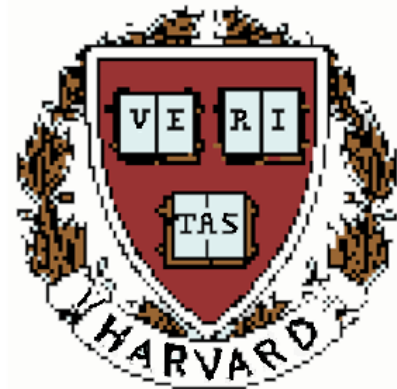
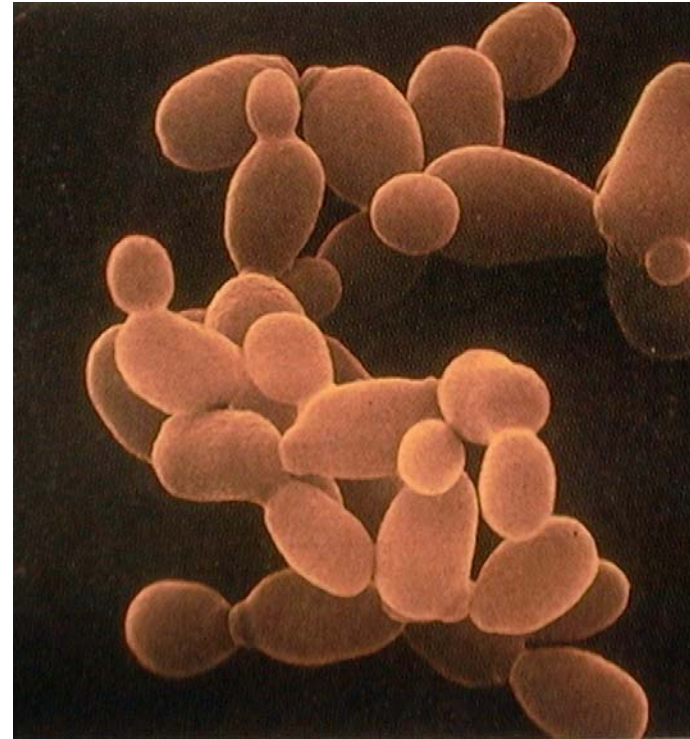


