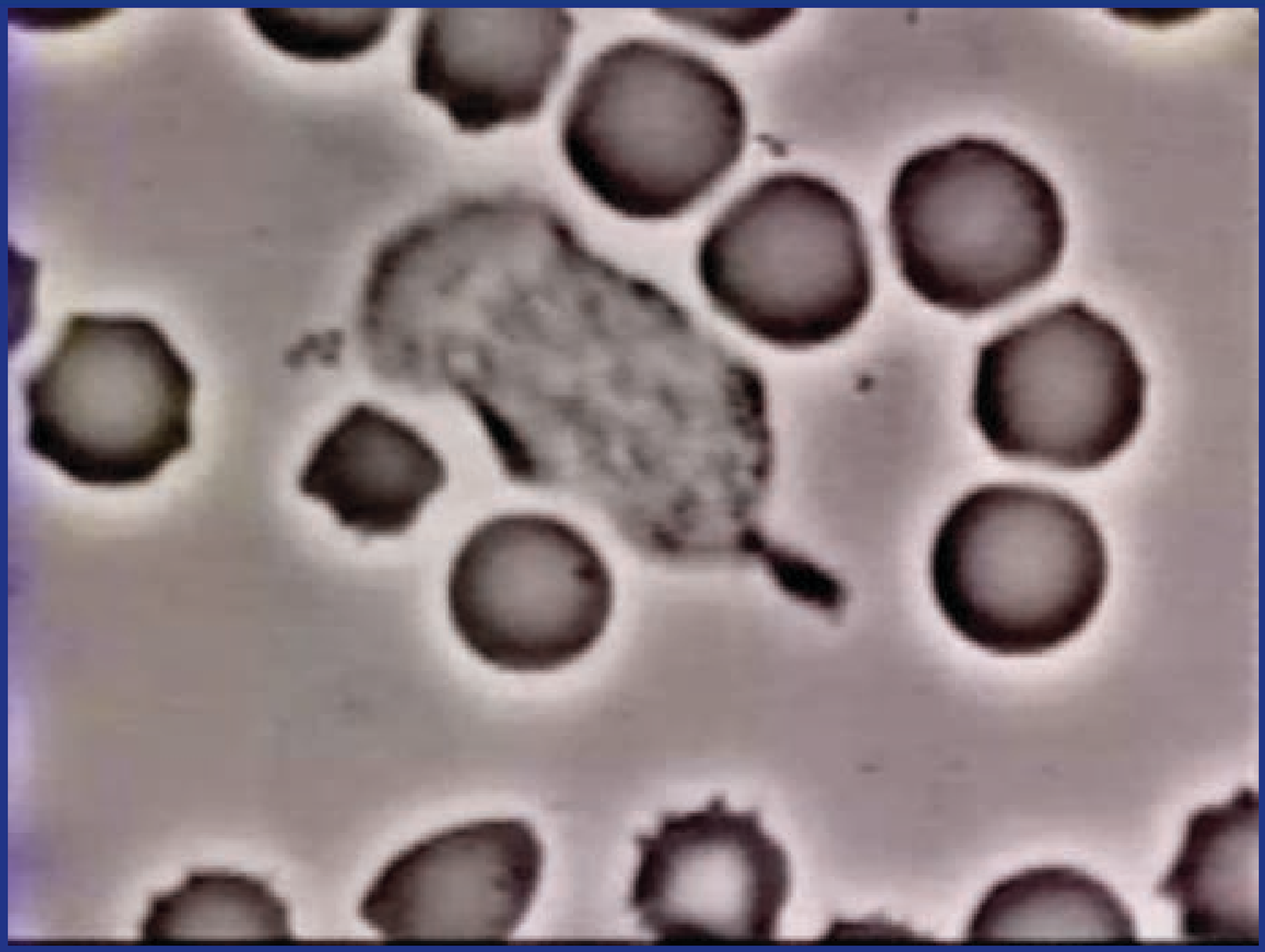


1837

3. Different parts are related by evolution.

INSTITUT PASTEUR
Benno Schwikowski

What is complex about biological systems?

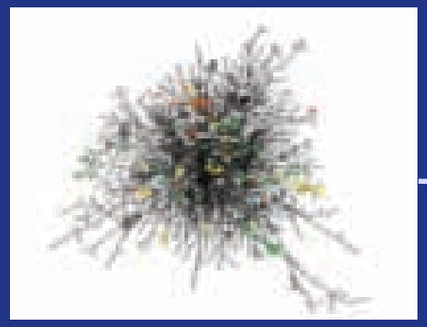


4. Interacting timescales.

INSTITUT PASTEUR
Benno Schwikowski

What is complex about biological systems?

Four nonlinear contexts.

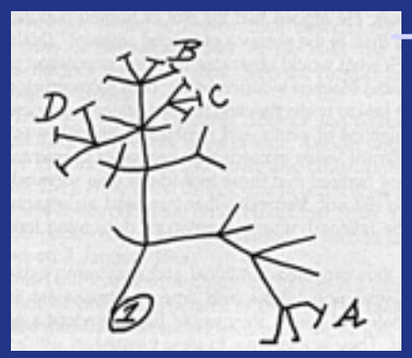


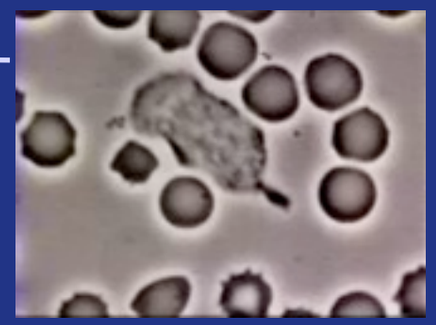
Genes

Levels

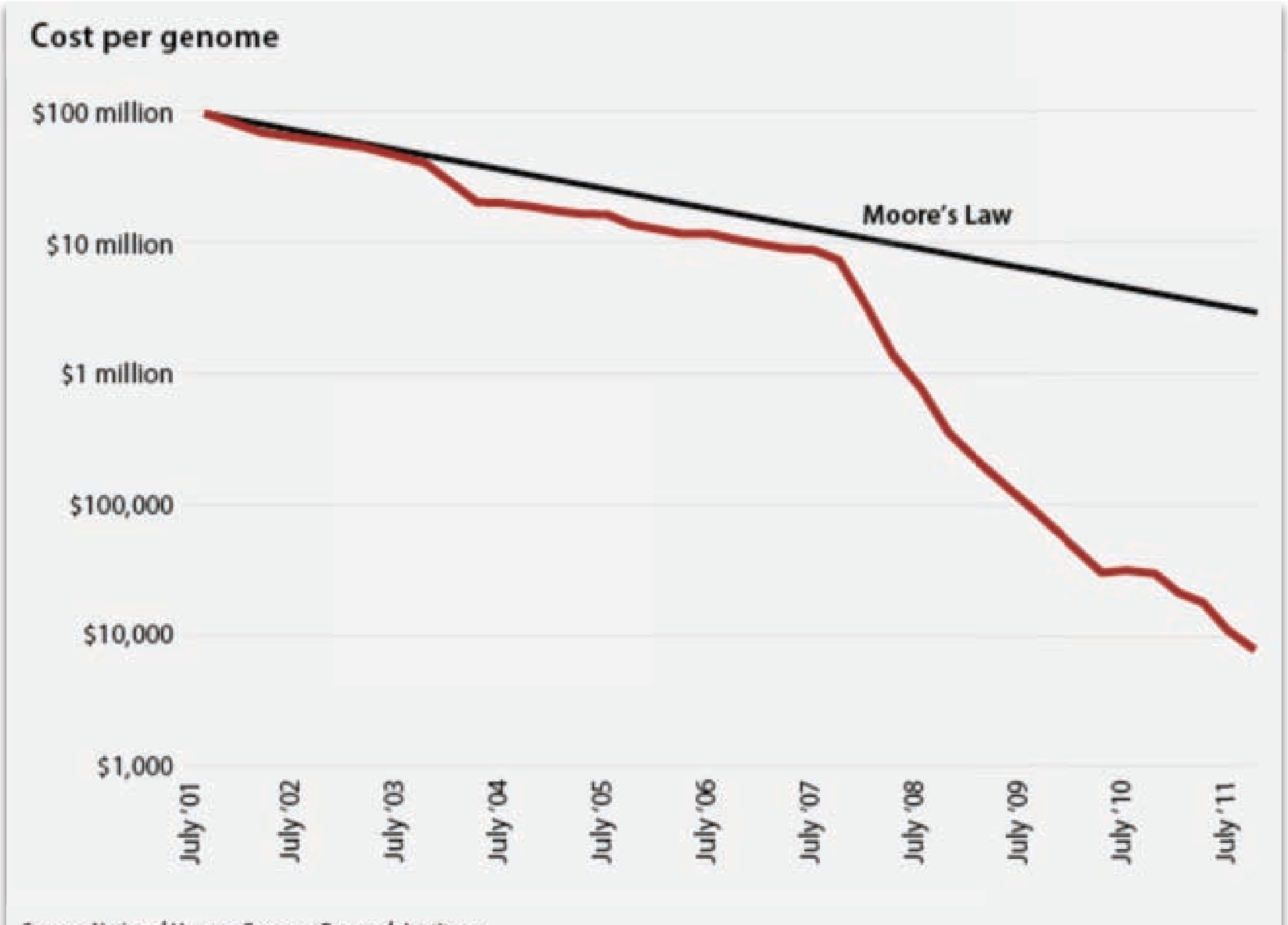
Evolution

Timescales

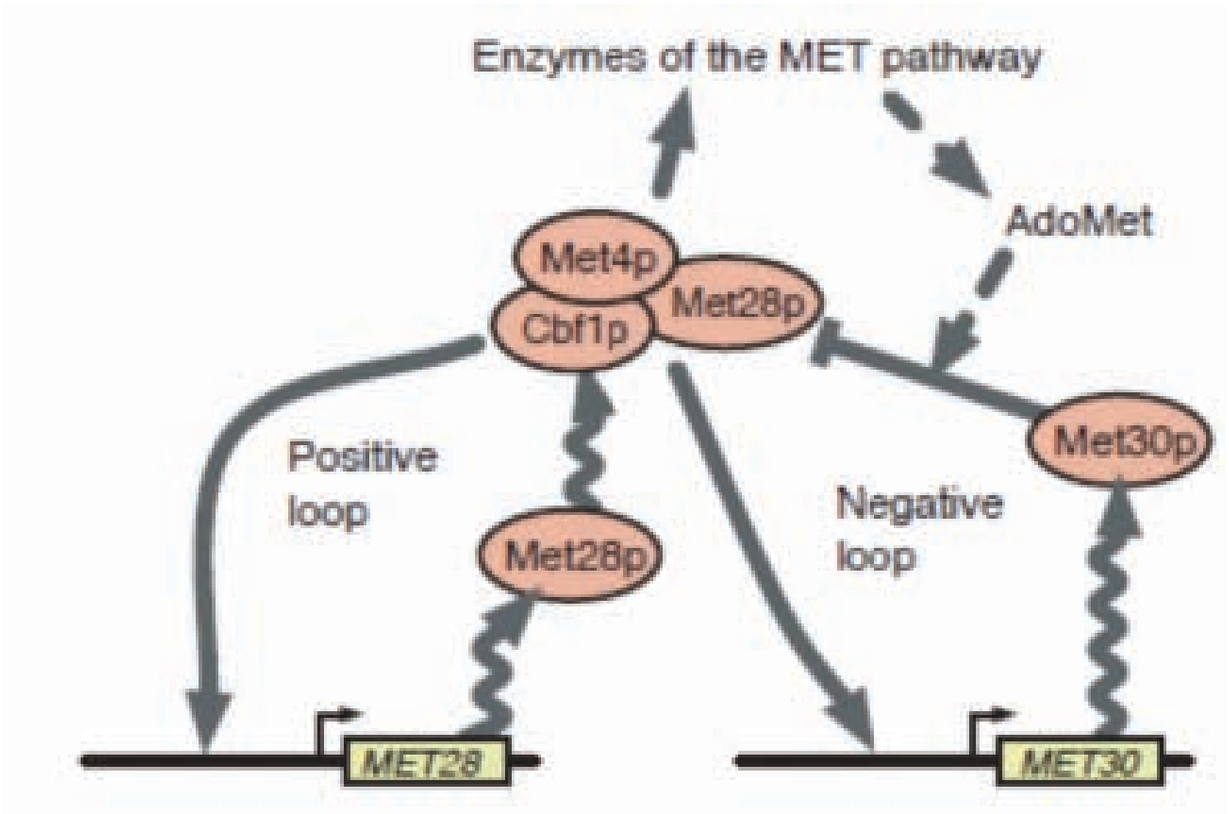


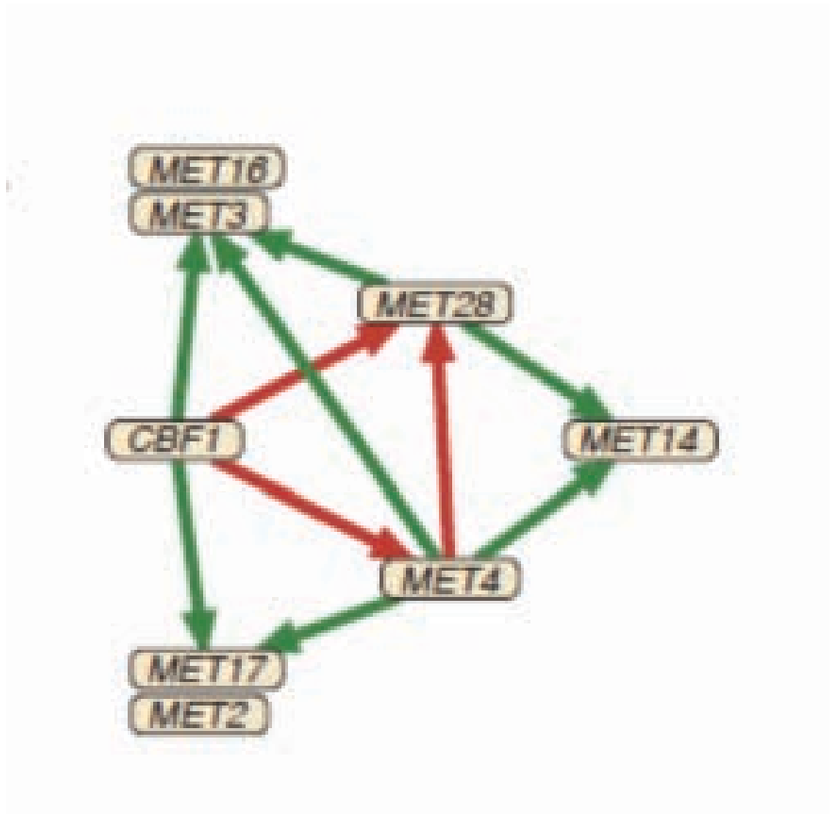


DNA/RNA sequencing technology



Textbook dream and large-scale reality

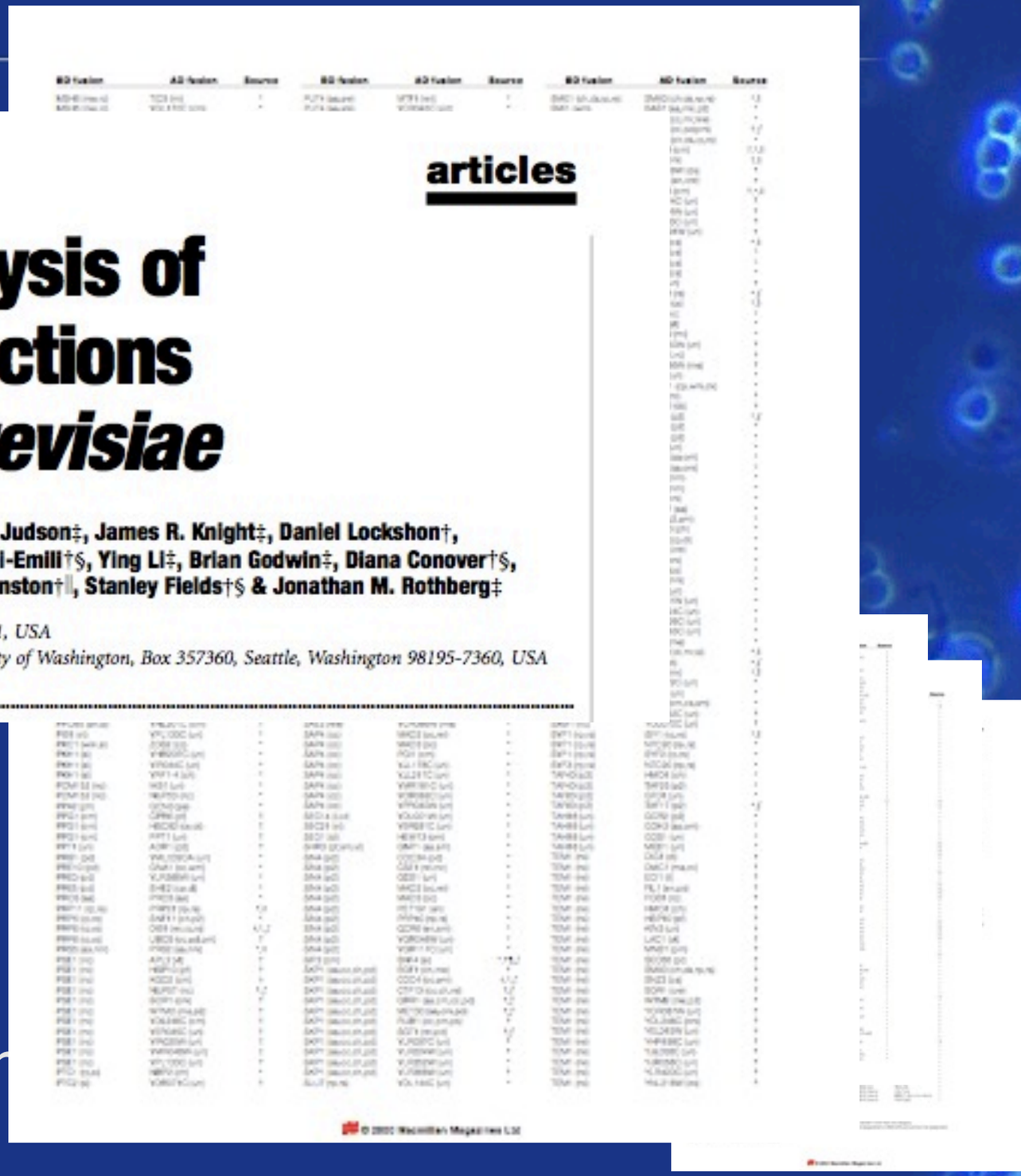




Mazurie, Bottani, Vergassola (2005), Genome Biology

Visualization of interaction networks

The first large-scale PPI dataset



articles

A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*

Peter Uetz*†, Loic Giot*‡, Gerard Cagney†, Traci A. Mansfield‡, Richard S. Judson‡, James R. Knight‡, Daniel Lockshon†, Vaibhav Narayan‡, Maithreyan Srinivasan‡, Pascale Pochart‡, Alia Qureshi-Emili†§, Ying Li‡, Brian Godwin‡, Diana Conover†§, Theodore Kalbfleisch‡, Govindan Vijayadamodar‡, Meijia Yang‡, Mark Johnston†‖, Stanley Fields†§ & Jonathan M. Rothberg‡

‡ CuraGen Corporation, 555 Long Wharf Drive, 11th Floor, New Haven, Connecticut 06511, USA
† Departments of Genetics and Medicine and § Howard Hughes Medical Institute, University of Washington, Box 357360, Seattle, Washington 98195-7360, USA
** These authors contributed equally to this work*

- ~50% false positives/ false negatives
- Is this data good for anyth

INSTITUT PASTEUR
Benno Schwikowski

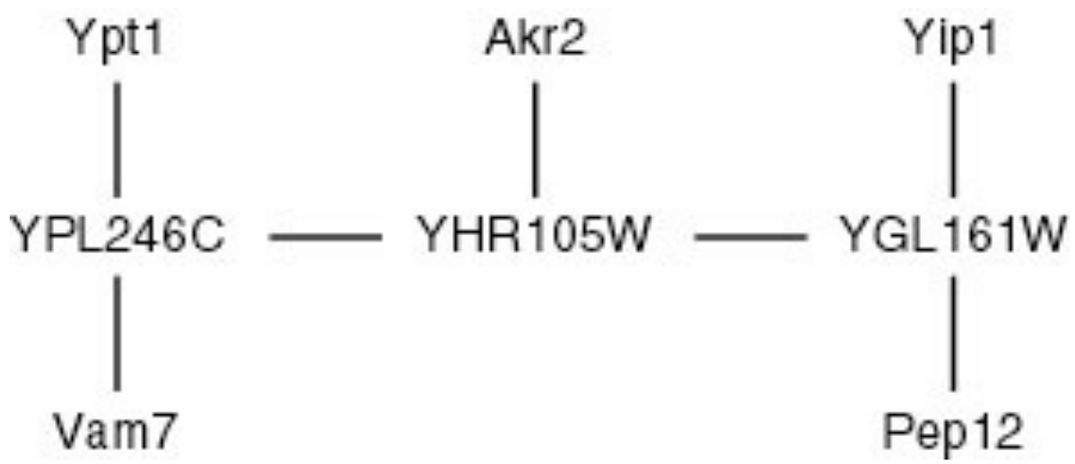
Uetz et al. (2000) Nature.

Advantages of graphical representation

Linear

Ypt1 — YPL246C
Akr2 — YHR105W
Yip1 — YGL161W
YPL246C — Vam7
YGL161W — Pep12
YPL246C — YHR105W
YHR105W — YGL161W

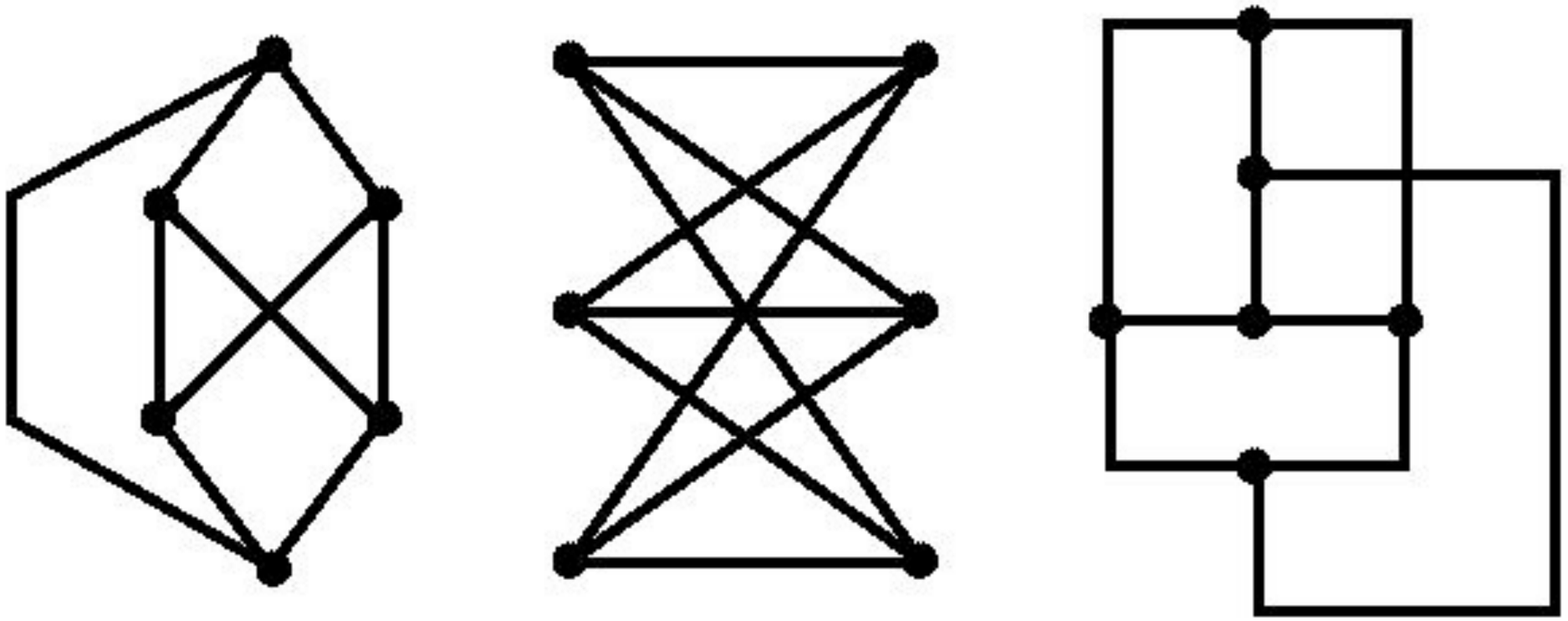
Graphical



Questions to assess usefulness

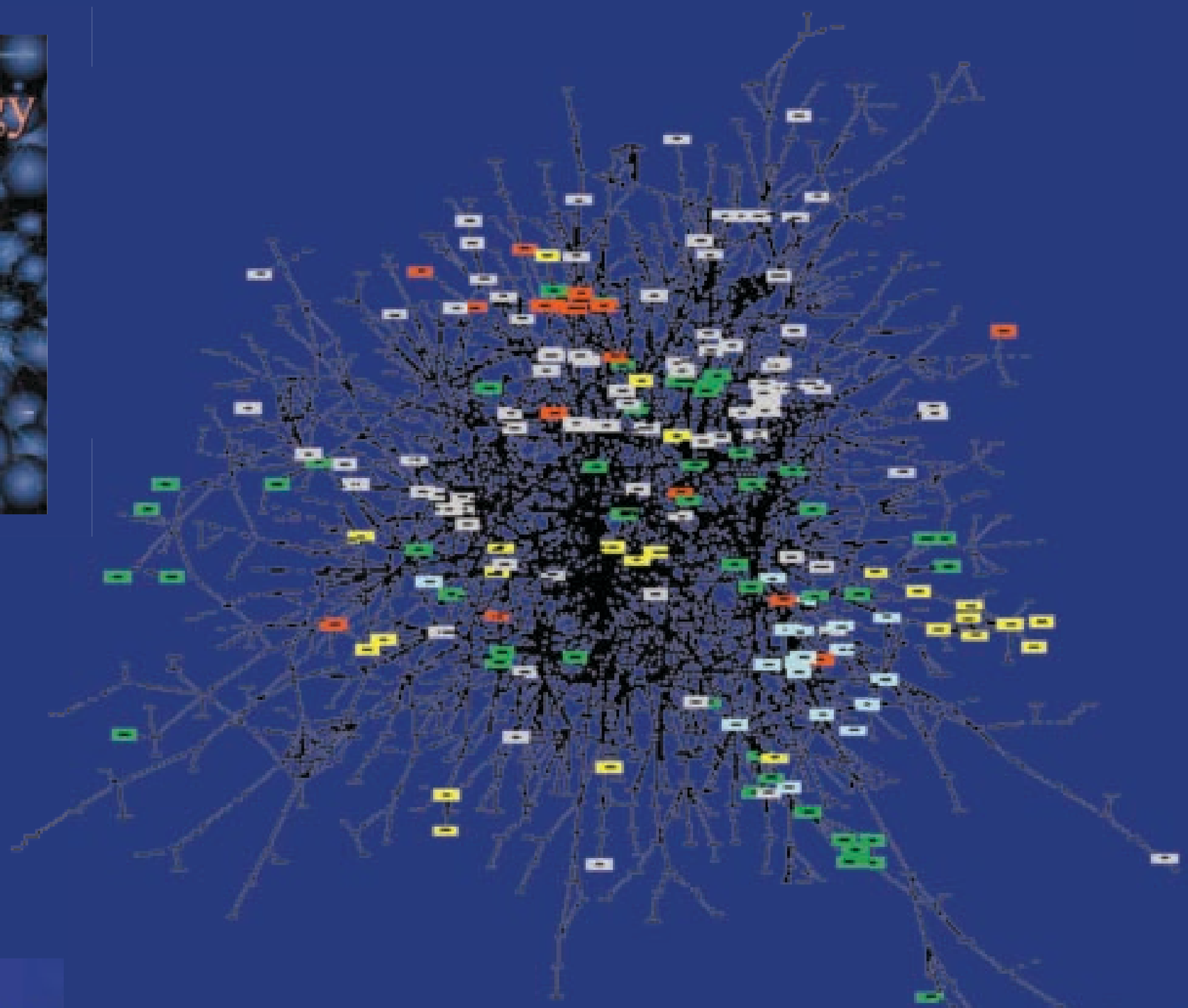
- How many interactions involve YHR105W?
- Are YHR105W and Pep12 closely related?

One graph, many visualizations



(a) (b) (c)

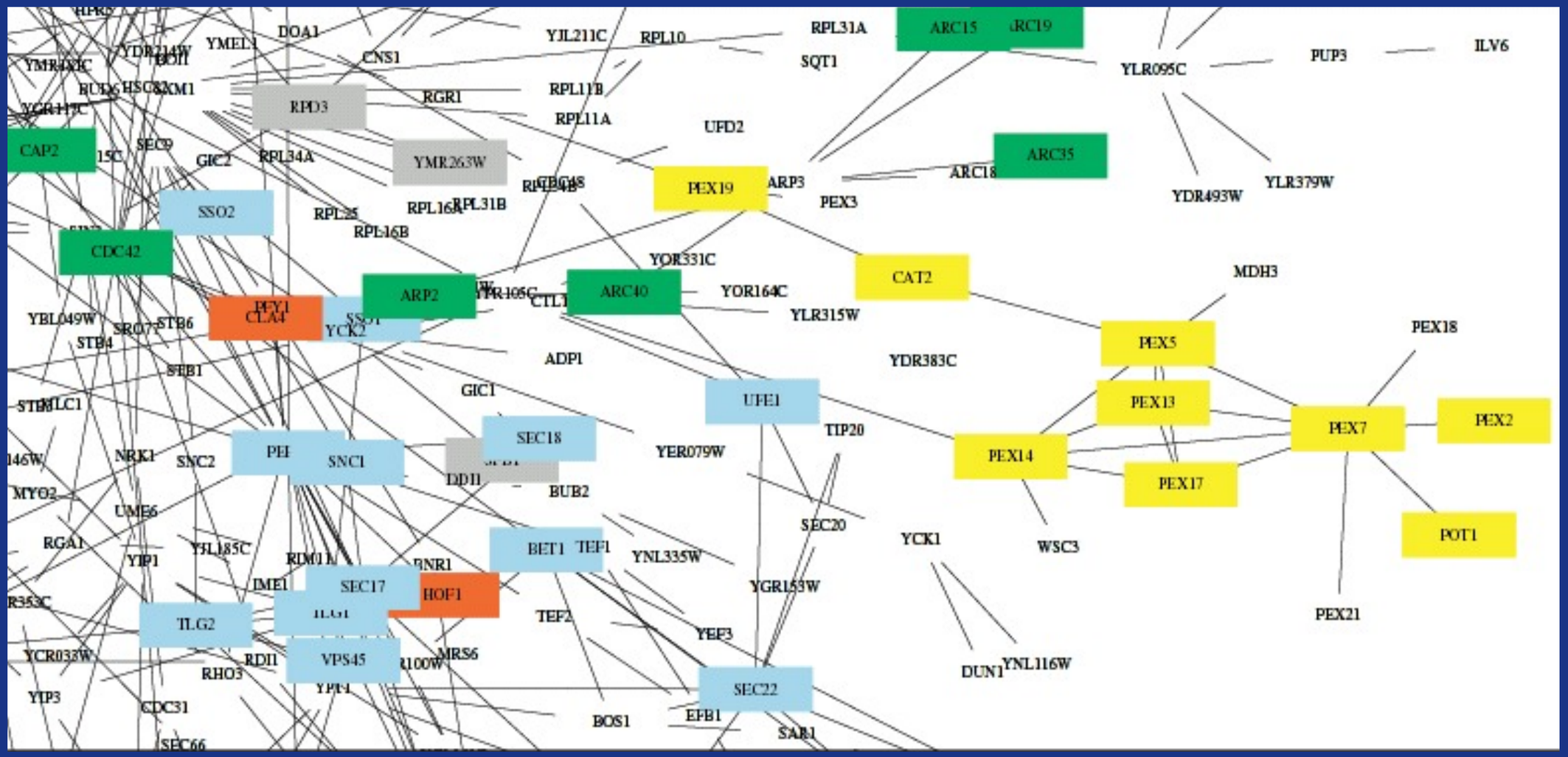
Visualization



Schwikowski, Uetz, Fields (2000) Nat. Biotech.

INSTITUT PASTEUR
Benno Schwikowski

Visualization



INSTITUT PASTEUR

Benno Schwikowski

Schwikowski, Uetz, Fields (2000) Nat. Biotech.

Statistical analysis

- Correlation interacting proteins/function?
- Is that correlation statistically *significant*?
- Exploitable for function prediction?

Schwikowski, Uetz, Fields (2000) Nat. Biotech.

INSTITUT PASTEUR
Benno Schwikowski

Protein interactions contain information about cellular roles

Simple prediction algorithm for the cellular role of a protein

1) Rank known cellular roles among the interactors
   from most frequent to least frequent.

2) Take the first three (or less) roles as predictions.

Accuracy on 1,393 out of 2,039 proteins: 72% (6 out of 8)
…on 100 scrambled networks: 12% (1 out of 8).

INSTITUT PASTEUR
Benno Schwikowski

Schwikowski, Uetz, Fields (2000) Nat. Biotech.

Protein networks—built by association

Melanie L. Mayer and Philip Hieter

The total genome sequence era has provided complete catalogs of the genes of several organisms and offered the challenge of understanding the functions of large numbers of previously uncharacterized proteins. Historically, the functions of genes (i.e., their encoded proteins) have been determined by analysis of mutant phenotypes, genetic interactions, biochemical activities, homology to other proteins of known function, and physical interactions with other proteins. Schwikowski et al.[1] have compiled comprehensive protein–protein interaction data sets from the yeast community and find that these interactions form one large network of 2,358 interactions among 1,548 proteins and several smaller networks. Analysis of these networks allows assignment of potential function to uncharacterized proteins and the discovery of potential interactions within and across cellular processes and compartments. These connections represent a gold mine for formulating and experimentally testing specific hypotheses about gene function.

The total genome sequence era has also made possible the ongoing development (and validation) of methodologies that address gene function on a genome-wide scale (functional genomics)[2]. Several new approaches are aimed at determining the function of large sets of proteins and defining how these macromolecules interact within complex networks. These include computational biology driven approaches, such as correlated phylogenetic profiles (which predict that proteins that function in a common pathway or complex will evolve in a similar fashion and be either preserved or eliminated in a given genome)[3], structure-based functional genomics (which aims to assign functions to uncharacterized proteins based on structure prediction), and the analysis of domain fusion events (which is based on the premise that two domains that are fused in one organism are likely to interact in another organism in which both domains are in separate proteins)[4,5]. Functional assignments for newly discovered proteins have also been made by partnering them with proteins of "known" function by analyzing large

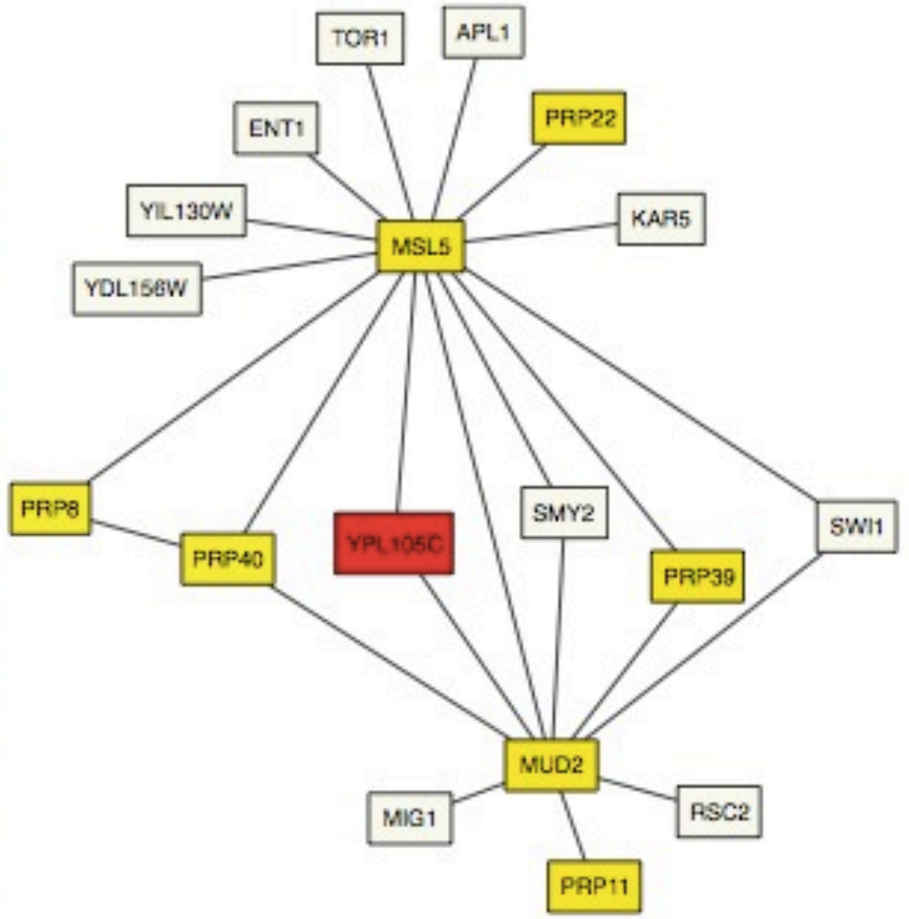
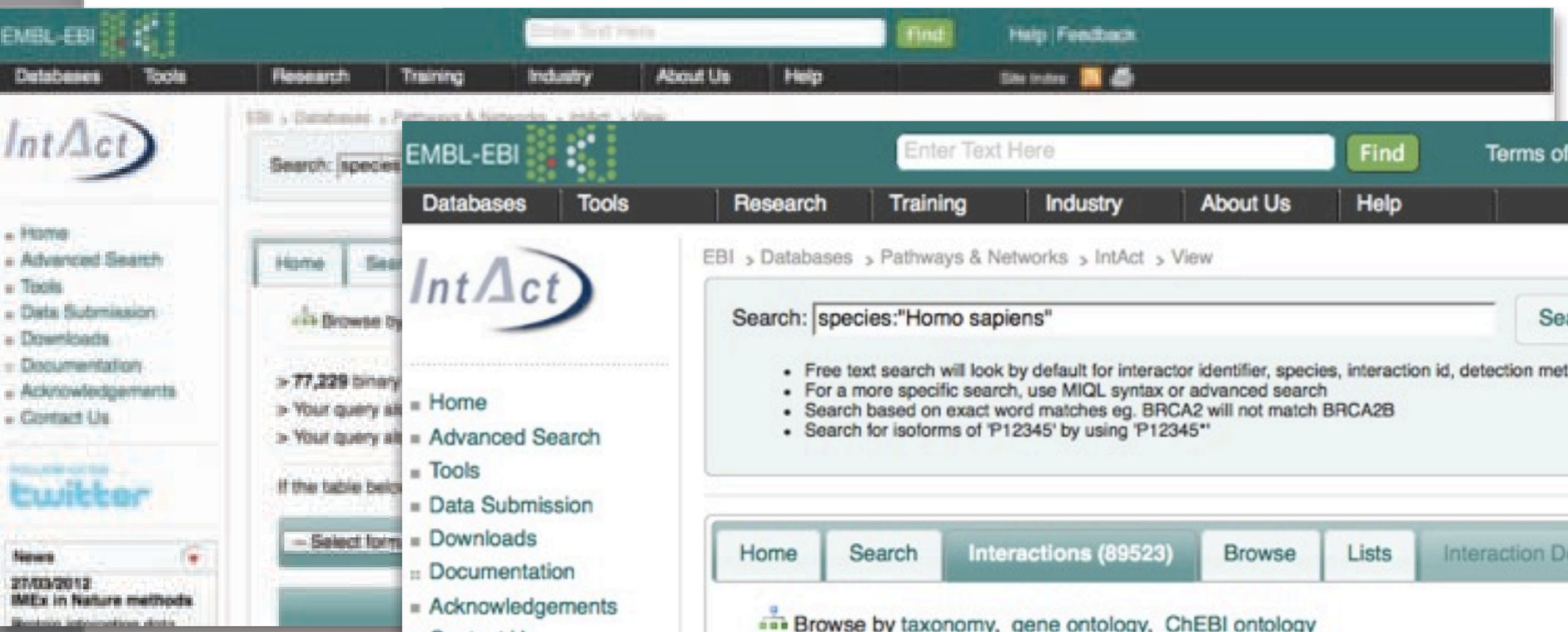


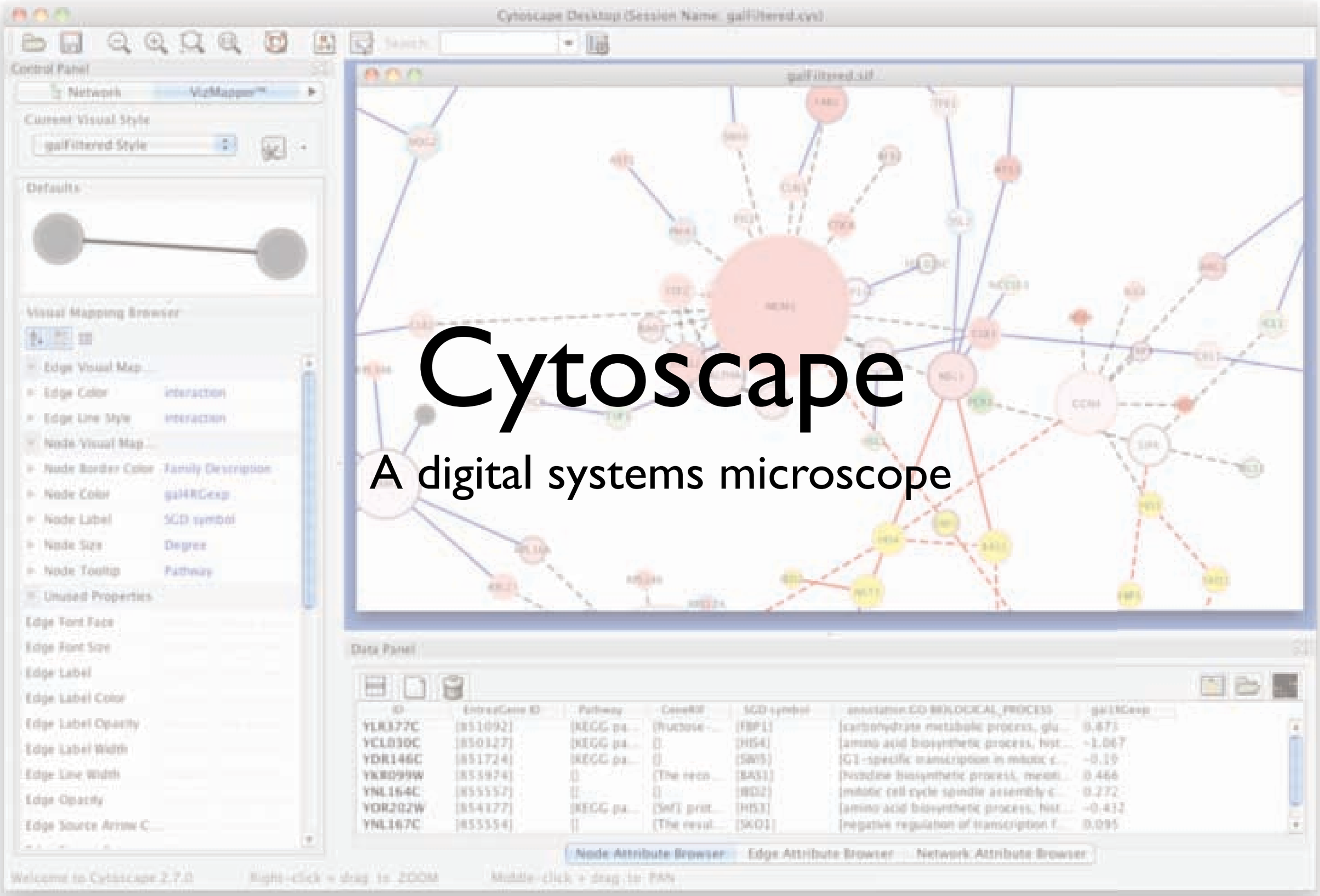
Figure 1. Enhanced protein function prediction via an annotated protein interaction web. A network of protein interactions is diagrammed up to a distance of two steps around YPL105c (marked red), an ORF that encodes a protein of unknown function. Proteins known to play a role in RNA splicing are marked yellow. This local web is connected at multiple points around the periphery to a larger network of 2358 interactions among 1548 proteins.