Pathway Based Analysis Of Microarray Data

Duccio Cavalieri

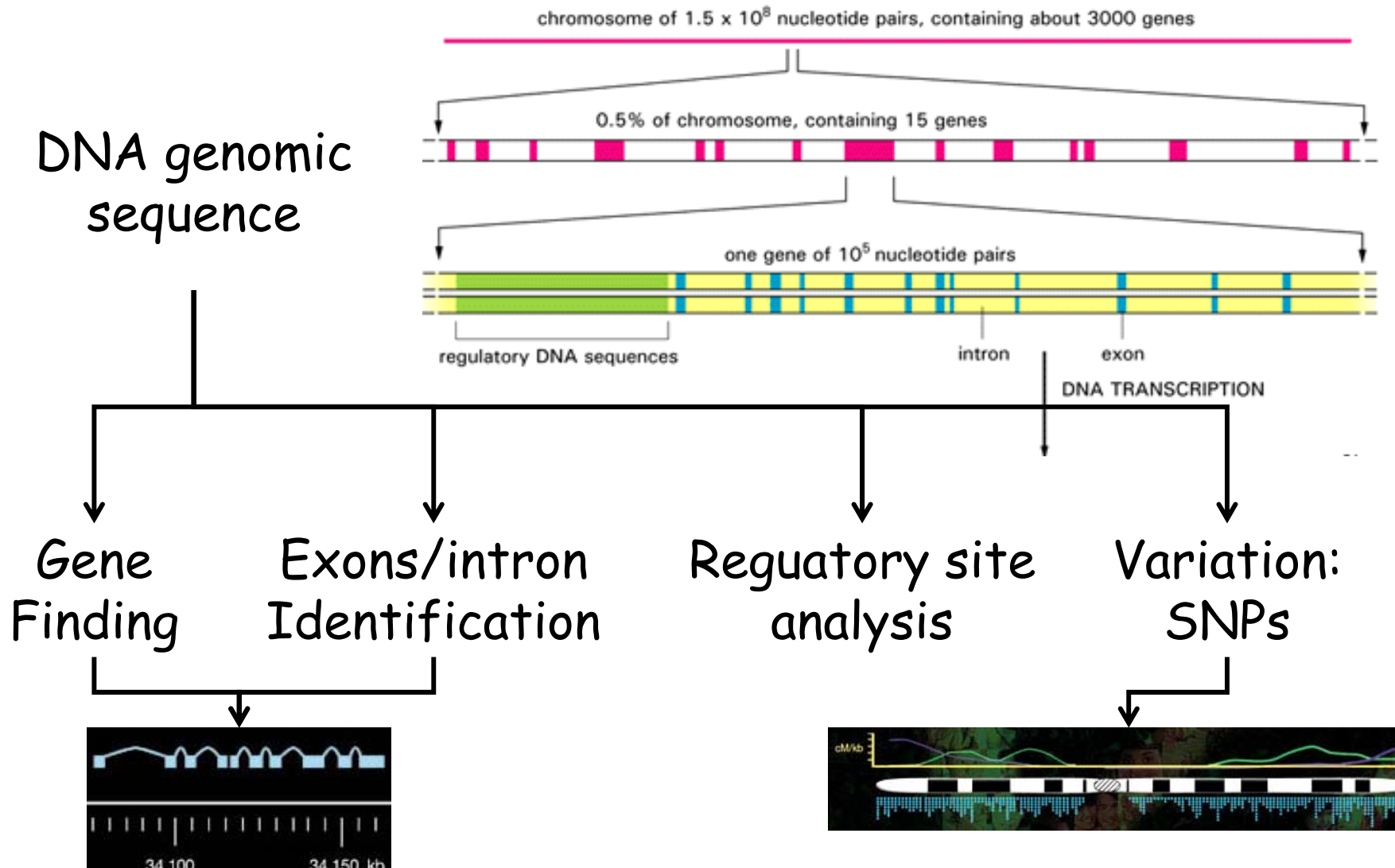
Dept of Pharmacology
University of Florence



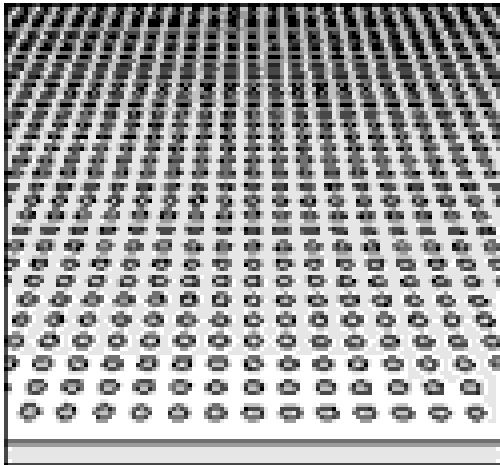
The Nature of Biological Information

- Hierarchical levels of information
- At every level
 - Redundancy
 - Multiplicity
 - Variability
- A MOLECULAR parts-list approach

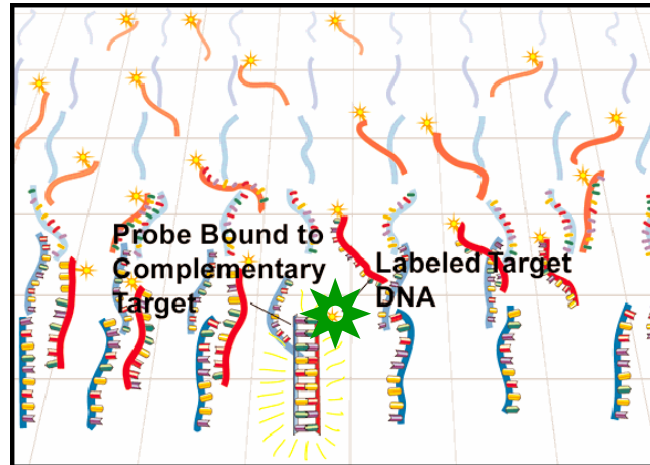
The molecular parts list making sense of the Sequence ...