Protein interactions today



May 21, 2012

Nov 28, 2012

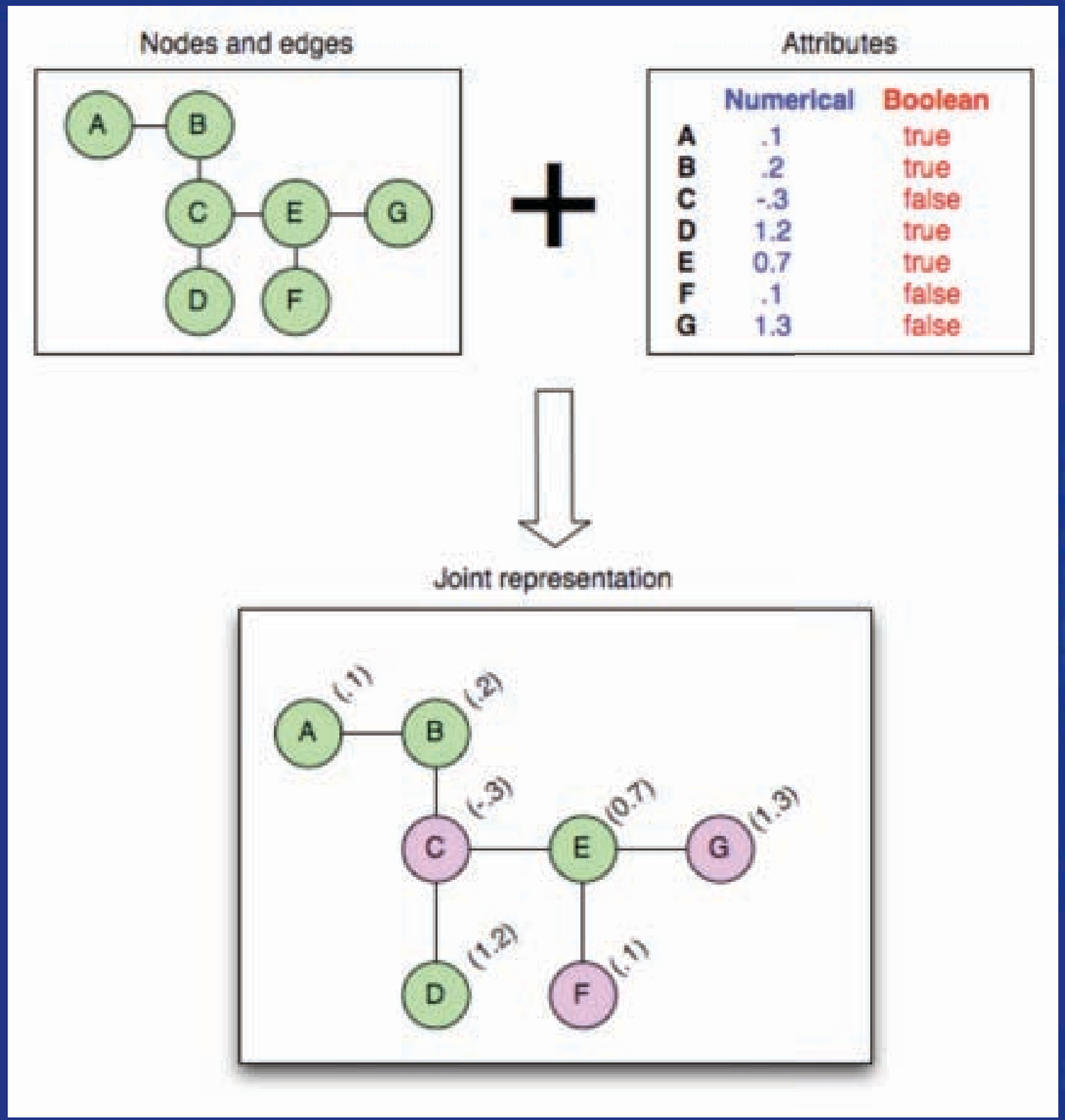


Cytoscape

A digital systems microscope

Cytoscape core functionality
Data model

- Network (Graph)
  - Nodes
  - Edges
- Attributes



Cytoscape file formats

Sample interaction file

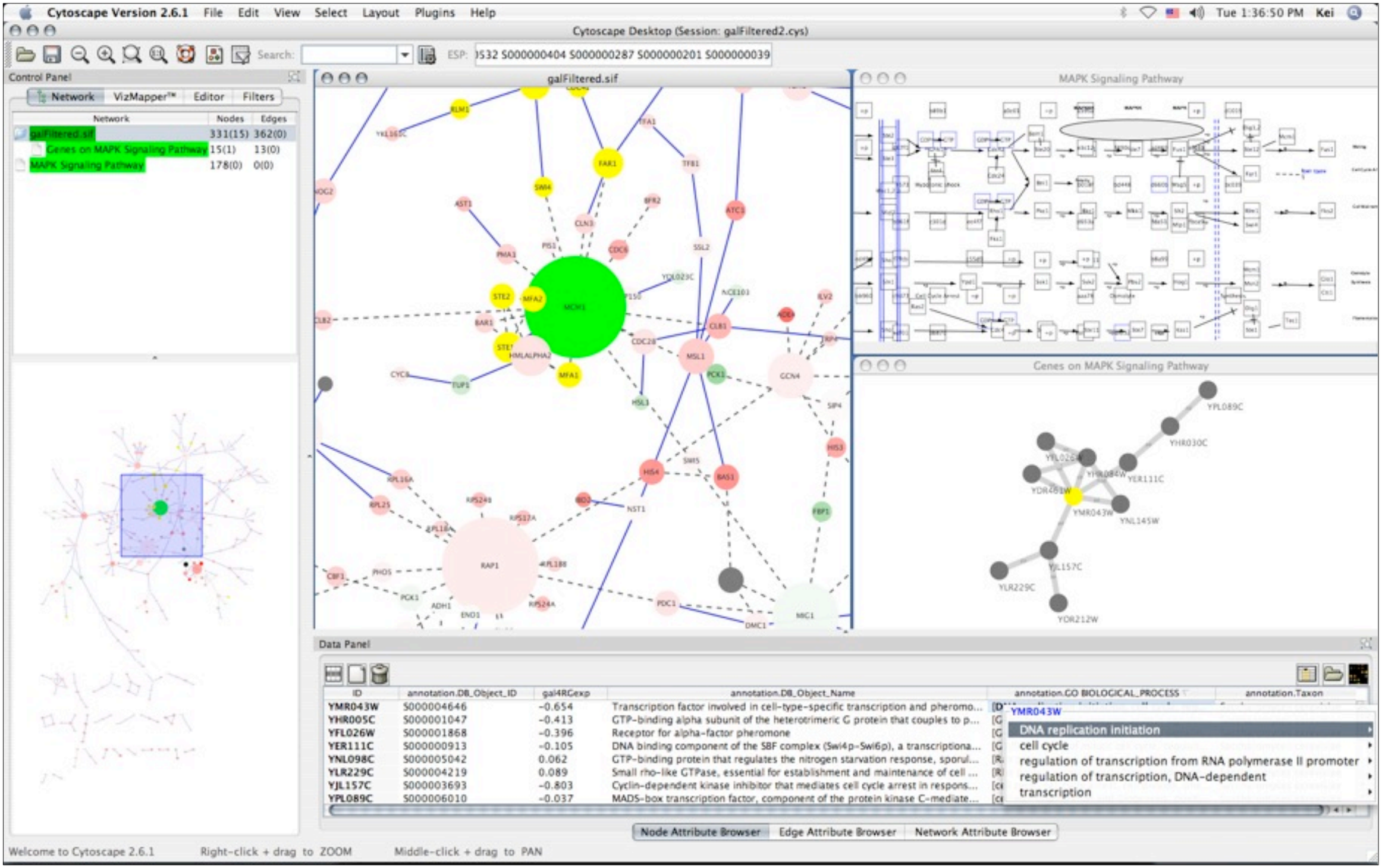
```
YDR216W pd YIL056W
YDR216W pd YKR042W
YDR216W pd YGL096W
YDR216W pd YDR077W

[...]
```

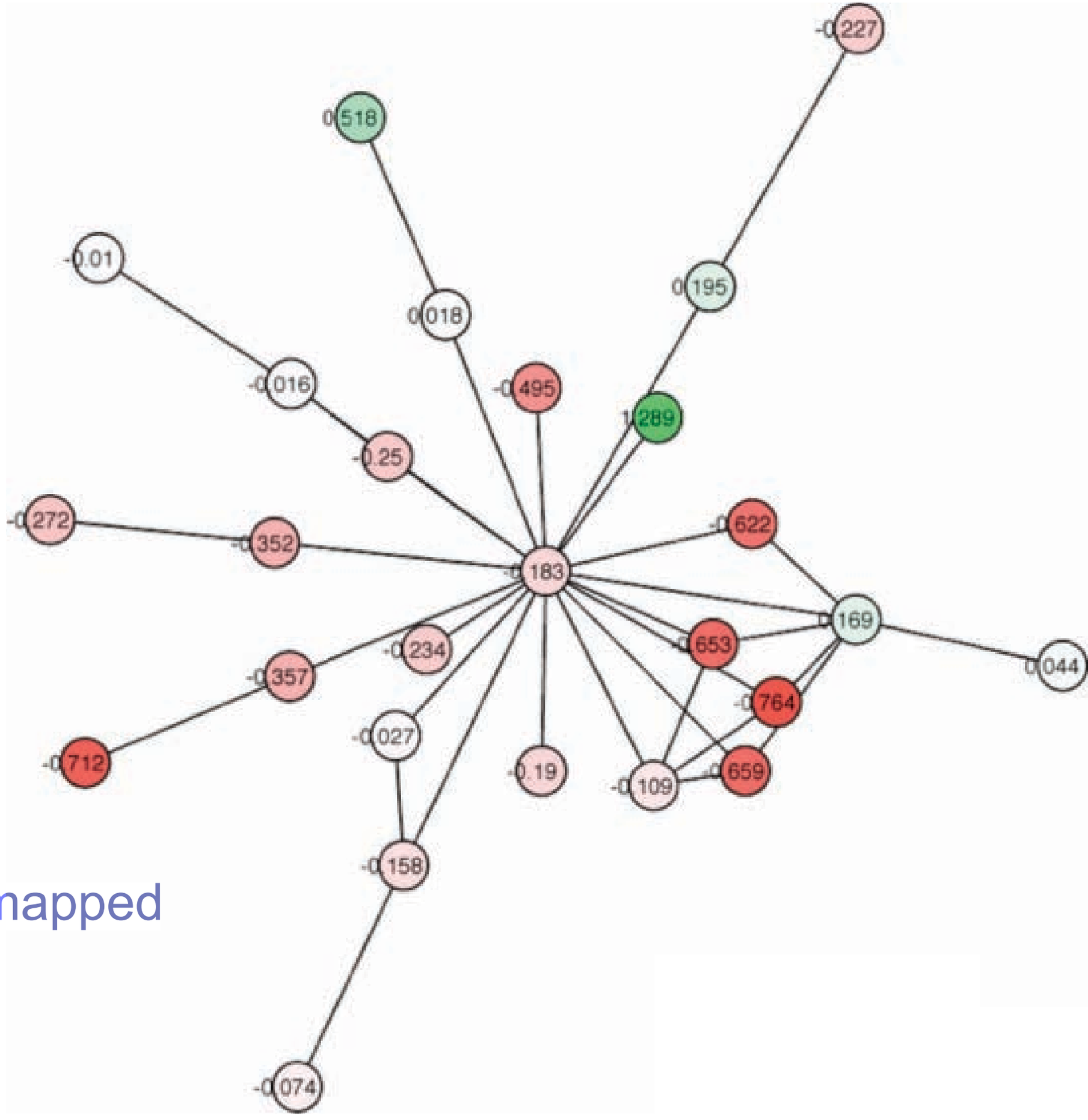
Sample expression file

GENE	DESC	exp0.sig	exp1.sig	exp0.sig	exp1.sig	
GENE0		G0	0.0	0.0	23.2	11.5
GENE1		G1	0.0	0.0	34.6	5.2
GENE2		G2	0.0	0.0	10.0	28.0
GENE3		G3	0.0	0.0	1.64	4.77
[...]						

Cytoscape interface



Visual Styles

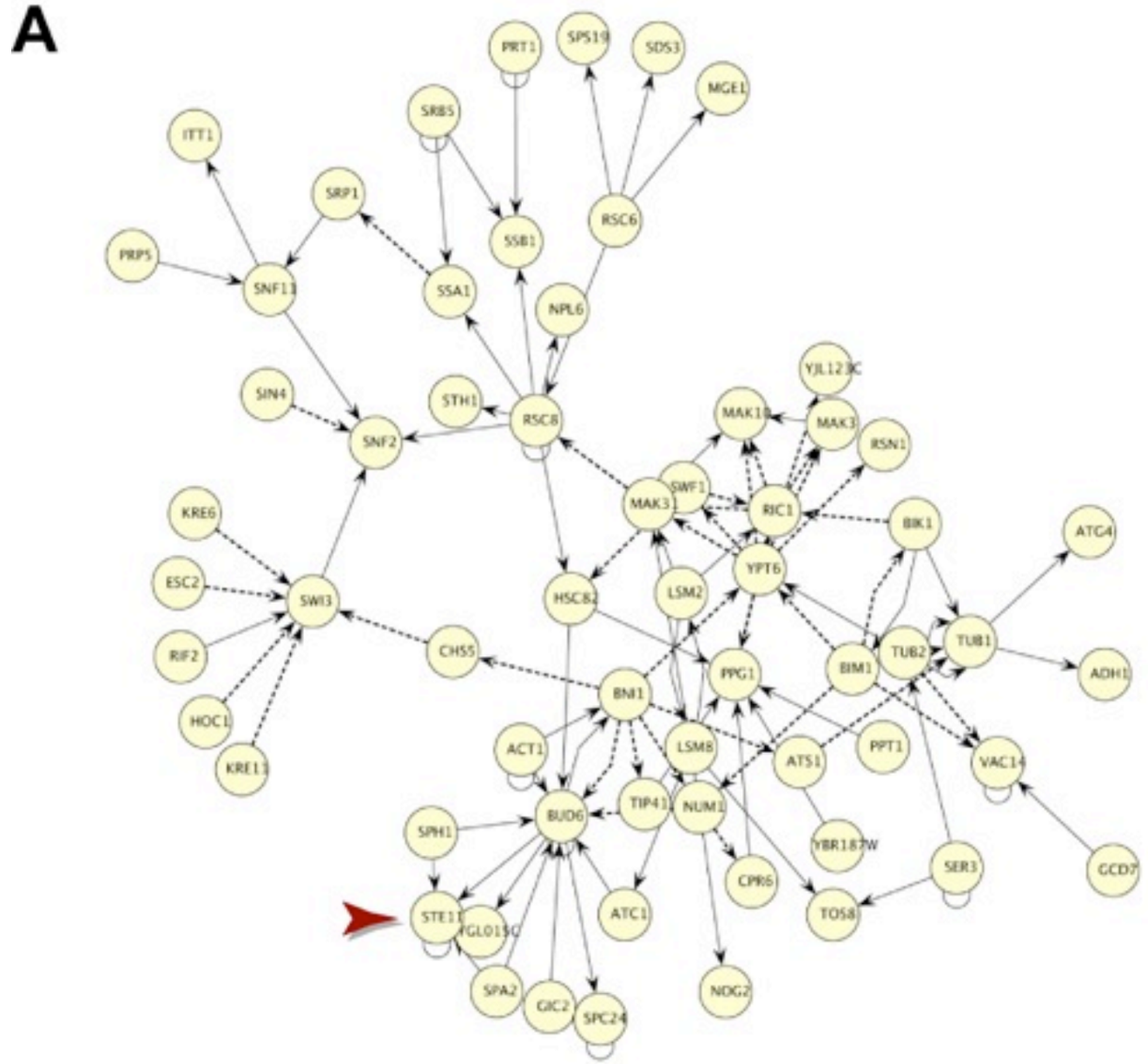


Expression data mapped to node colors

Visual Styles

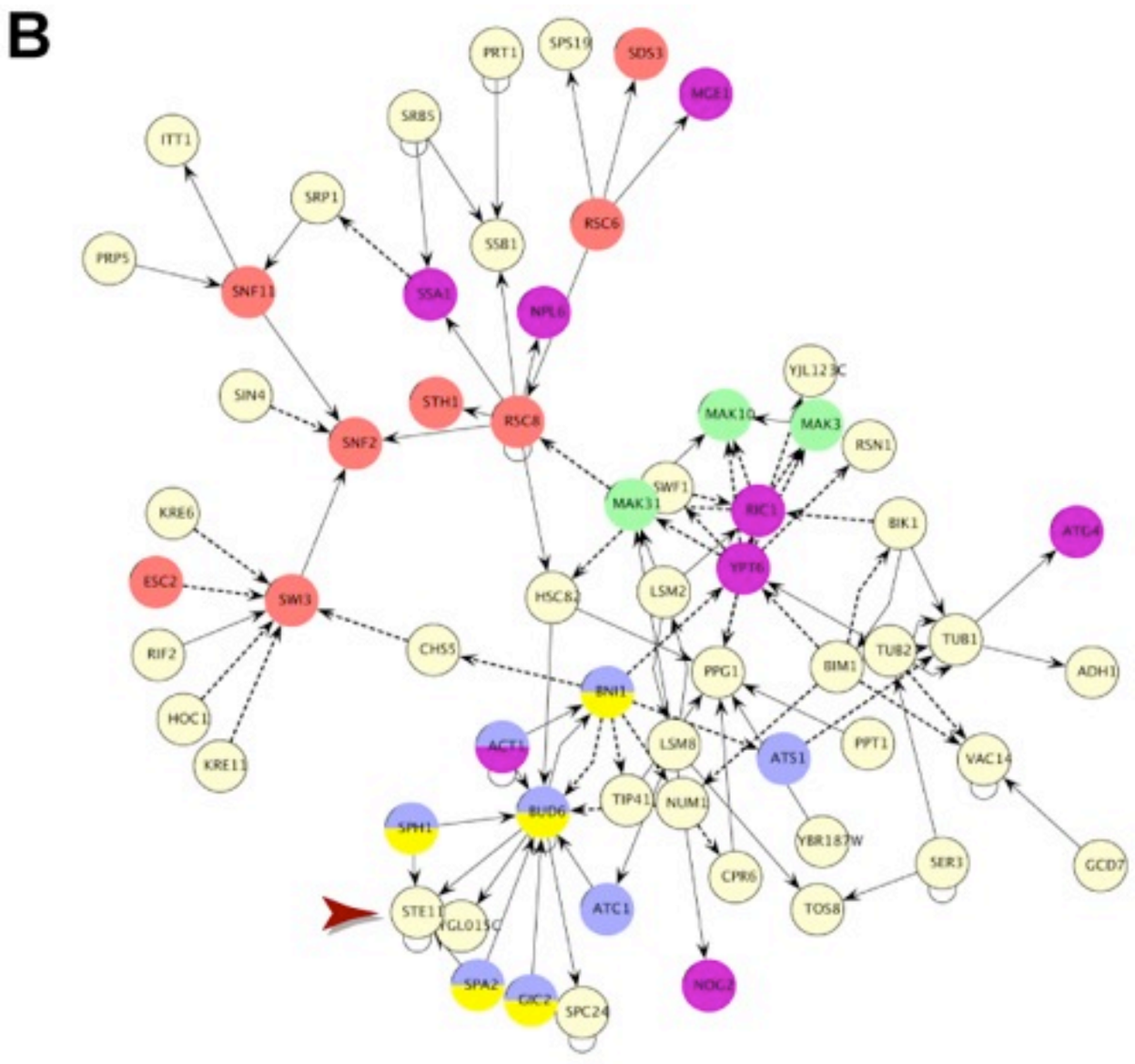
Map expression values to node colors using a continuous mapper

Visual Styles



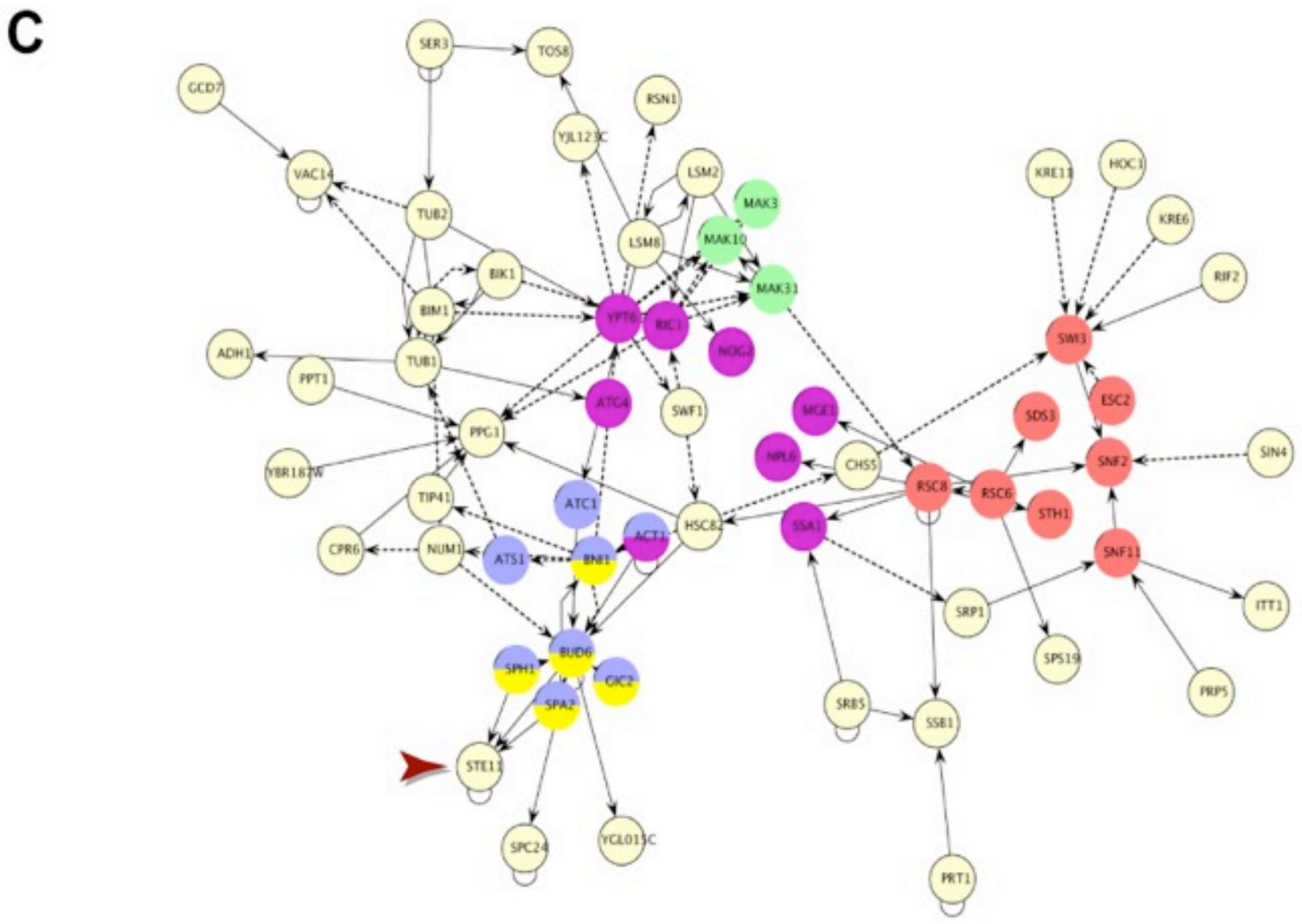
Spring-embedded layout

Color reflecting GO classes



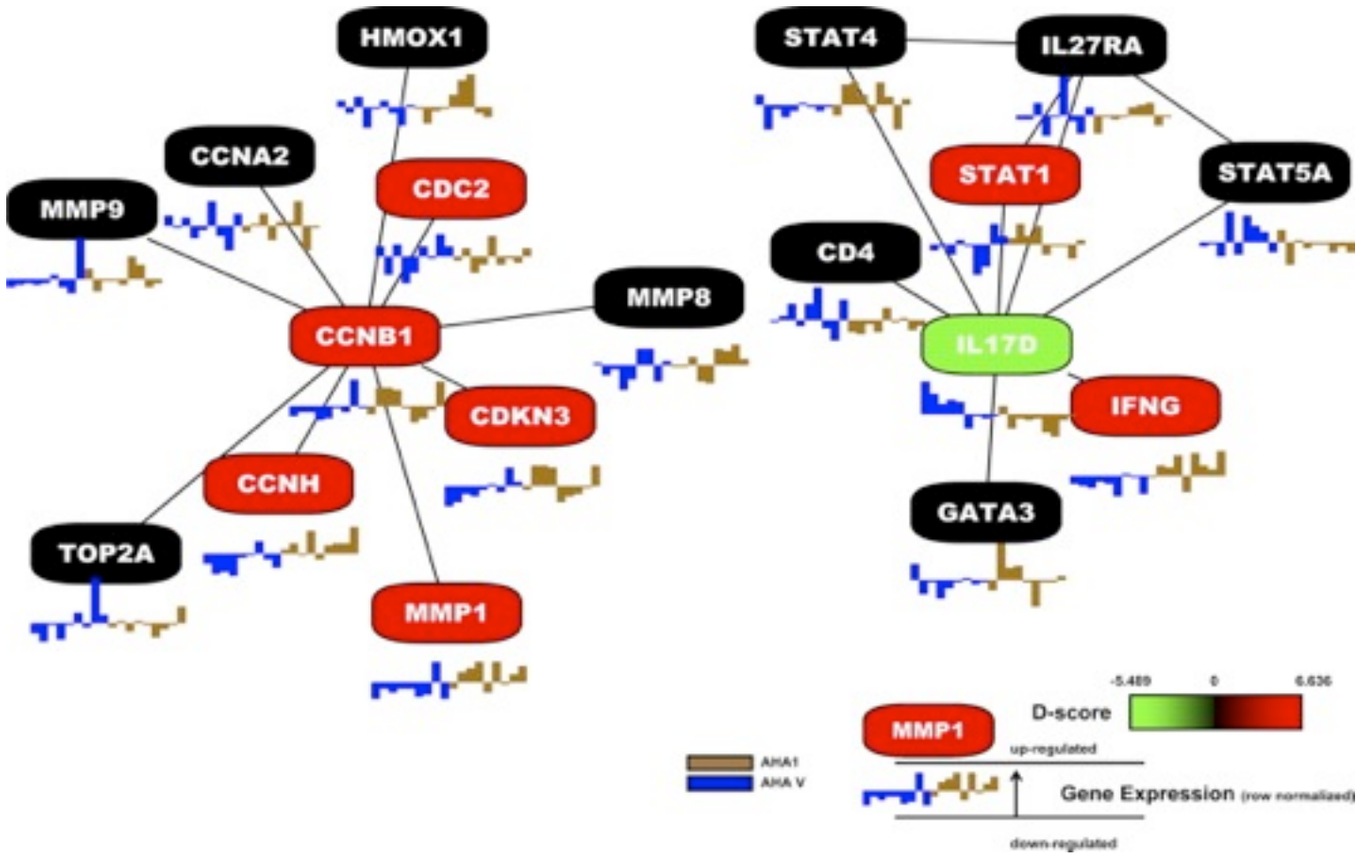
Spring-embedded layout, supplemented with color-coding

GO categories driving layout



GOlorize layout algorithm, informed by the Gene Ontology classes
Garcia et al. Bioinformatics 2007

Visualizing dense information



King et. al., Physiol Genomics. 2005 Sep 21;23(1):103-18.

Cytoscape, a digital microscope for cell biology

- Different qualitative and quantitative information accessible in a visualization and analysis platform
- Organized into interaction networks that represent systems
- Visual exploration goes hand in hand with statistical data exploration and analysis
- Simple, open-standard file formats, links with many databases
- Active community exchanging networks, know-how, and new functionality

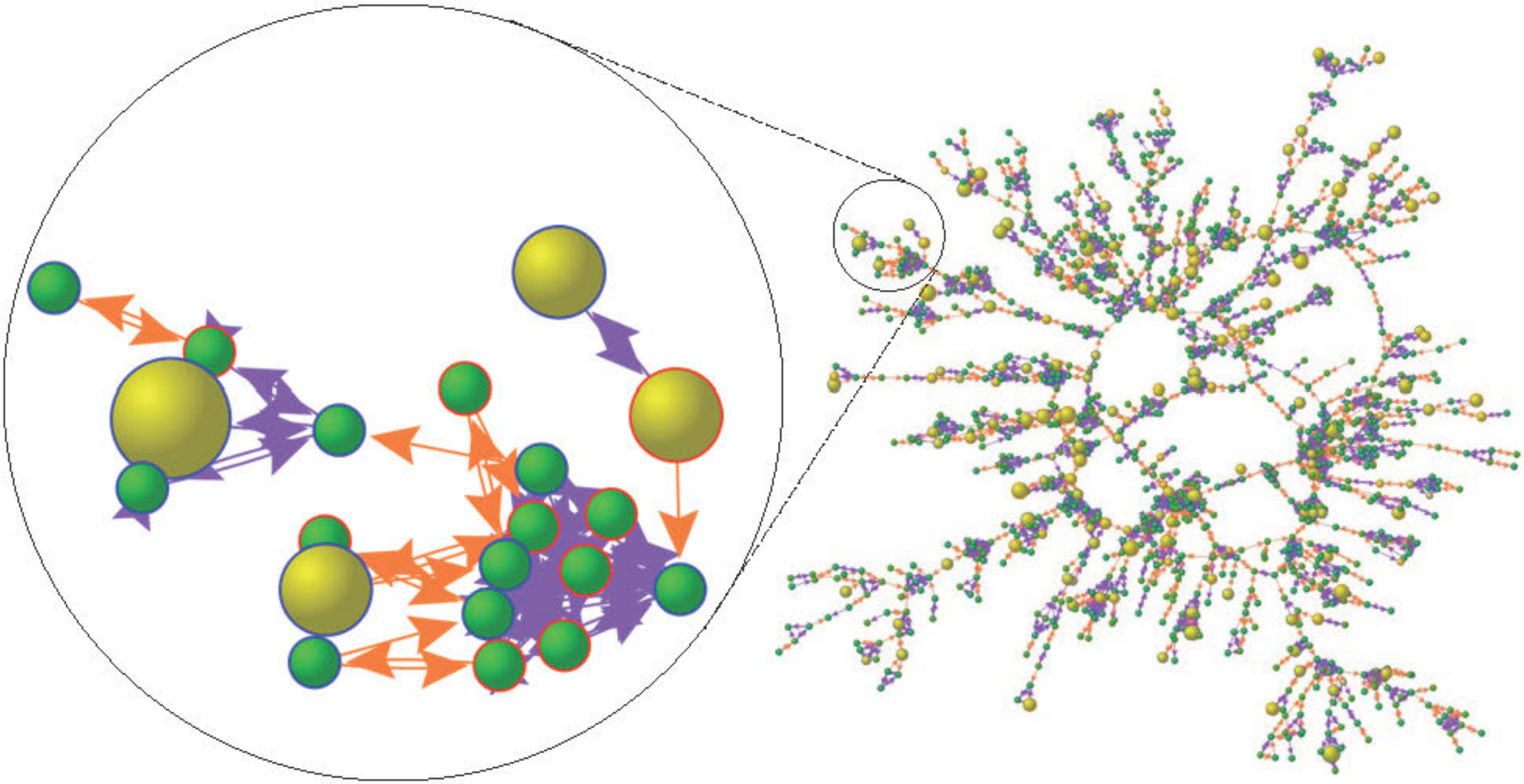
Cytoscape

- Cross-platform (Java)
- Plug-in architecture that allows external developers to easily extend core platform
- Nearly 100 plugins available through our website: http://cytoscape.org
- Downloaded ~2500 times per month.
- Very popular in the Systems Biology community, but also used in other domains

The Collective Dynamics of Smoking in a Large Social Network

Nicholas A. Christakis, M.D., Ph.D., M.P.H., and James H. Fowler, Ph.D.

N ENGL J MED 358;21 WWW.NEJM.ORG MAY 22, 2008



Dissemination

- Cytoscape is released under the LGPL software license - it is ***free*** software available for download from our website.
- The hub of our dissemination efforts is the project Website: http://cytoscape.org

Cytoscape: An Open Source Platform for Complex-Network Analysis and Visualization

http://cytoscape.org/

UW ZL BV SB PM News Personal IP Science 3d Rendering Of Thr... 3d Rendering Of Thr... 3d Rendering Of 3 D... 3d Rendering Of 3 D...

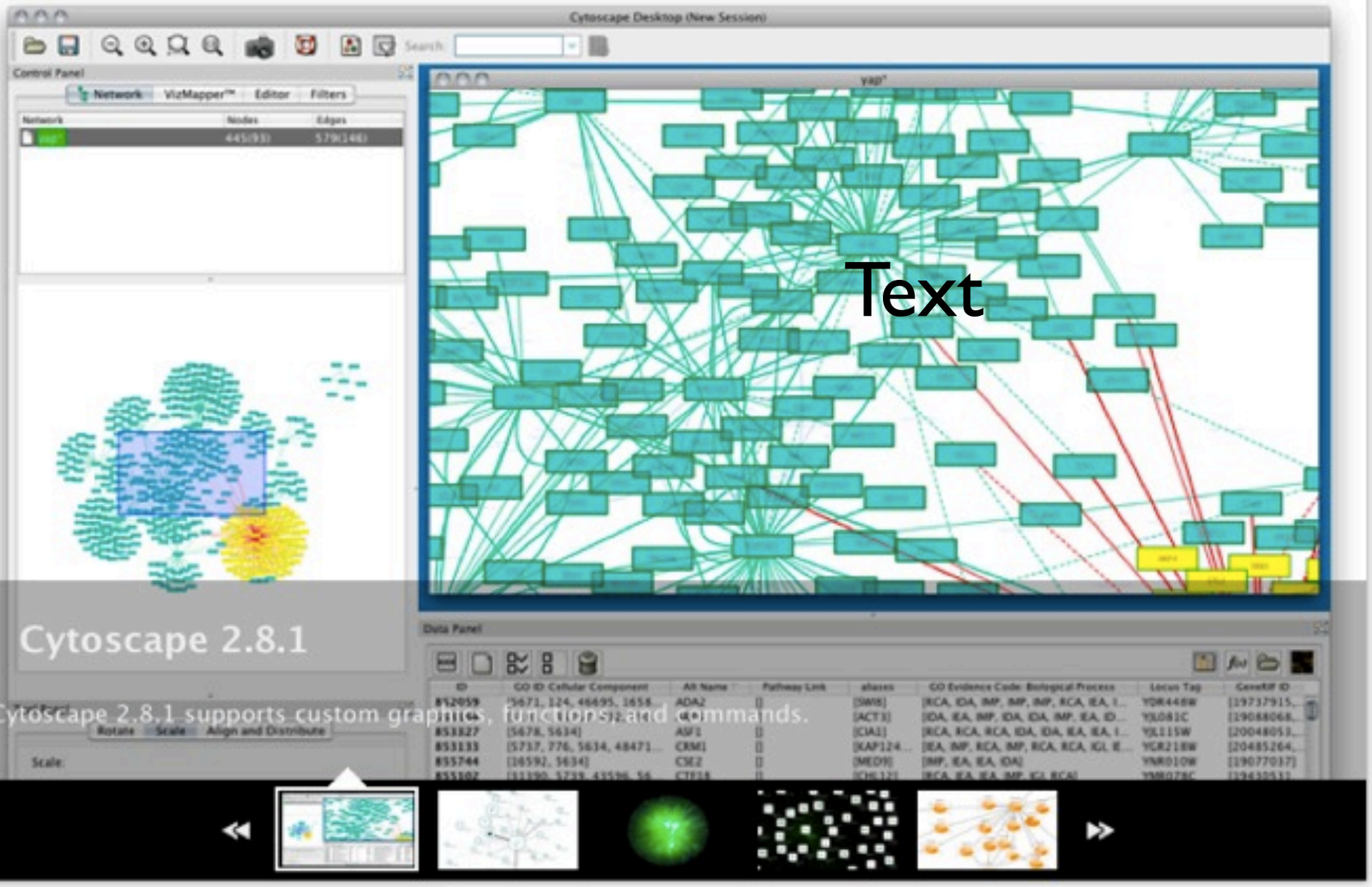
0 results - TinEye | Save the date: Cytoscape retrea... | Cytoscape: An Open Source Pla...

Cytoscape

Cytoscape: An Open Source Platform for Complex-Network Analysis and Visualization

Search... Go

Home Introduction Download Plugins Documentation Community Report a Bug Getting Help



Network Data Integration, Analysis, and Visualization in a Box

Cytoscape is an open source software platform for visualizing complex-networks and integrating these with any type of attribute data. A lot of plugins are available for various kinds of problem domains, including bioinformatics, social network analysis, and semantic web.

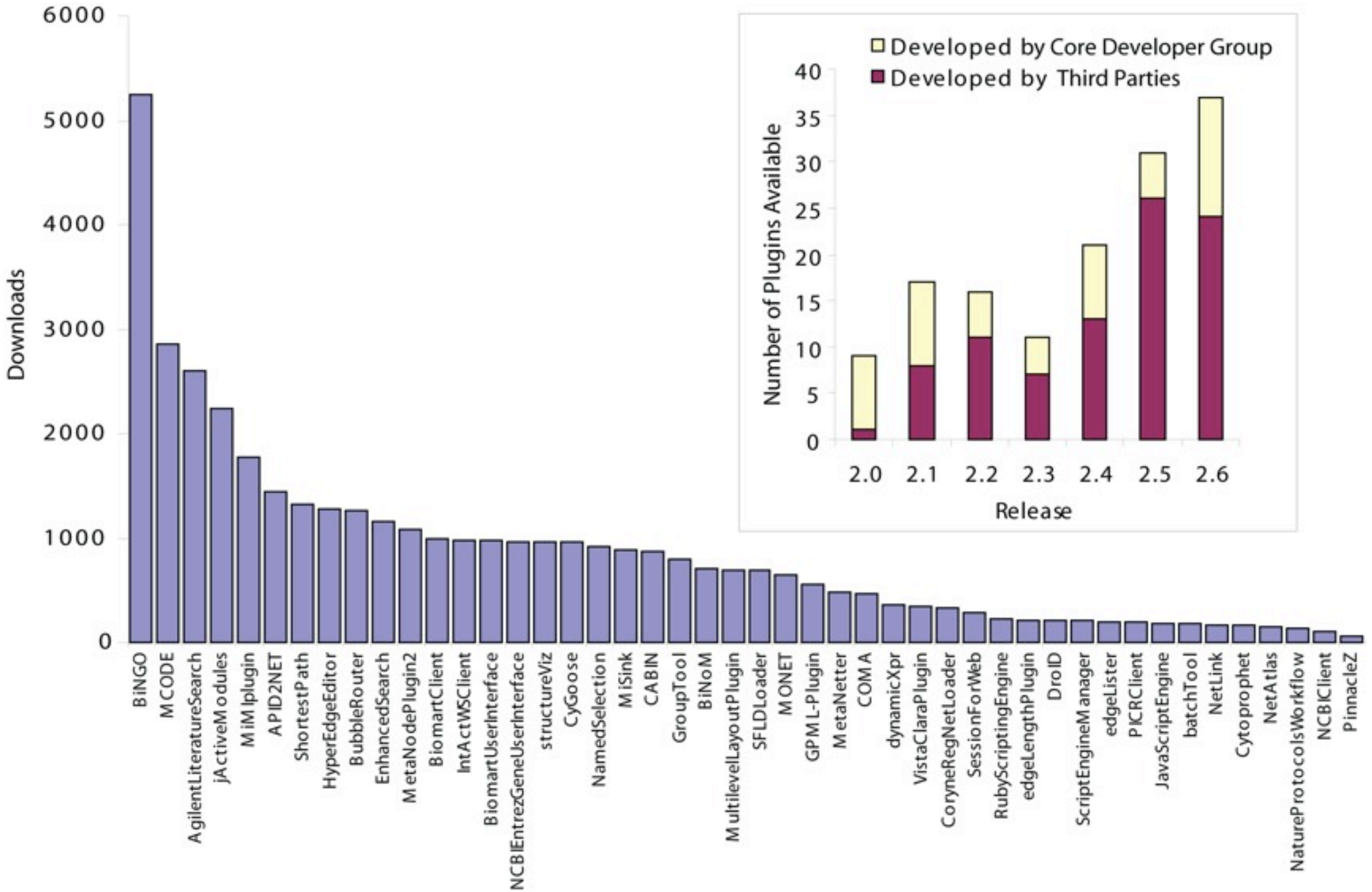
Download Cytoscape Now

Learn More

How to Cite Cytoscape

What Can You Do With Cytoscape?

Cytoscape Plugins

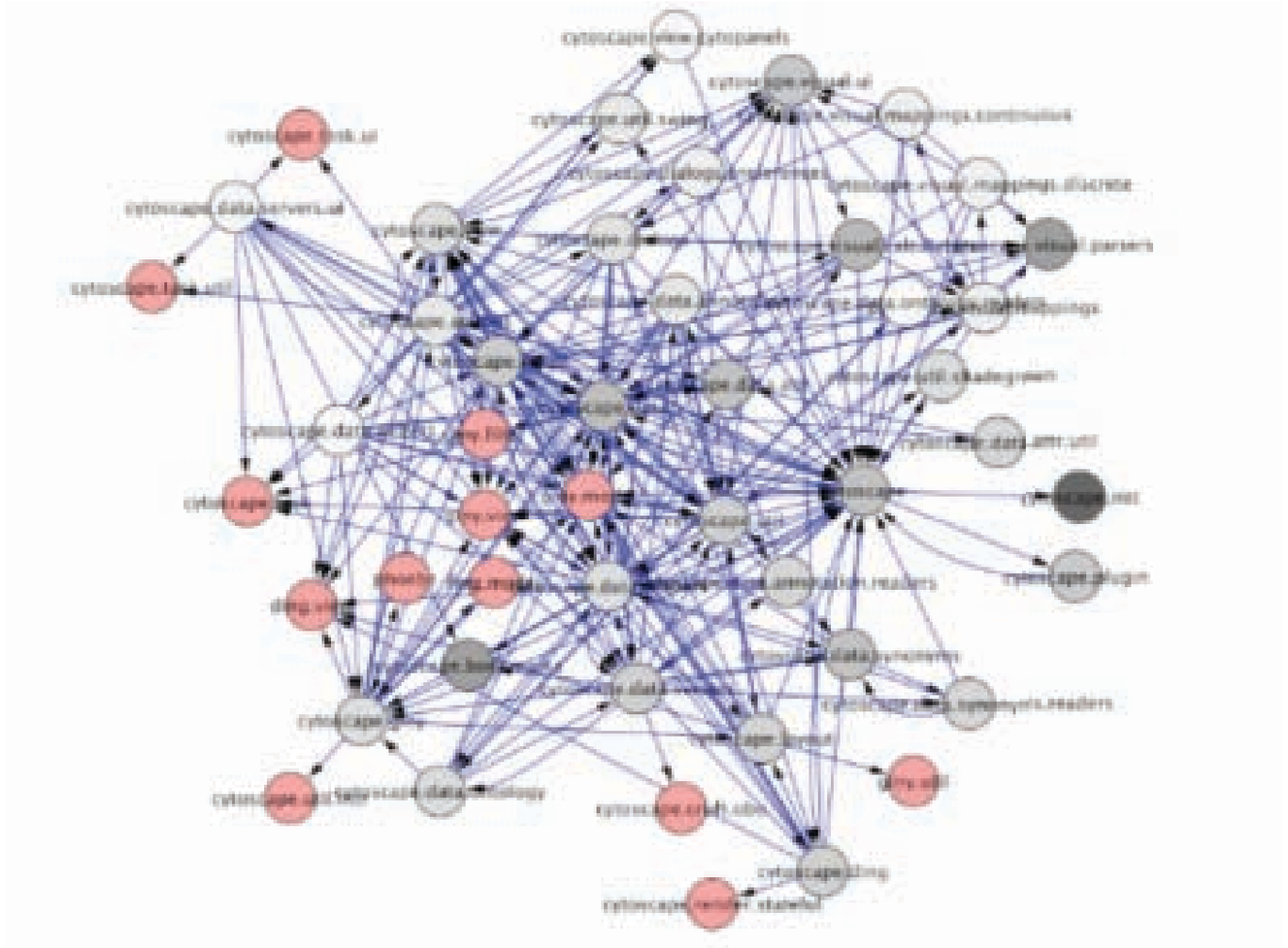


Training and Education

- Numerous tutorials and video lectures available on http://cytoscape.org
- Annual Symposium and Developer's retreat.
  - Held each year.
  - Features invited talks, a plugin expo, tutorials, demonstrations, and user feedback forums.
- Participation in the Google Summer of Code.
- Taught in graduate, undergraduate and other classes (e.g. UCSD, UCSF, Lund University, DTU, ISB, ...).

Cytoscape status

- The Cytoscape codebase is very feature-rich, but is becoming increasingly ***difficult to maintain***, and more importantly ***difficult to extend***.



What's next?

Modular Network Biology Toolkit

(aka Cytoscape 3.0)

What does this mean?

- A set of independent Java modules (jar files).
- A well defined and principled API - culmination of our team's collective experience with Cytoscape.
- Use of OSGi to support and enforce modularity.
- Use of Spring-DM to abstract away the complexities of OSGi.
- Use of Maven to facilitate distribution and integration of modules.
- Follow accepted best programming practices (information hiding, code to interfaces, dependency injection, extensive unit testing, scrum, code quality metrics, semantic versioning, thread safe, ...).

Overall Goals

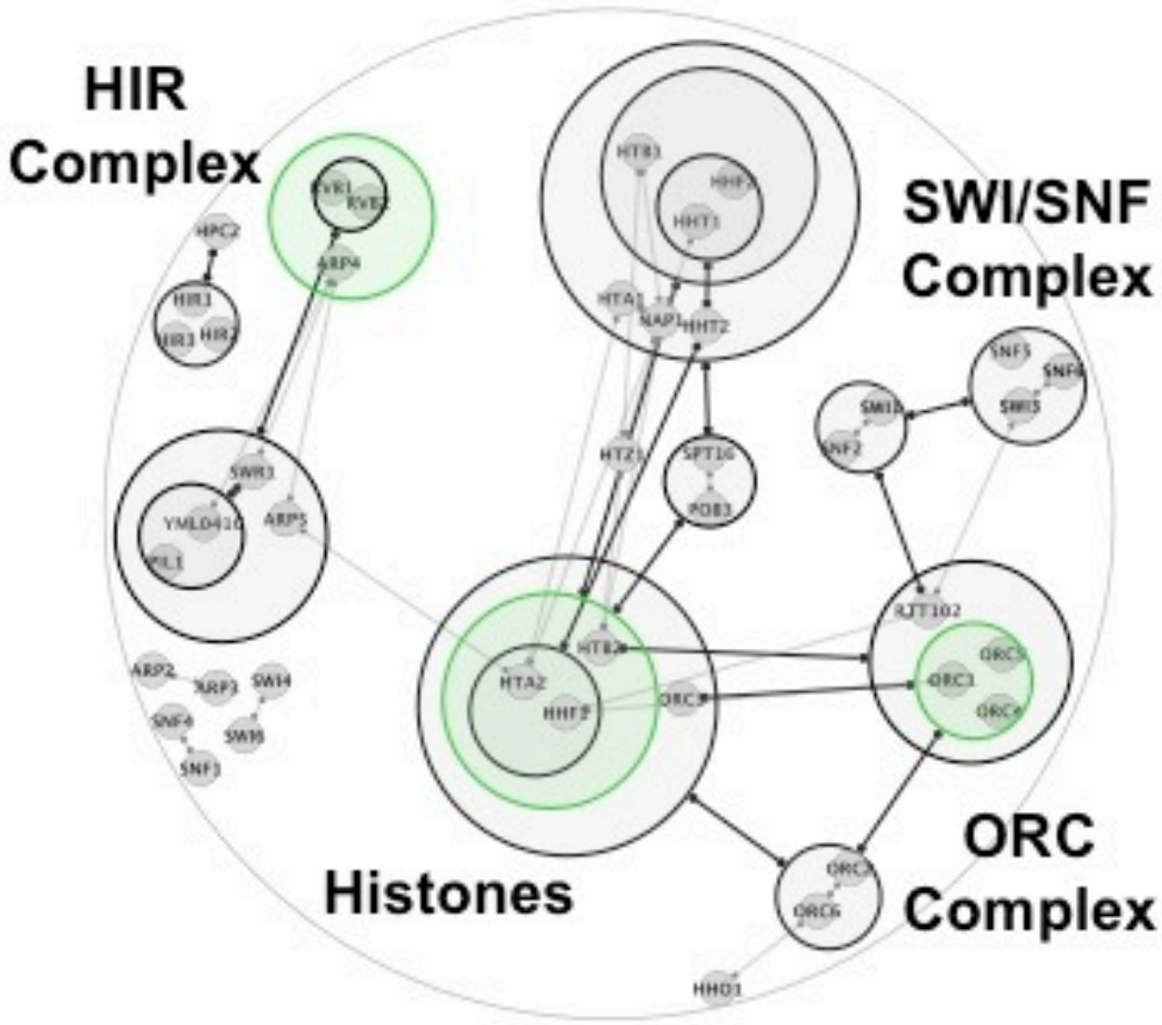
Make Cytoscape...

- Easier to ***use***
  - Simple programming model = more consistent user interface.
- Easier to ***understand***
  - Well defined APIs, well defined dependencies, Maven archetypes.
- Easier to ***maintain and extend***
  - Clear APIs, separate API and implementation, semantic versioning, well understood dependencies.