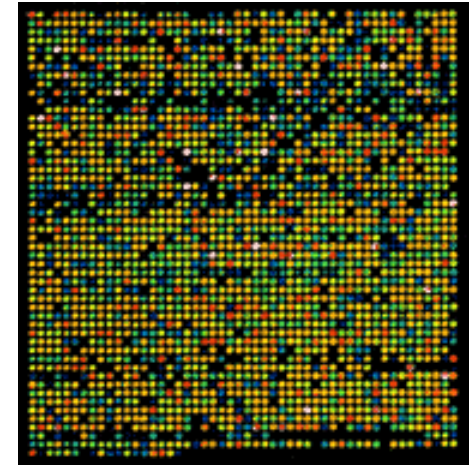
The molecular “parts-list”: Transcriptosome Snapshots: Expression Profiling



cDNA bonded on
a glass surface



Label RNA from
cell and hybridize
to array

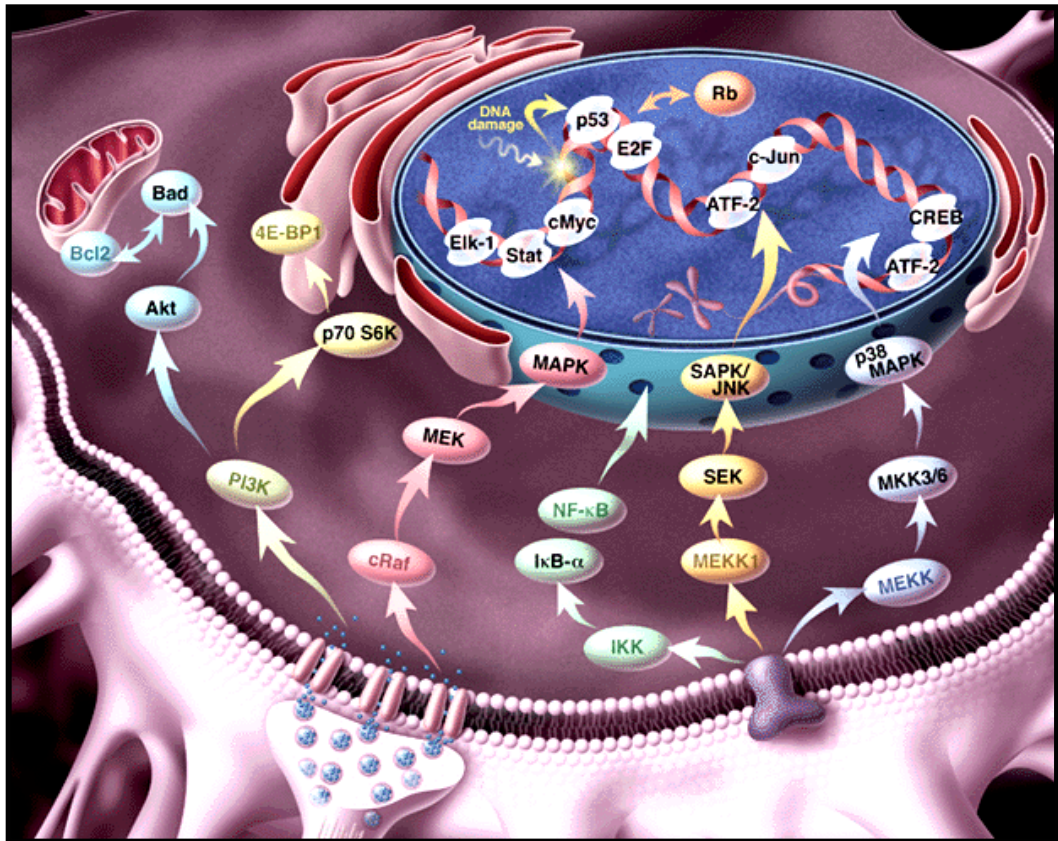


Scanned,
hybridized array

Camera
(Microarray)