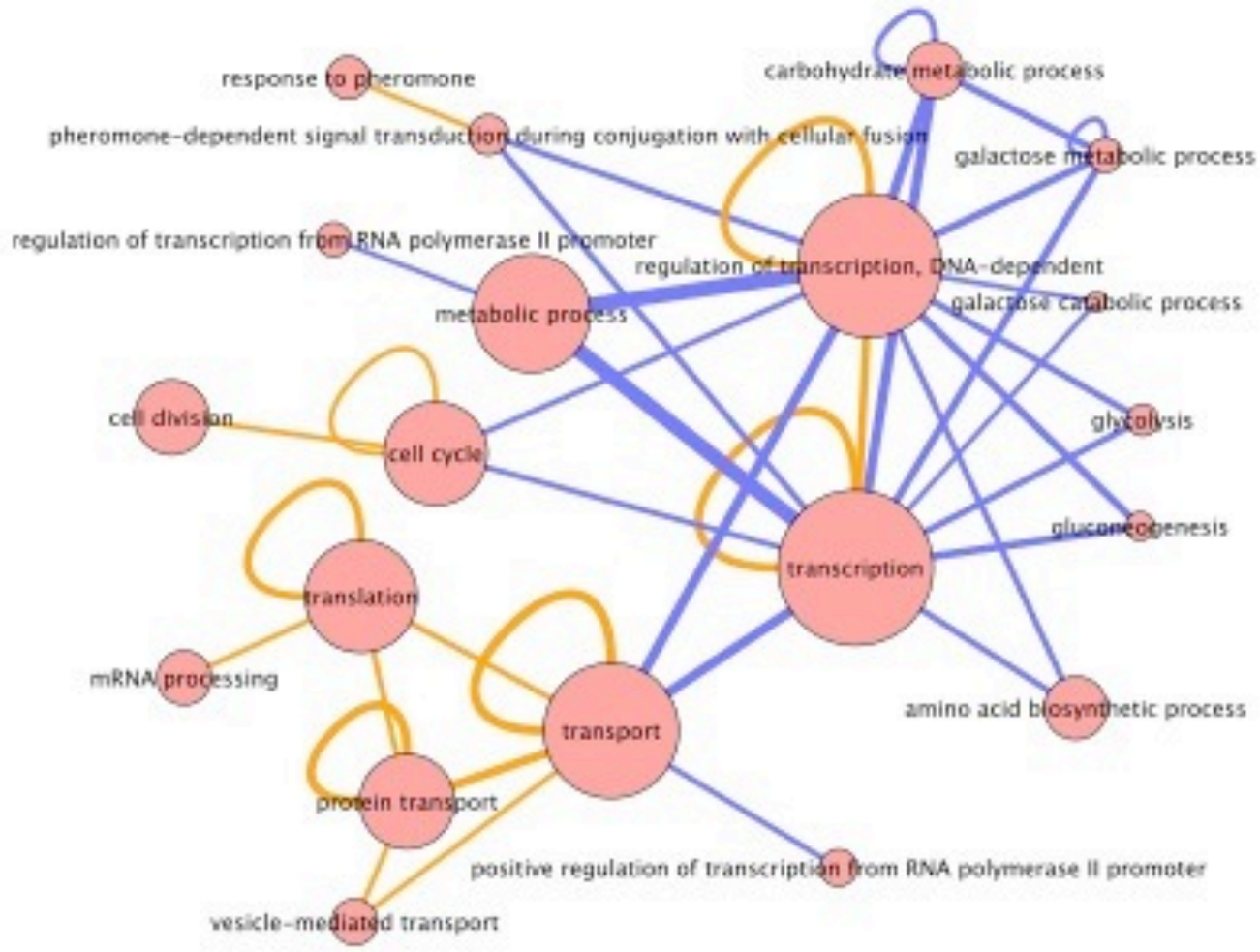
Capabilities

- Do everything the current desktop version of Cytoscape does.
- Run in headless mode to support batch operations.
- Run in daemon mode to support backend web services.
- Take advantage of multithreaded and/or clustered environments.
- Many new features (scripting in different languages, 3D rendering, custom graphics, ...).

1A. Modular Layouts and Views

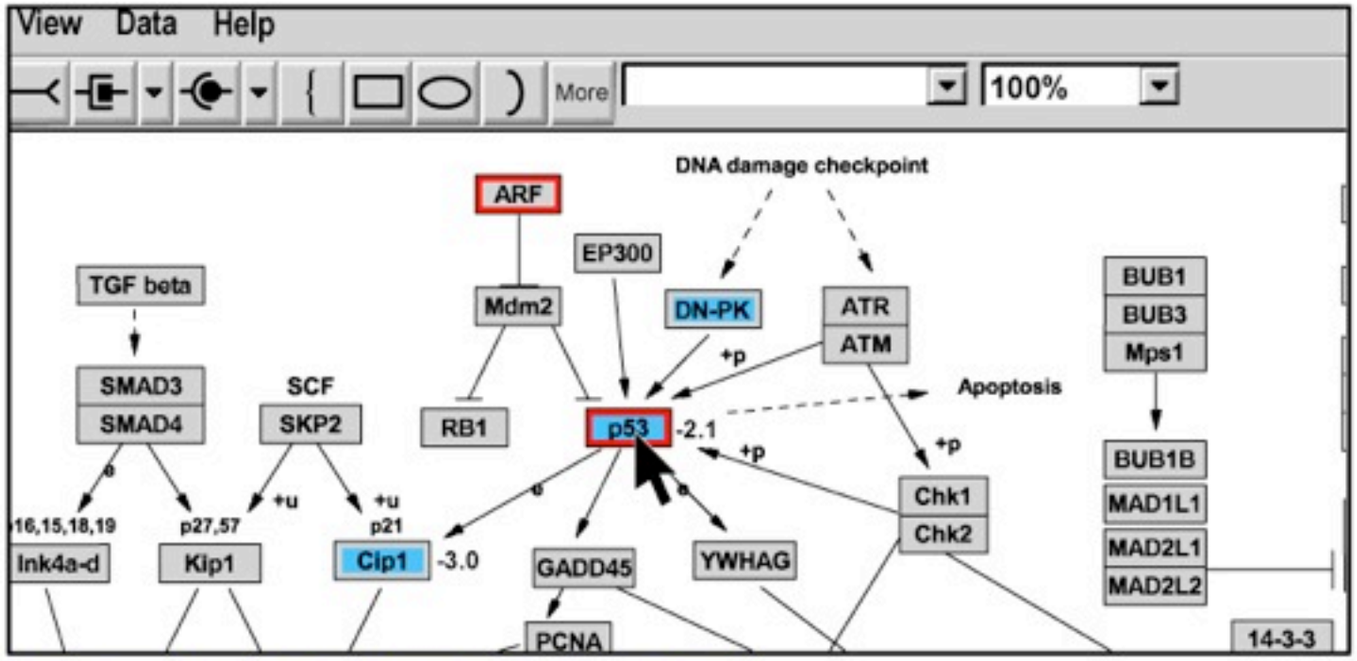


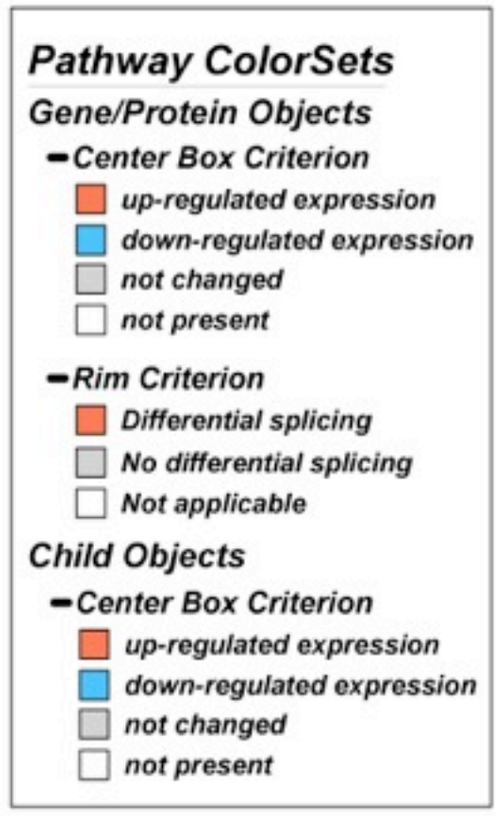
Hierarchical network view



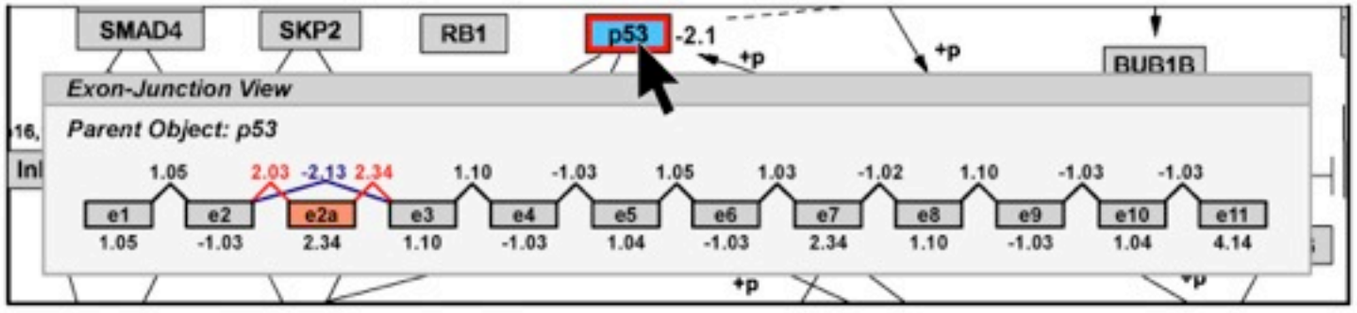
Thematic map based on node attributes

2A. Semantic Zooming:
from genes to exons

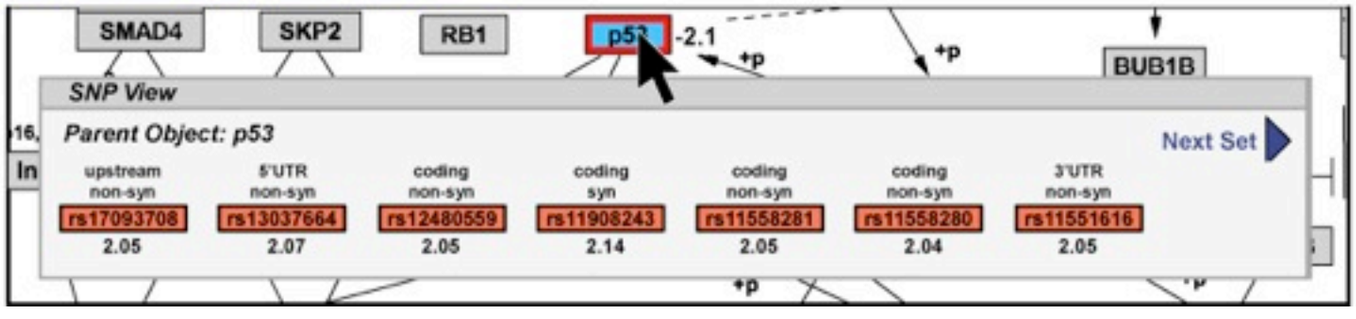


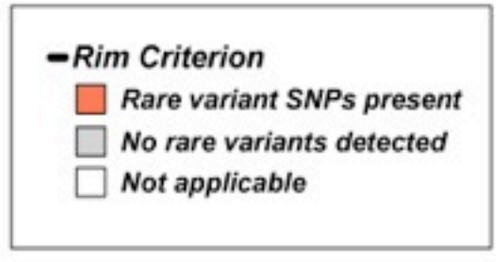


Exon and splice junction analysis

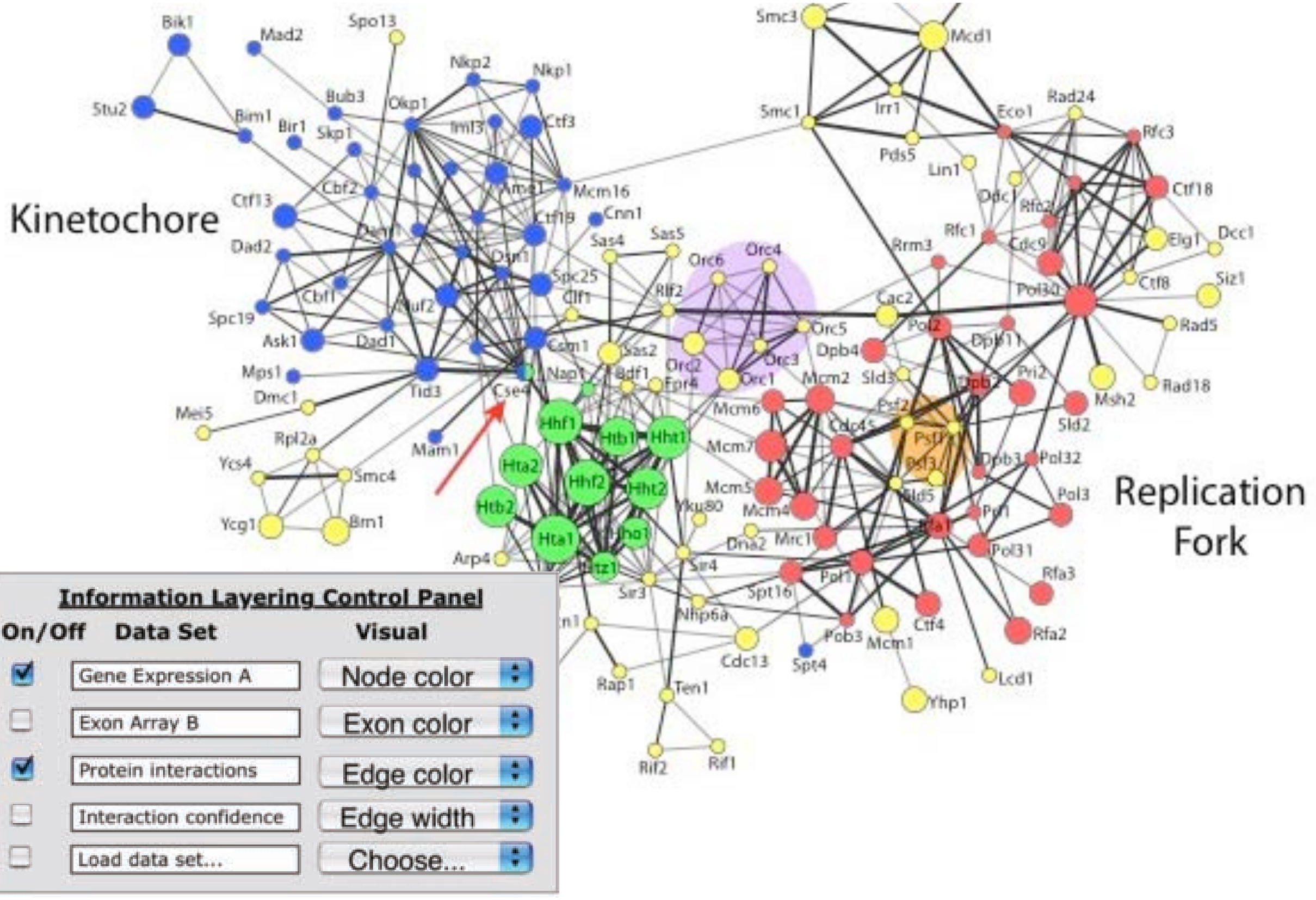


Polymorphism analysis





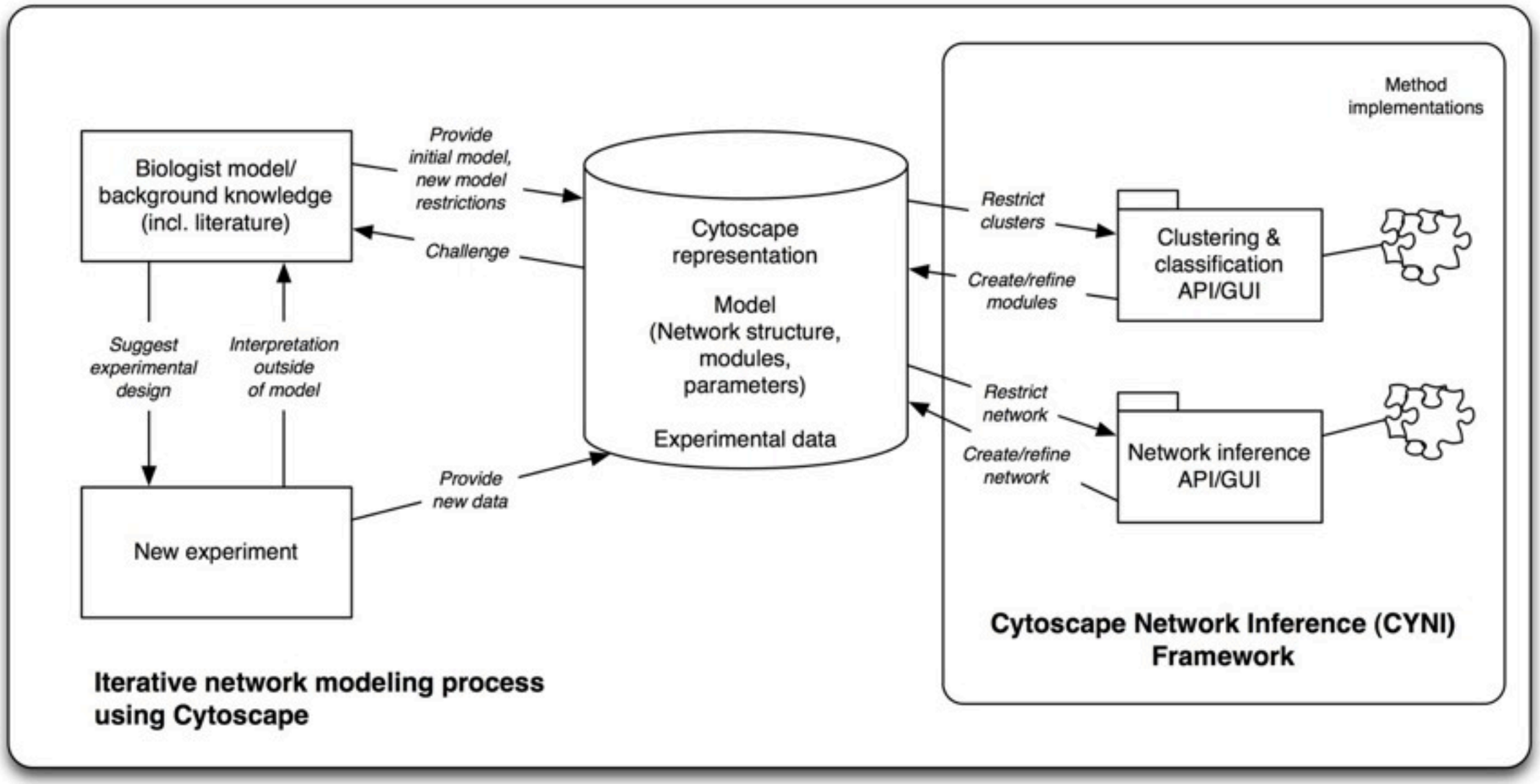
5 (Vt w p t tf wz a



Information Layering Control Panel

On/Off	Data Set	Visual
☑	Gene Expression A	Node color
☐	Exon Array B	Exon color
☑	Protein interactions	Edge color
☐	Interaction confidence	Edge width
☐	Load data set...	Choose...

Cytoscape Network Inference (CYNI)

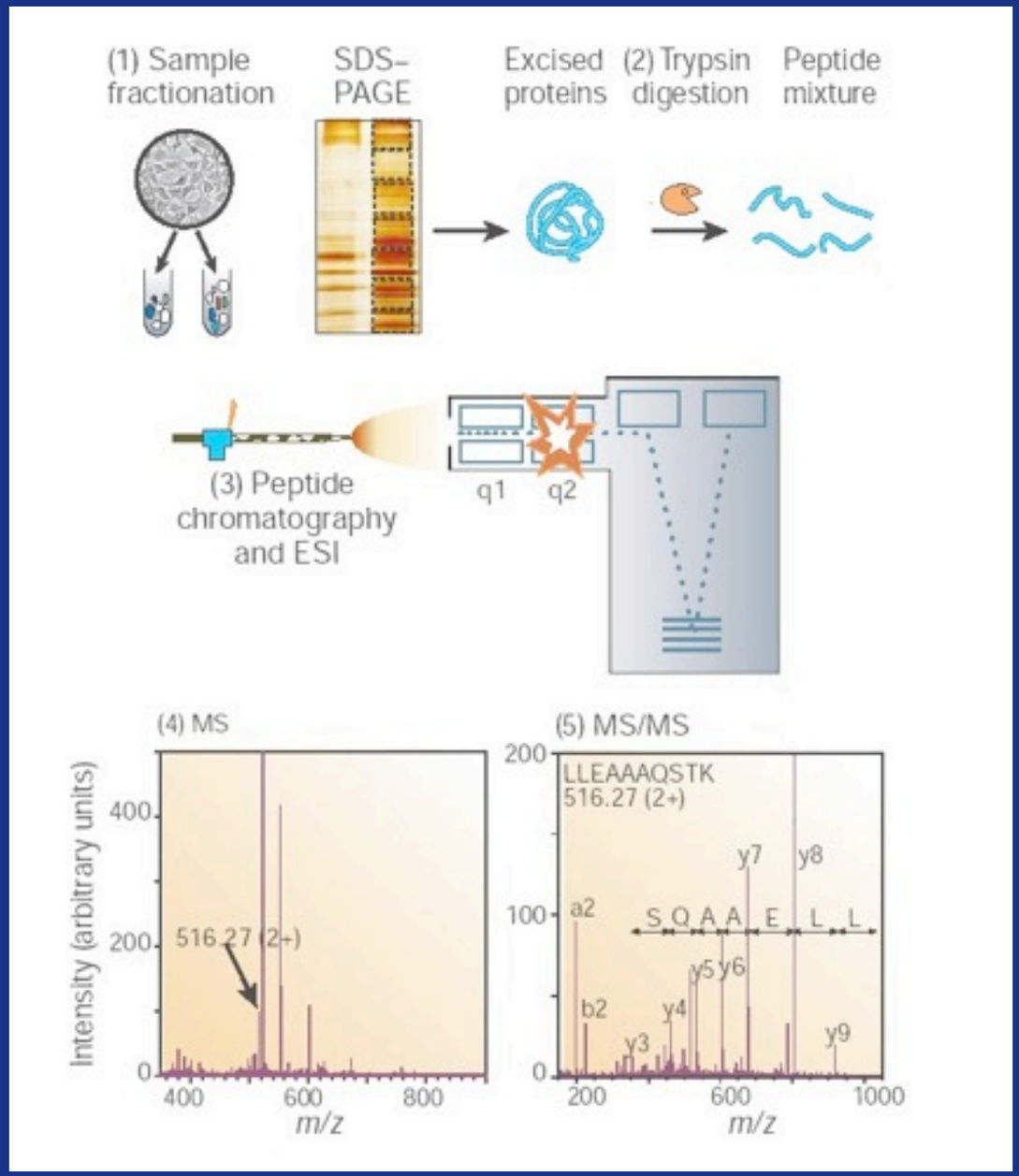


Questions we'd like biologists to ask

- What is the best network explaining the data?
- Which *parts* of this network are well-supported?
- Is there a well-supported subnetwork?
- Which experiment could be done to better distinguish different possible models?
- Given a model, which parts are consistent/inconsistent with the data?
- Which interactions could be added (removed) to make the data compatible with the model?

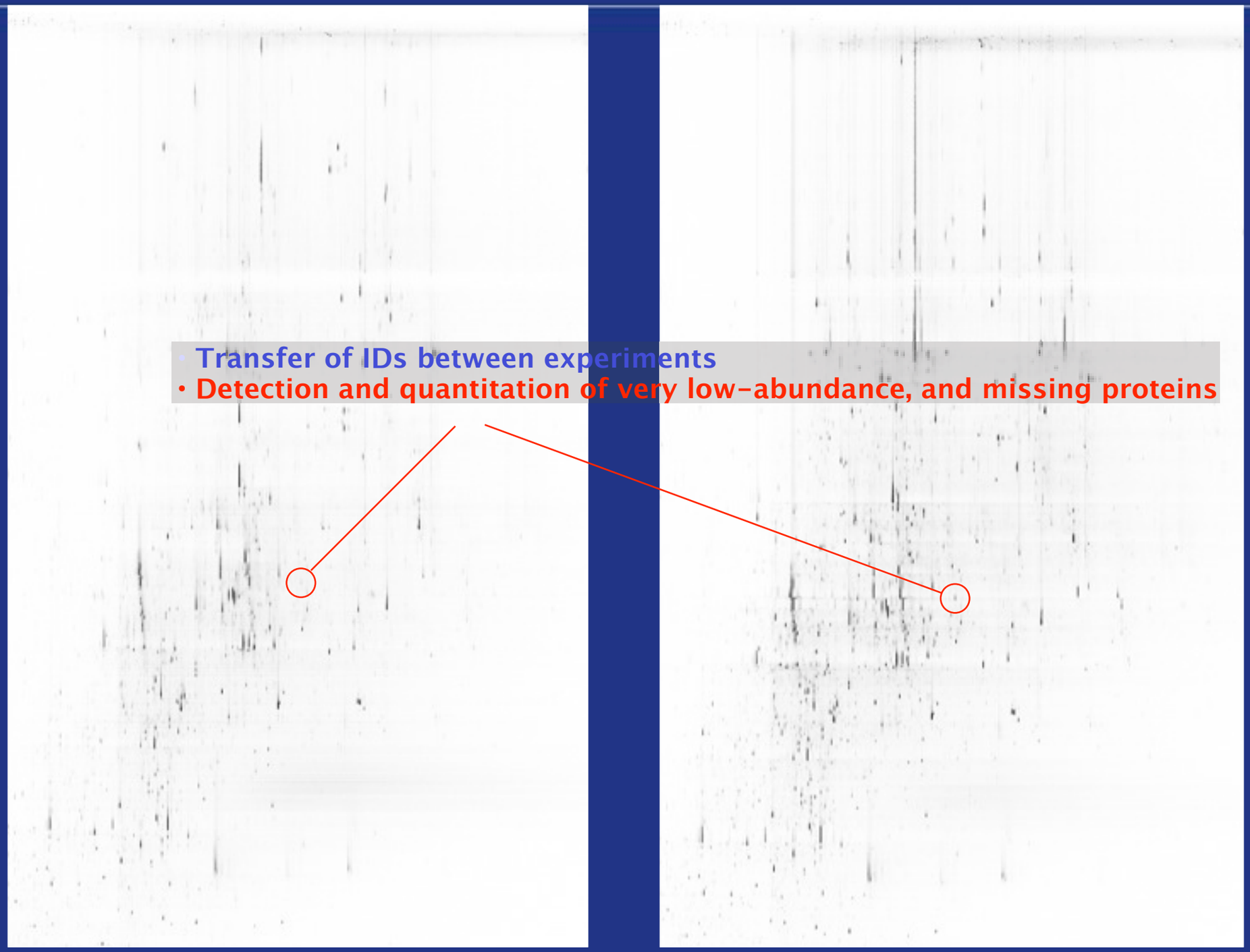
Visualization example in proteomic data analysis

LC-MS/MS is throughput-limited



Aebersold and Mann (2003)

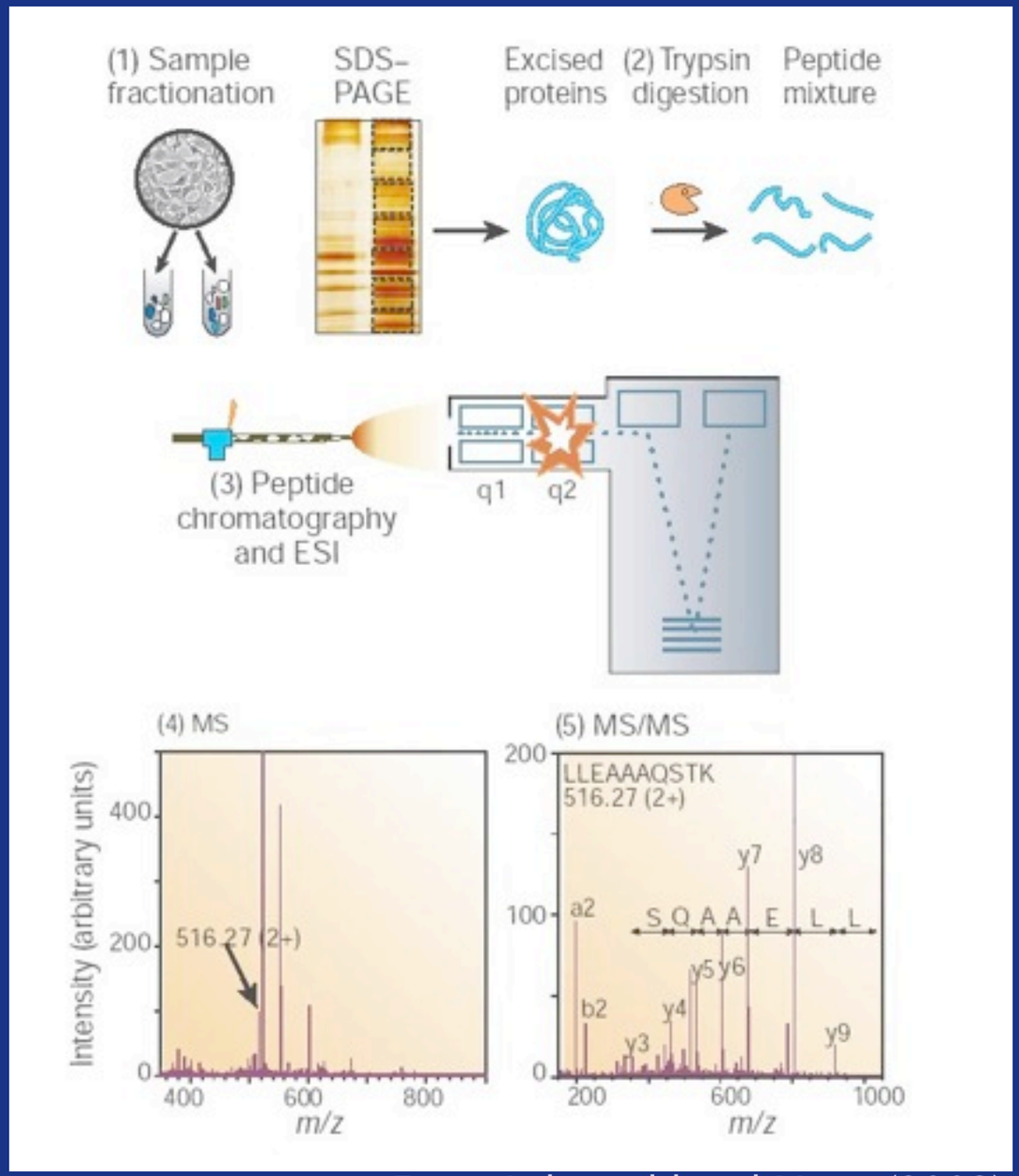
Label-free proteomic analysis



Experiment 1

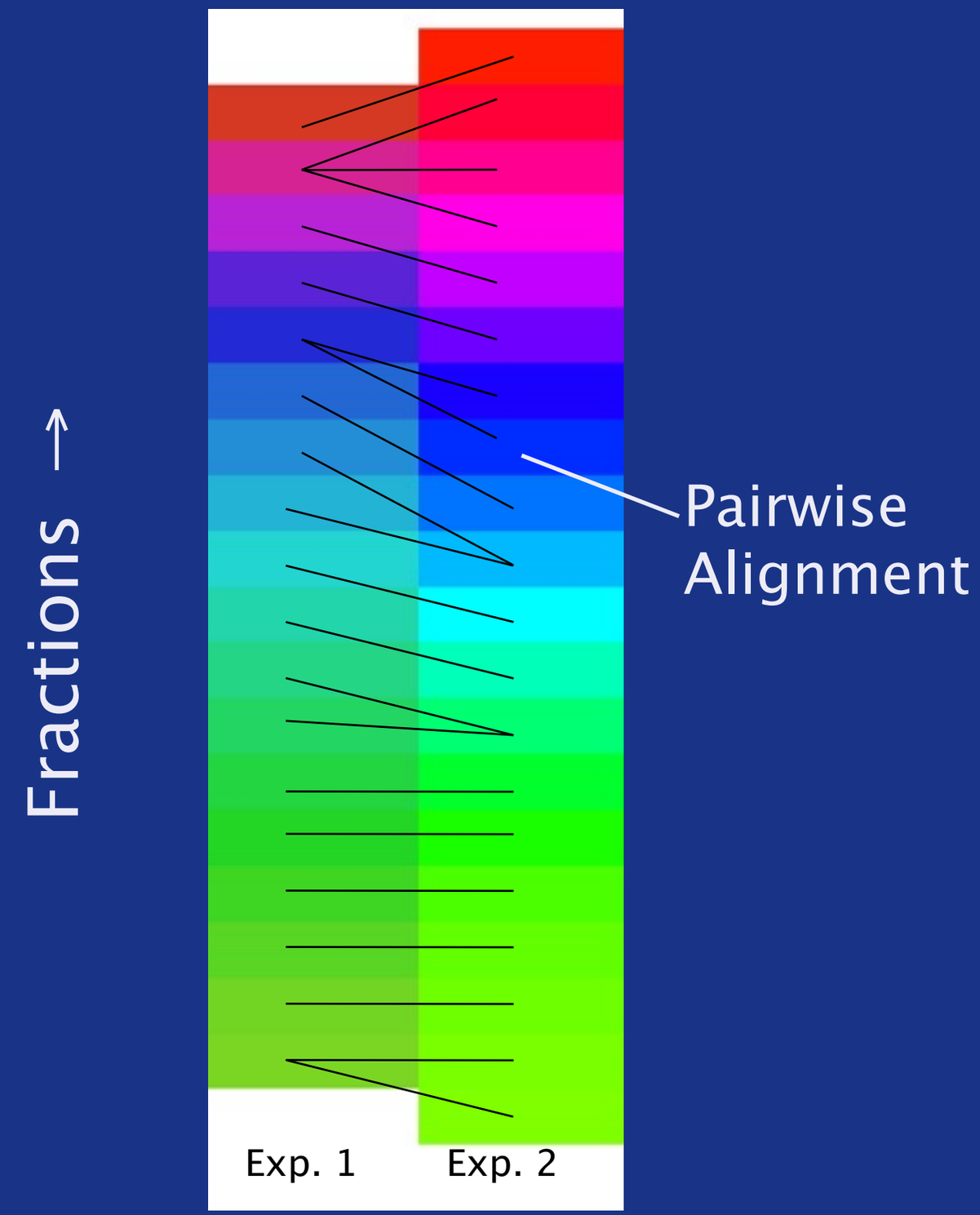
Experiment 2

LC-MS reproducibility



From Aebersold and Mann (2003)

Implementing the comparative approach: Sequence alignment



Sequence alignment

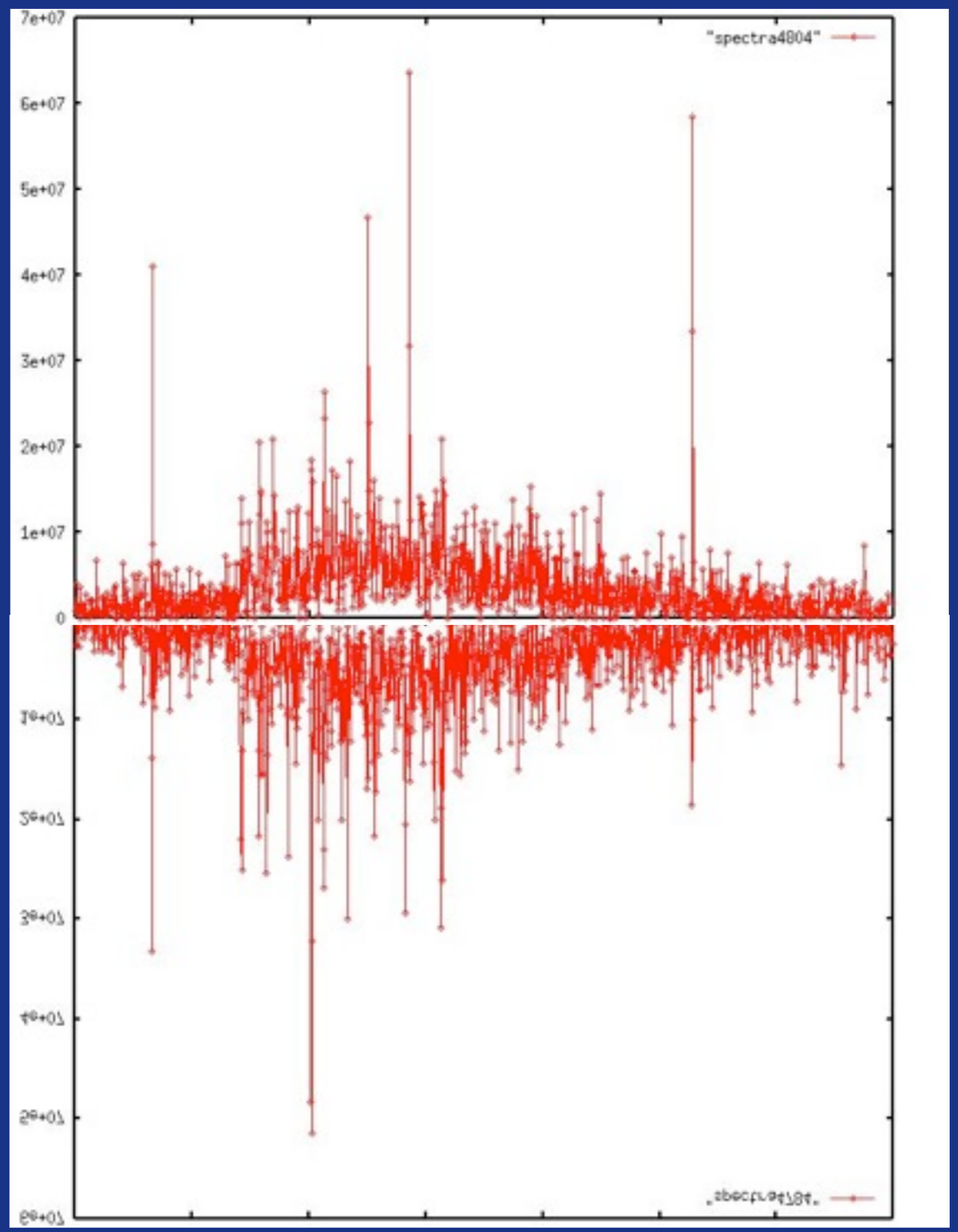
…LPGNARKMDKSTVLQKTIDF…
…LPGNARKMDKSTVLQ-EIDF…

- Scoring function BLOSUM62

BLOSUM62 (E,Z)= 5.0

BLOSUM62 (E,Z)= 1.0

The “dot product” score function for comparing two spectra



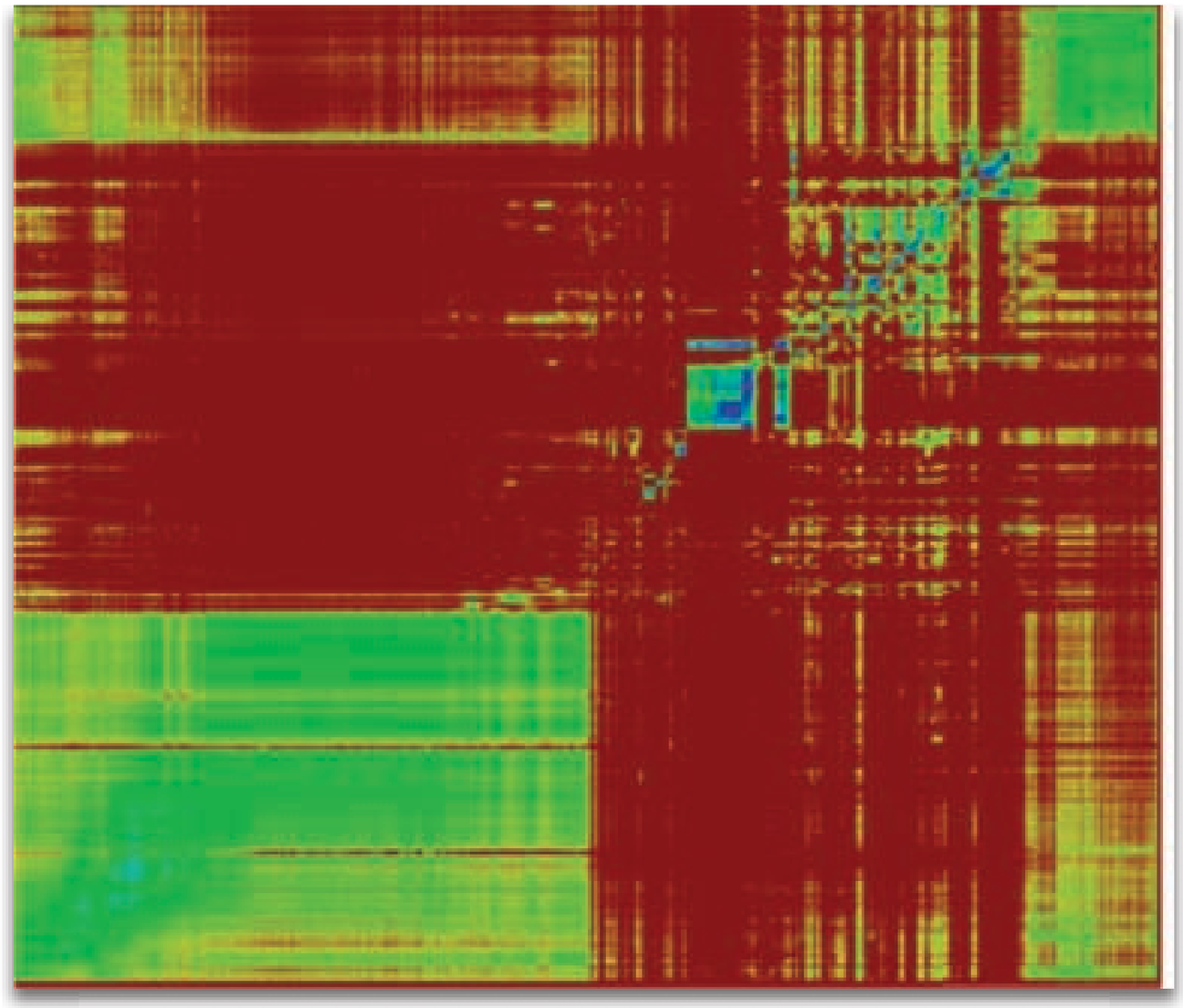
Idea
(Stein & Scott, 1994)

1. Reward all “close enough” peak pairs by amplitude product

2. Add over all such peak pairs

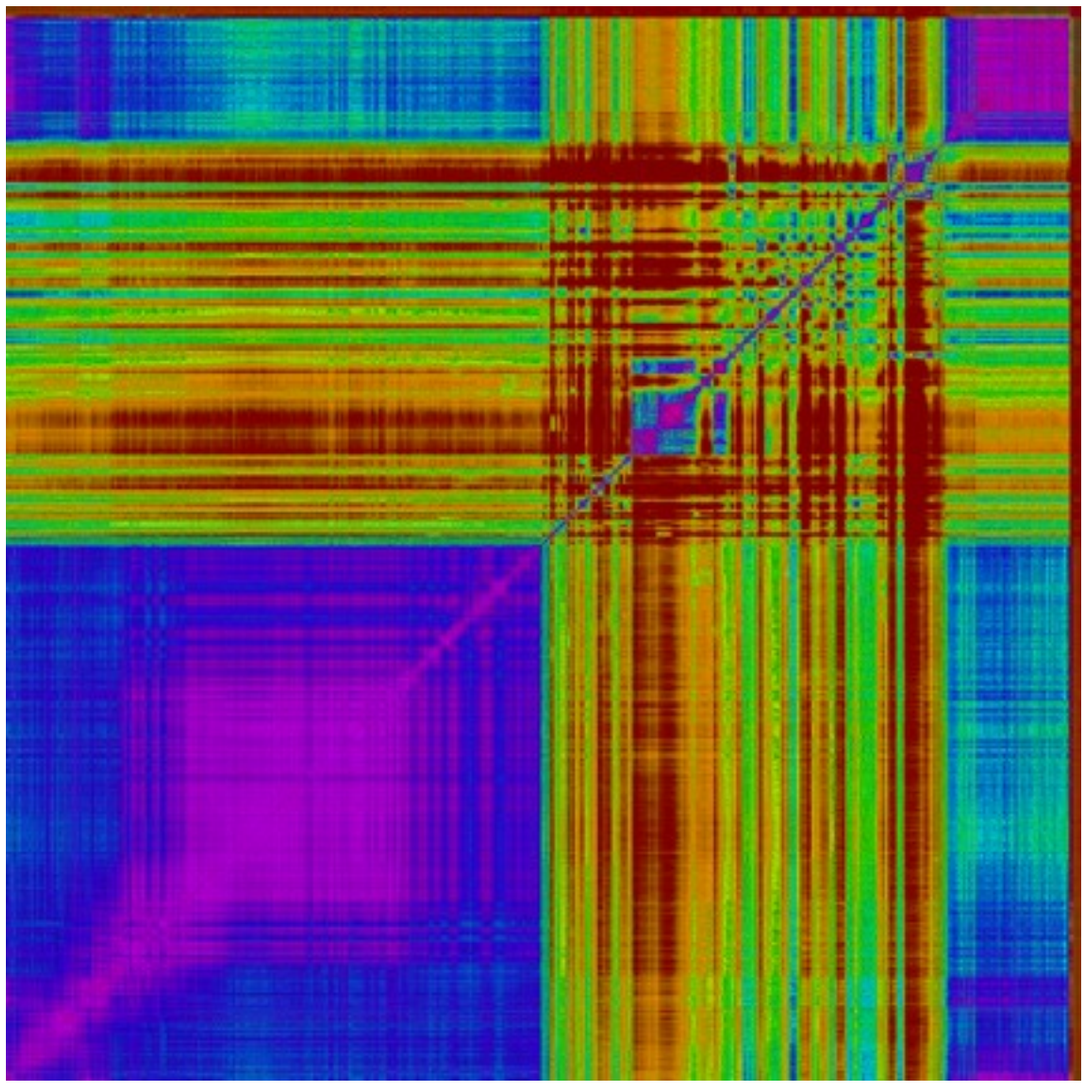
3. Normalize by total peak sum

?wv ?m?f A?? t zk?