Snapshot
(Expression profil

The molecular "parts-list": The proteome



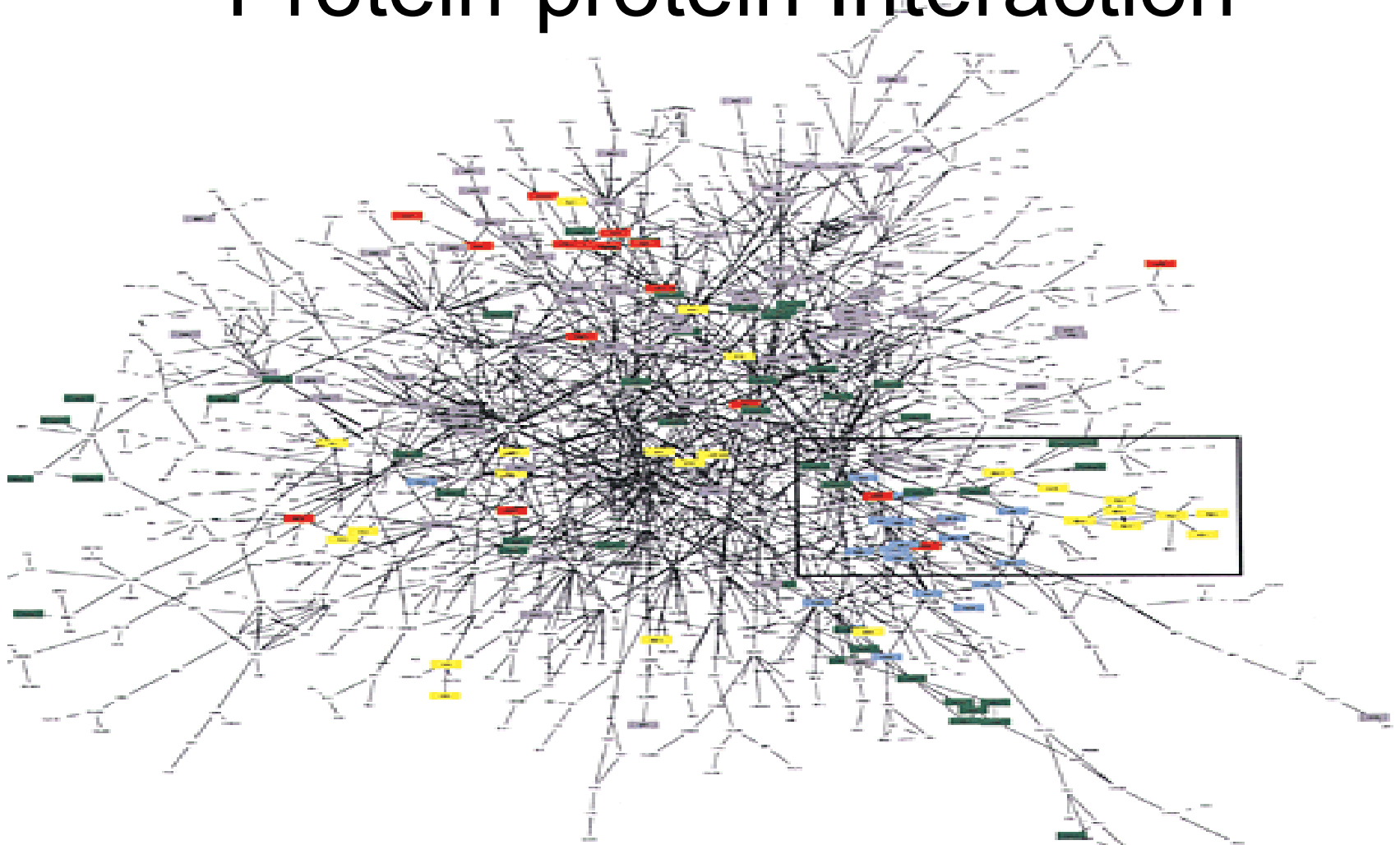
Translation
Degradation

Localization

Modification
(binding, cleavage,
covalent modification)

0.5-1X10⁶ variants

Multiple Faces of the Proteome: Protein-protein Interaction



High throughput 2-hybrid analysis of protein-protein
interaction in yeast



Metabolomics



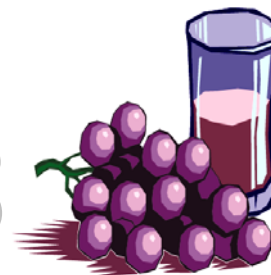
Vitis vinifera

Volatile organics important for wine flavor

Grapes: 466

Wine: 644

Difference: 178



More than 200 genomes have been sequenced

<http://www.ncbi.nlm.nih.gov/>

But we are still searching for the holy grail!!!