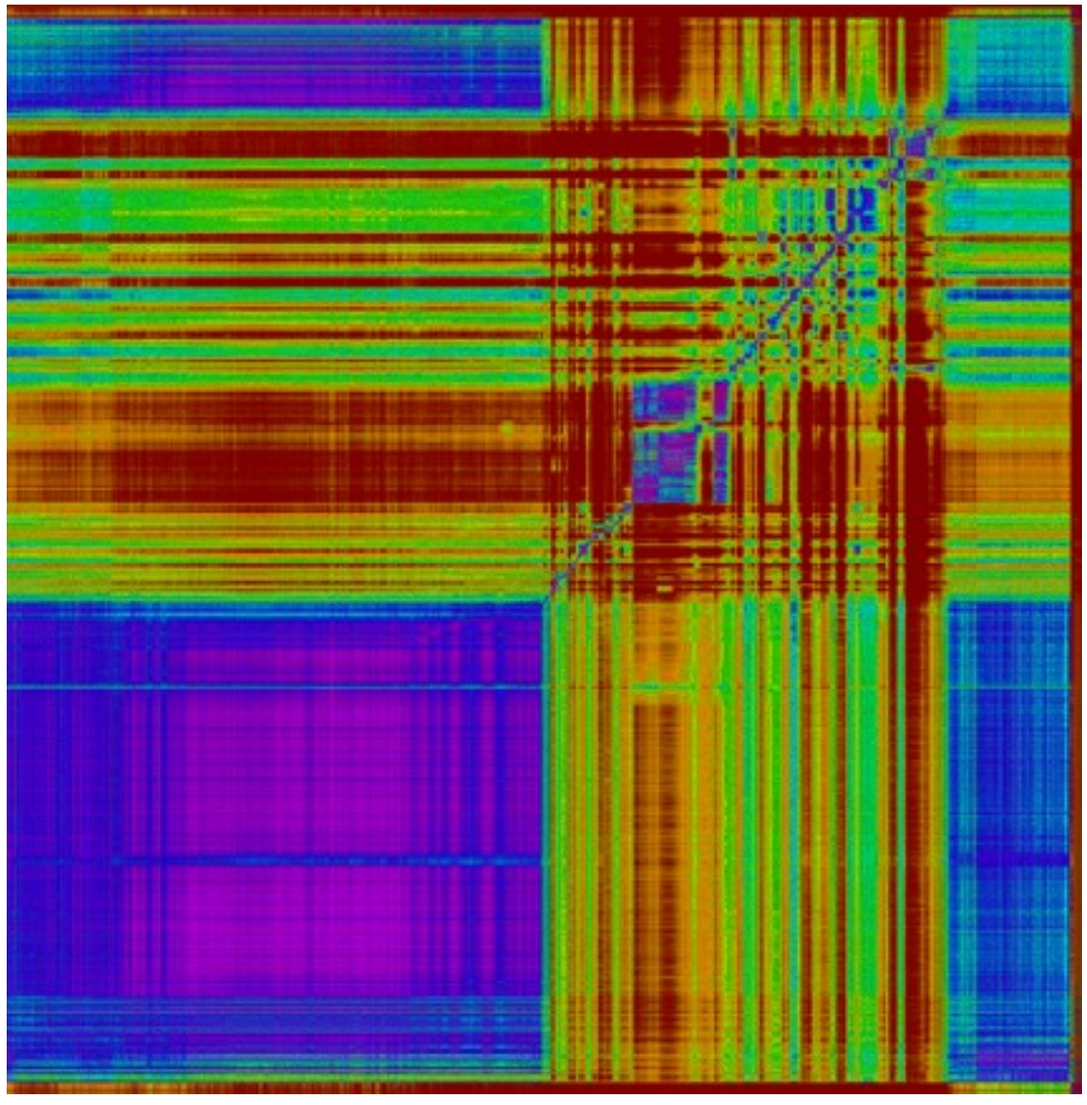


(Data from SCX fractions 23A and 24A, yeast cell cycle data, t=0, Mark Flory)

Problem: Noise



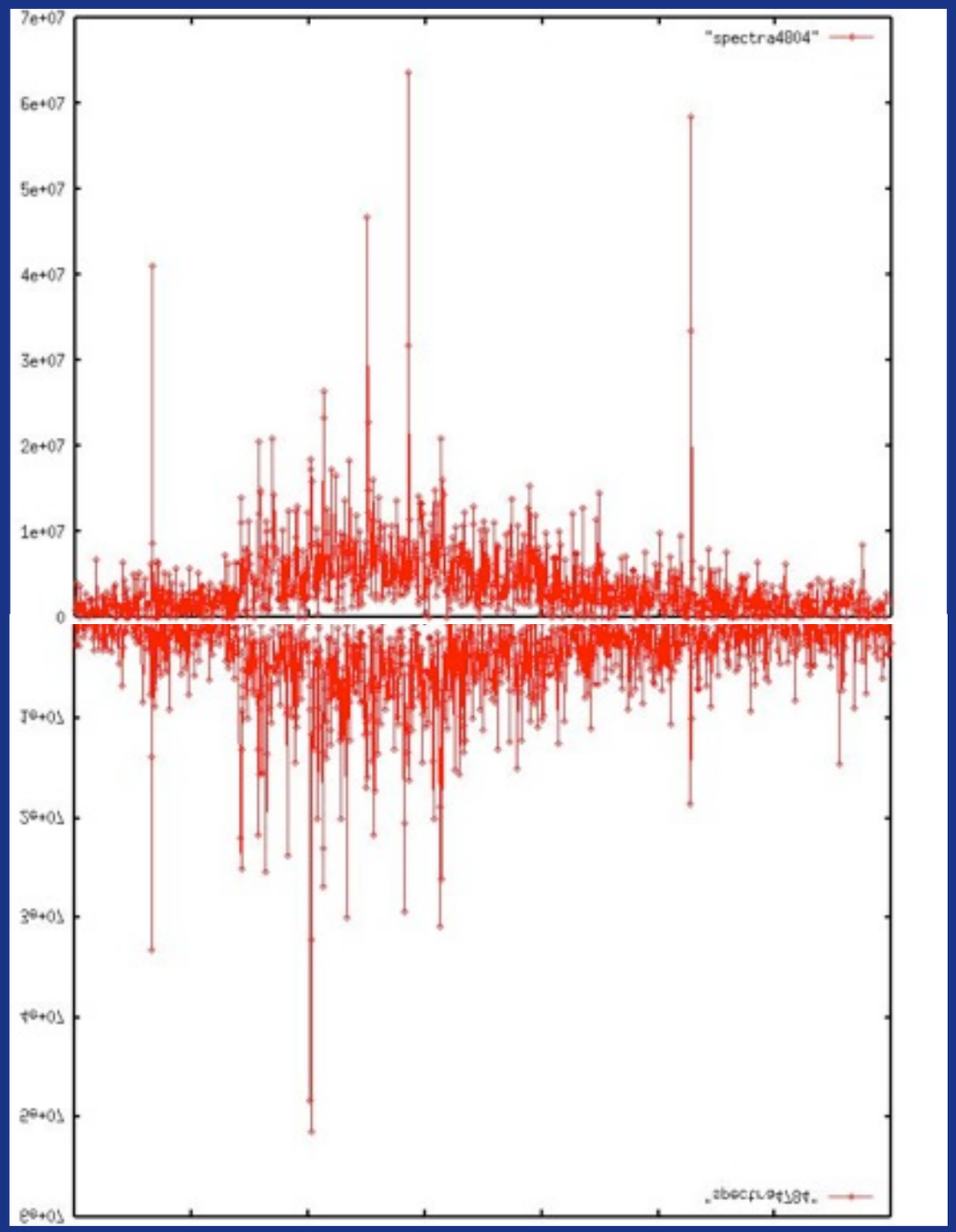
23A vs. 23A



23A vs. 23B

(Data from SCX fractions 23A and 23B, yeast cell cycle data, t=0, Mark Flory)

Modified score function



Modifications

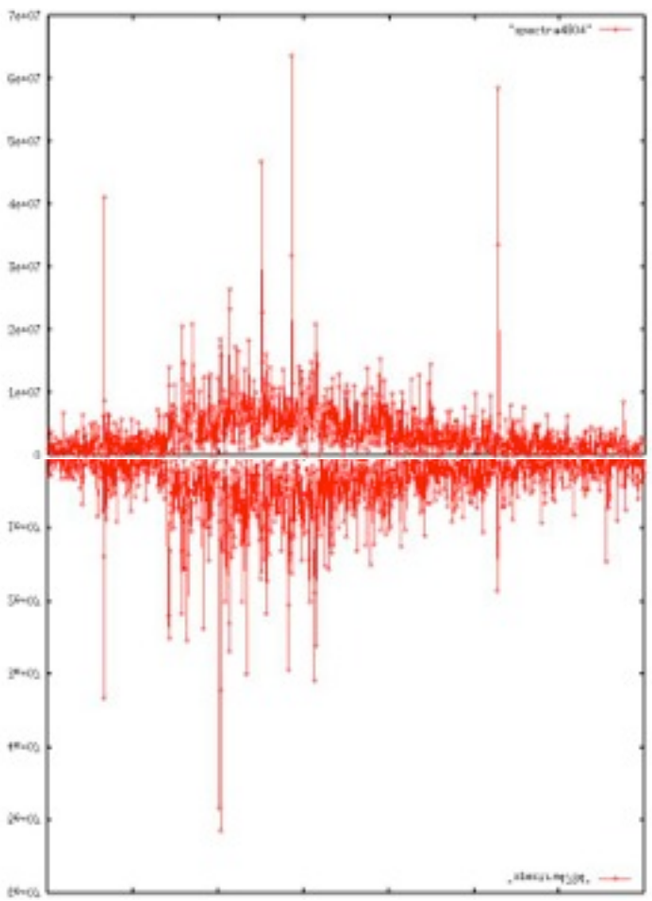
- "Background subtraction": Locally subtract spectrum "noise level" before comparison

- Subtract match score between randomized spectra before normalization, so expected value becomes 0

$$s(i,j) = \frac{M_i \times N_j - E(M_i \times N_j)}{\sqrt{(M_i \times M_i)(N_j \times N_j)}}$$

Computing the expected score

$$E(M_i \times N_j) = \sum_{\text{Peaks } s,t} p(s,t \text{ 'close enough'}) \cdot int(s) \cdot int(t)$$

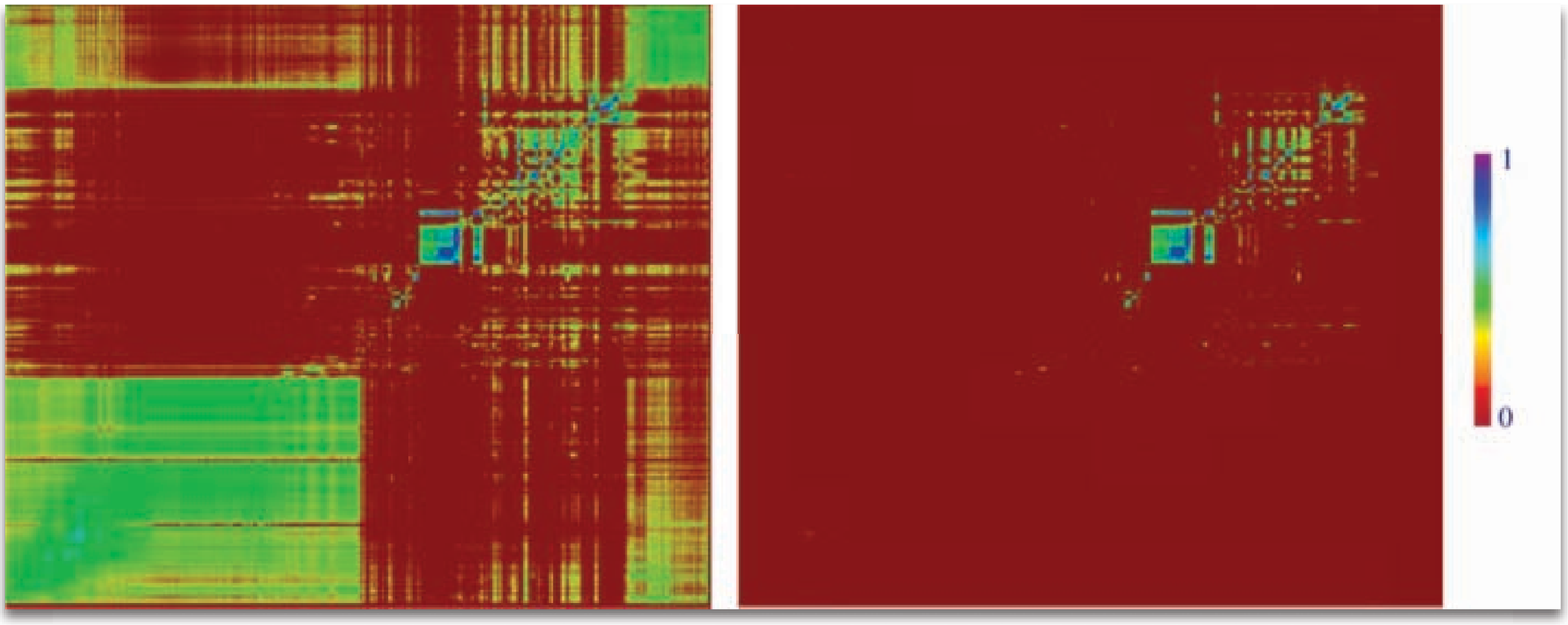


$$= \sum_{\text{Peaks } s,t} c \cdot int(s) \cdot int(t)$$

$$= c \cdot \left(\sum_s int(s) \right) \cdot \left(\sum_t int(t) \right)$$

$$= c$$

□ t □zs □□□□□zCf □Cwz□□k

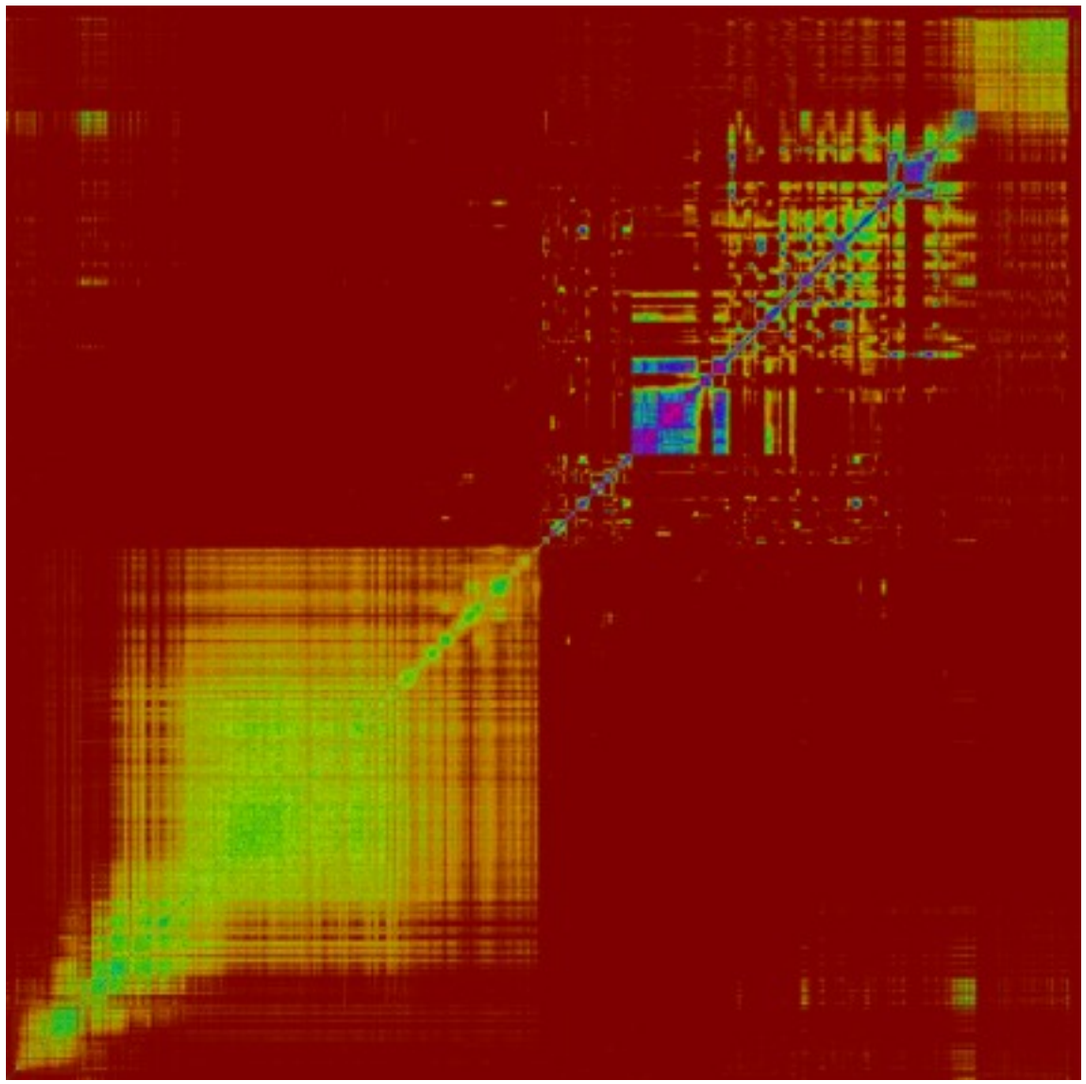


Stein-Scott score

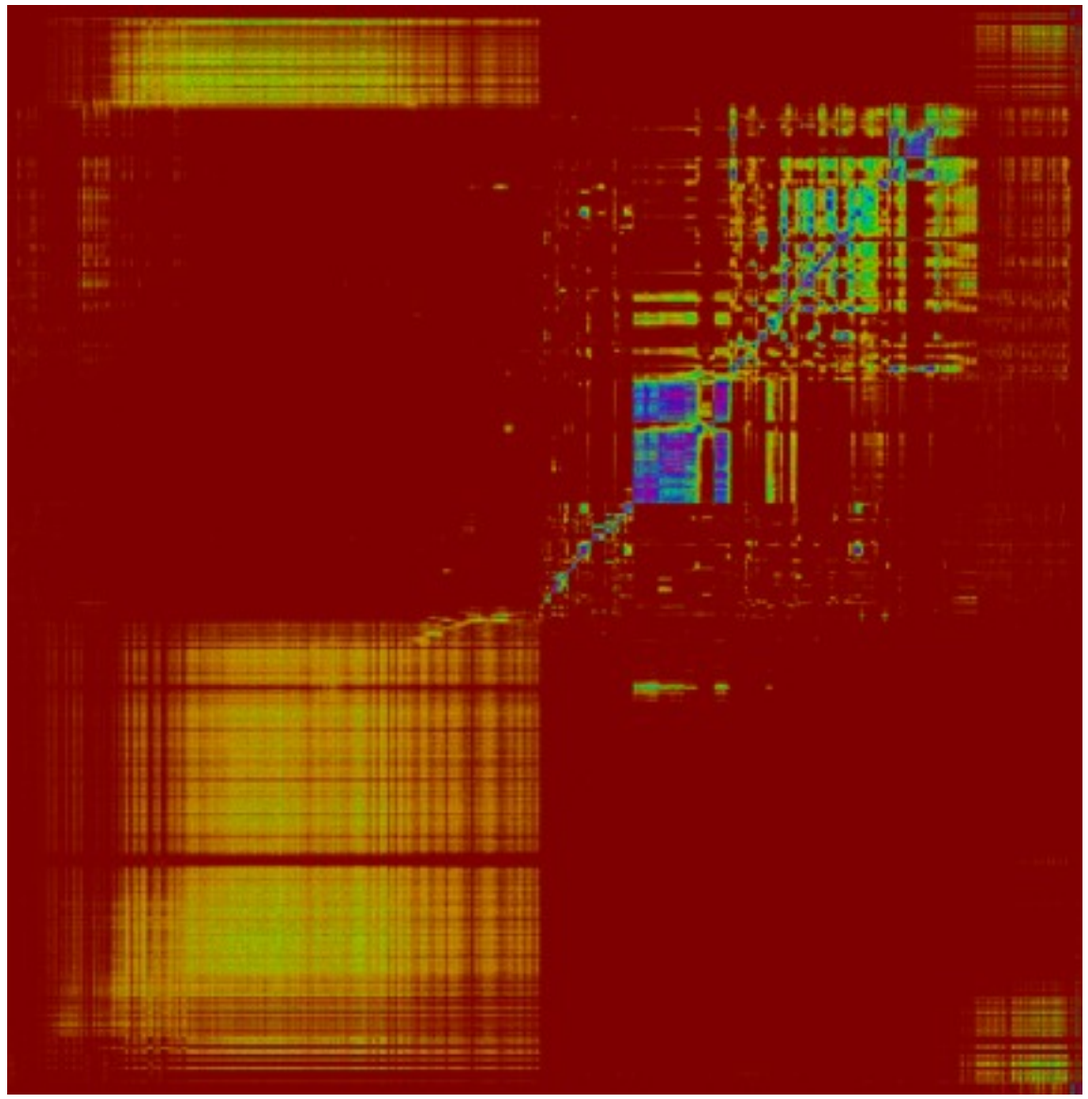
Modified Stein-Scott score

(Data from SCX fractions 23A and 24A, yeast cell cycle data, t=0, Mark Flory)

Modified edit matrices



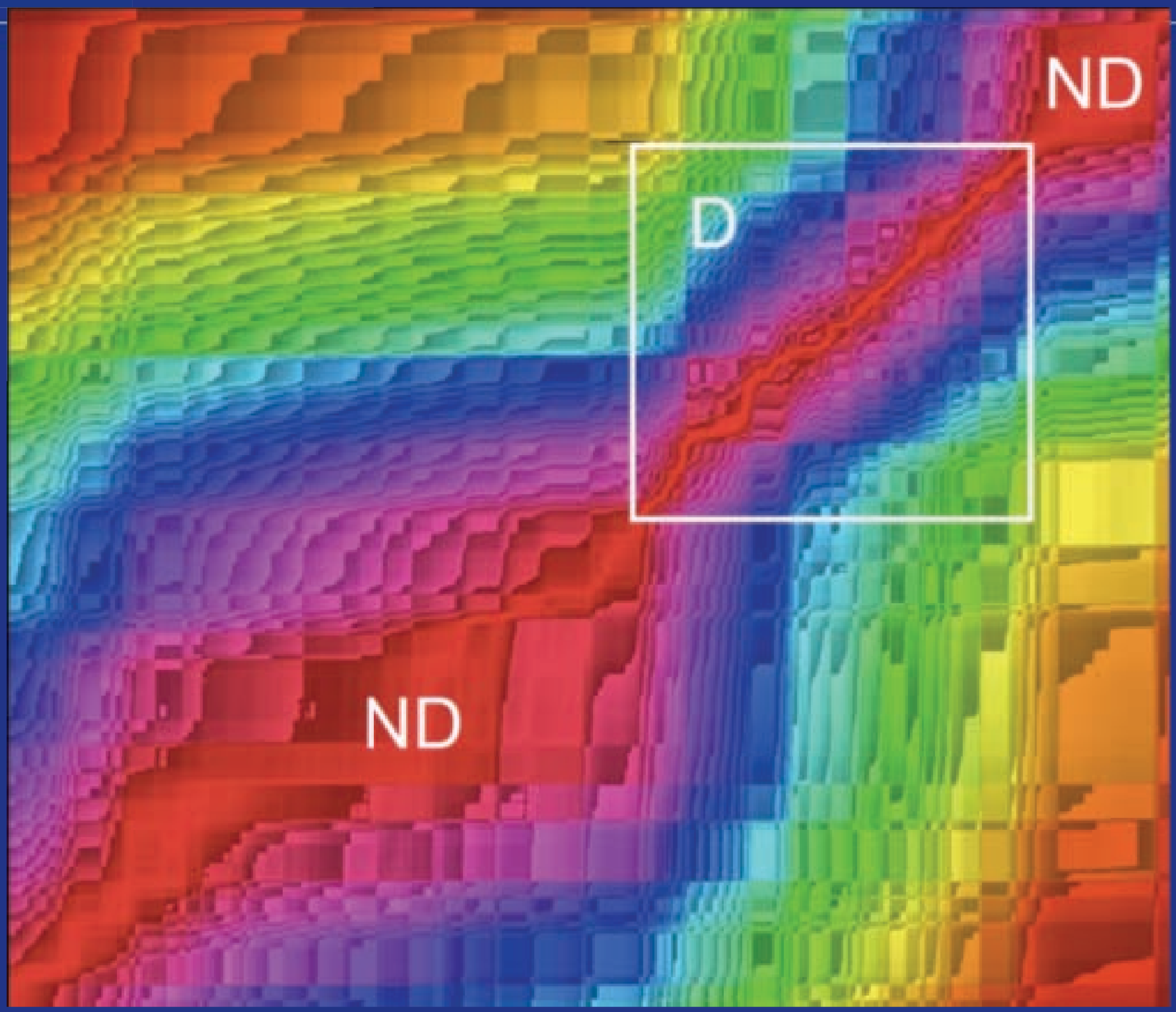
23A vs. 23A



23A vs. 23B

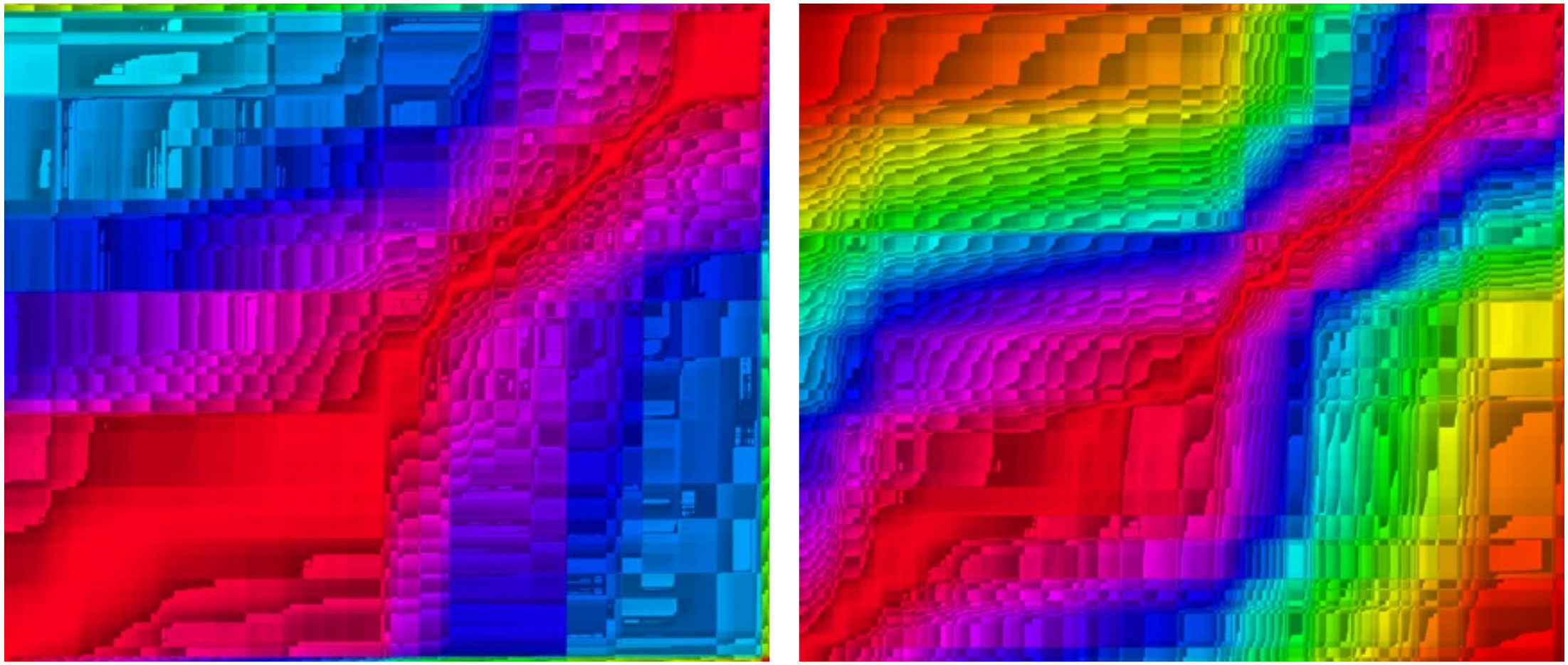
(Data from SCX fractions 23A and 23B, yeast cell cycle data, t=0, Mark Flory)

Regions of optimality and suboptimality



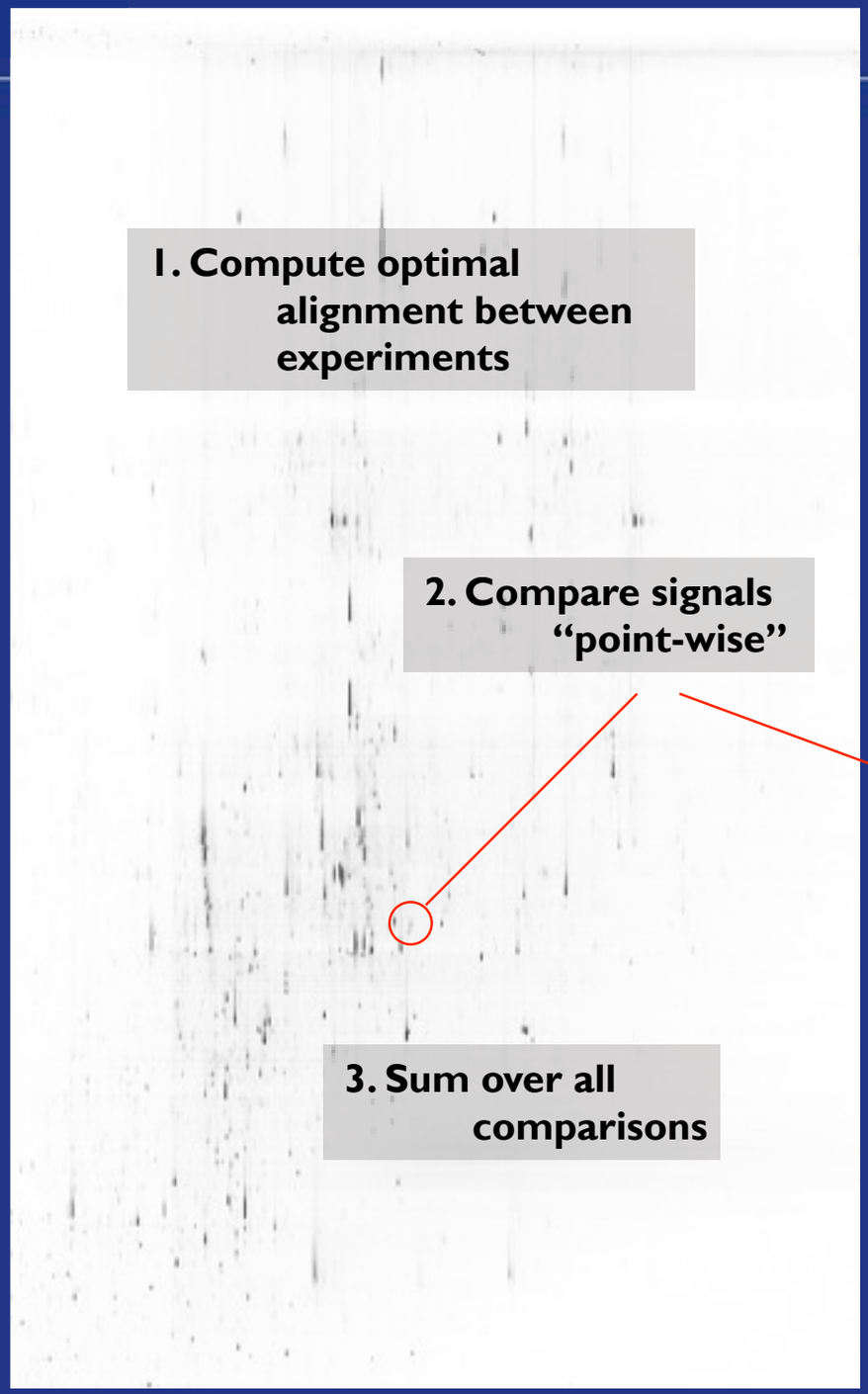
Prakash et al. (2006), Mol. Cell. Prot.

Improvement in alignment

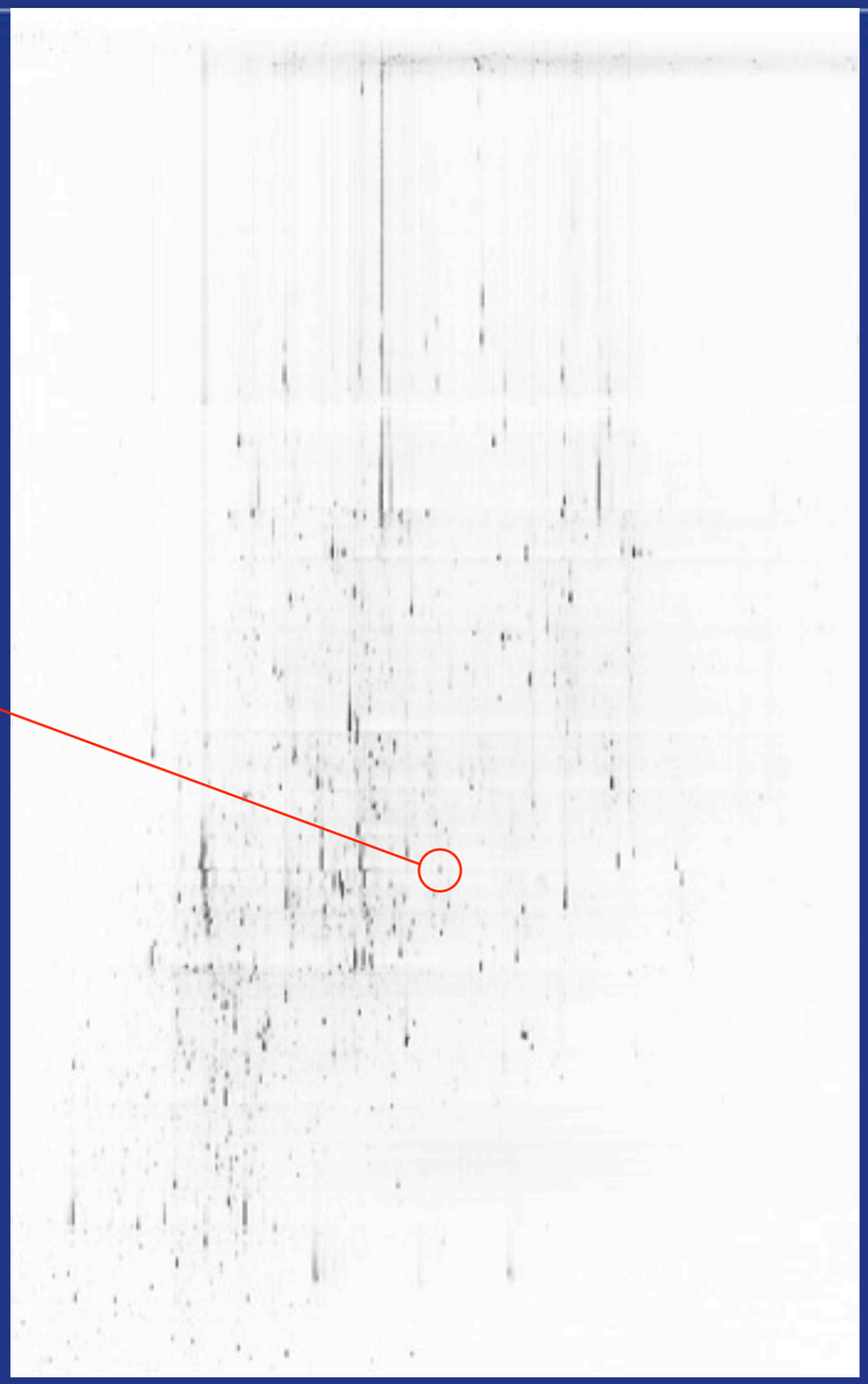


(Alignments of SCX fractions 23A and 24A, yeast cell cycle data, t=0, Mark Flory)

We are quantifying experimental similarity

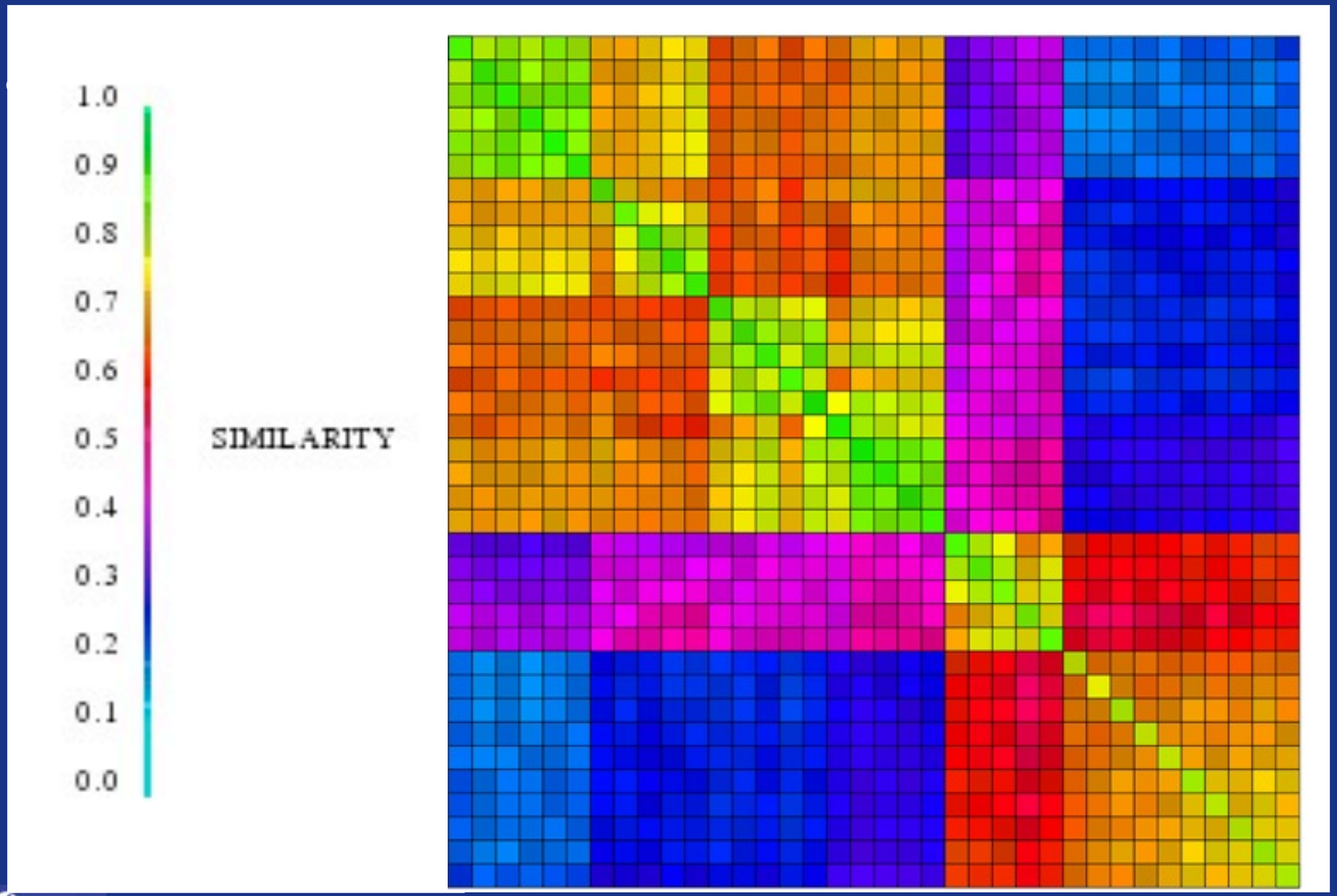


Time point 1



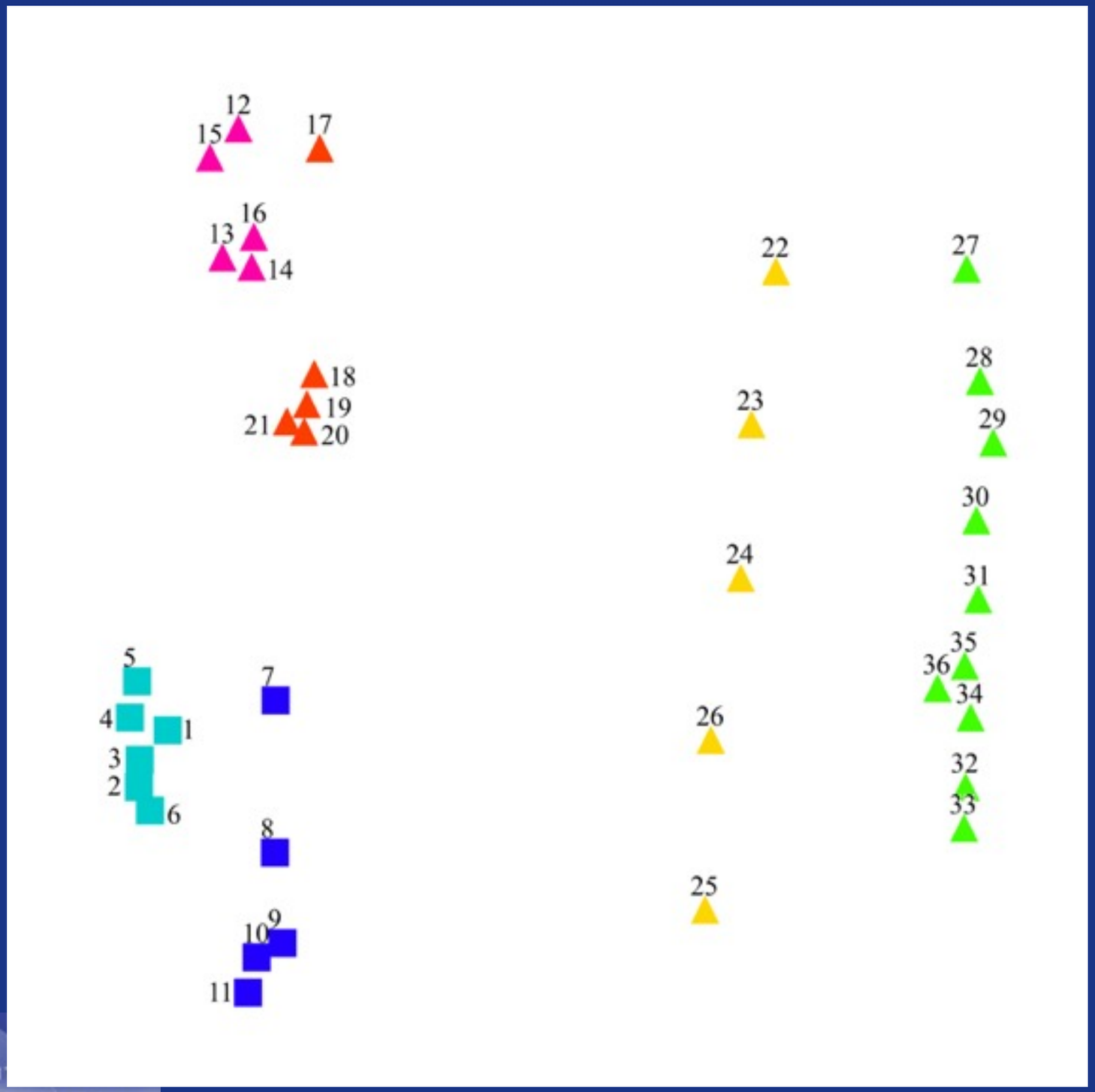
Time point 2

A distance matrix between experiments



Prakash et al. (2007), Mol. Cell. Prot.

A 2D embedding



Data
"Pure" angiotensin II

Purpose
QC

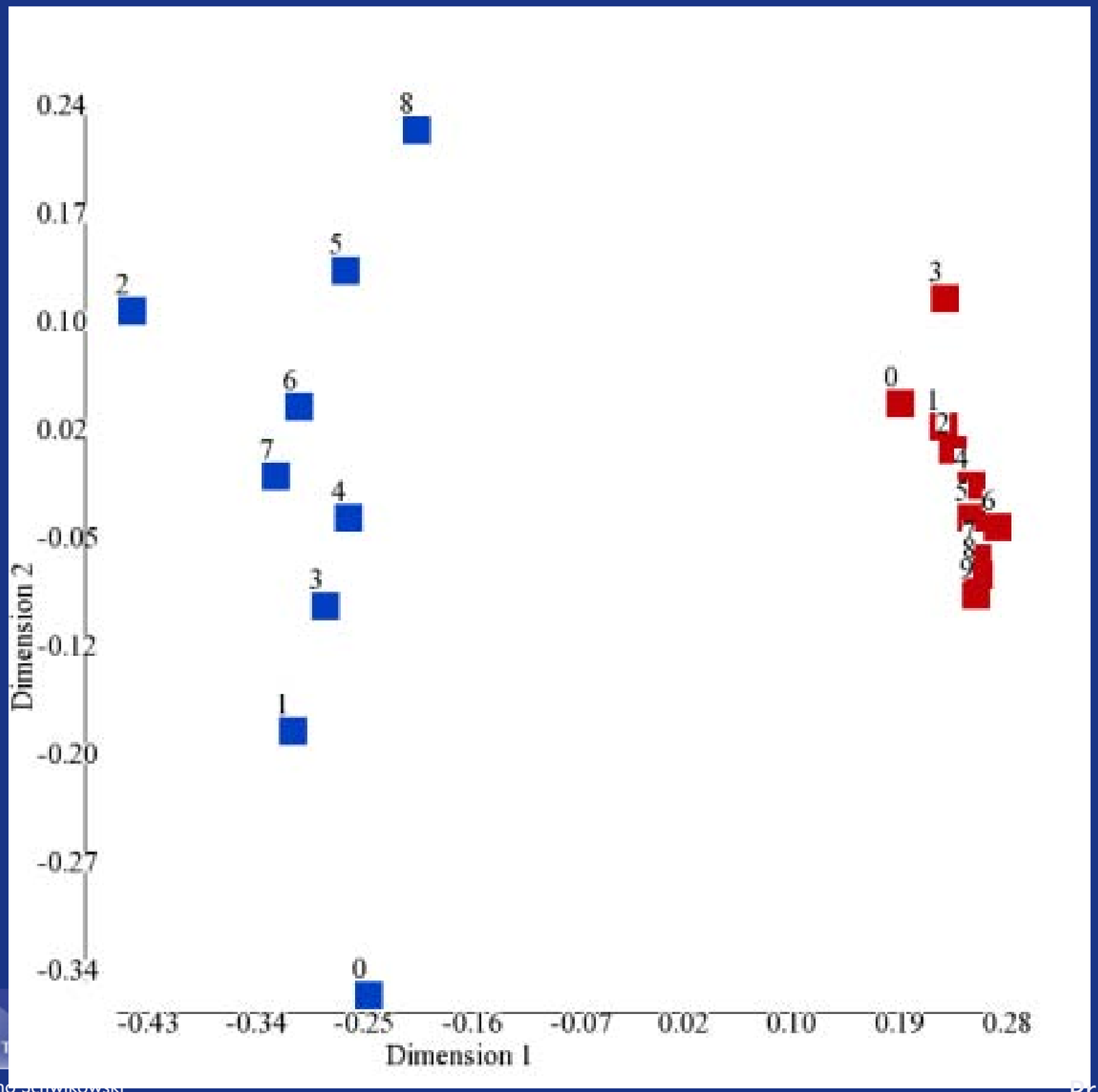
Instrument
FT-ICR

Colors
Different days

Shapes
Different LC columns

Prakash et al. (2007), Mol. Cell. Prot.

Cross-platform reproducibility



Data
Human blood serum
Identical samples

Blue/Red
QSTAR/QTOF

Numbers
Run order

Prakash et al. (2007), Mol. Cell. Prot.

Summary

- Systems Biology faces biological complexity
- Models must capture complexity along several axes
- Visualization helps...
  - Comprehend data in an integrative manner
  - Develop experimentally testable hypotheses and models
  - Comprehend complex relationships

Challenges around visualization in network biology

- Dense networks, “scrollable”
- Visualization capturing dynamic behavior
- Visualization of uncertainty (in the network)
- Standardized access to different data sources
- Many tools are not interactive
- More effective
- Development and maintenance of *tools* in an environment that rewards *publications*
- [...]

Acknowledgements

Cytoscape labs

Leroy Hood (ISB)
Trey Ideker (UCSD)
Gary Bader (U. Toronto)
Chris Sander (MSKCC New York)
Annette Adler (Agilent)

Mike Smoot (UCSD)
Oriol Guitart (IP)

Proteomics labs

Ruedi Aebersold (ETH Zürich)
David Goodlett (U. Washington)
Tina Guina (U. Washington)
Parag Mallick (LA)

Amol Prakash (U. Washington)
Freddy Cliquet (IP)

Thanks for your attention

http://systemsbiology.fr
http://cytoscape.org