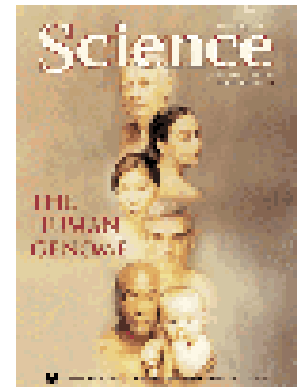
Bacteria
1.6Mb
1600 genes



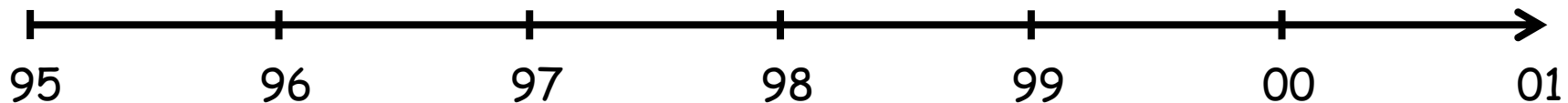
Eukaryote
13Mb
~6000 genes



Animal
100Mb
~20,000 genes



Human
3Gb
~30,000 genes?



GENOMES to LIFE

ACCELERATING BIOLOGICAL DISCOVERY

A NEW PROGRAM PROPOSED BY
THE U.S. DEPARTMENT OF ENERGY

DNA SEQUENCE DATA
FROM GENOME PROJECTS

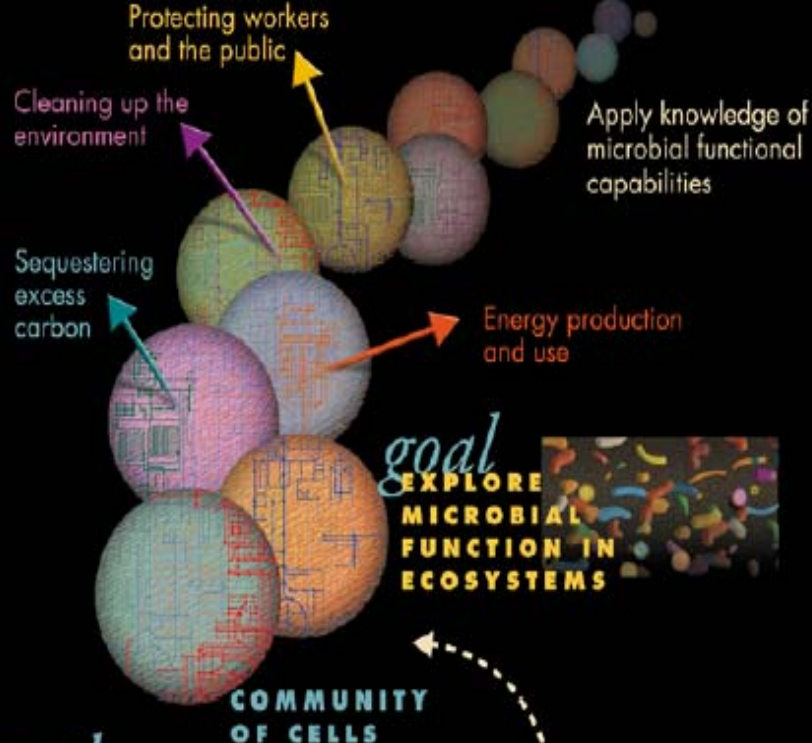
SCIENCE FOR THE 21ST CENTURY

Genes and other
DNA sequences
contain instructions
on how and when
to build proteins

goal
**IDENTIFY
PROTEIN
MACHINES**

PROTEINS

Proteins perform all of life's most essential functions. To carry out their specific roles, they often work together in the cell as protein machines.



goal
**DEVELOP COMPUTATIONAL
TOOLS TO MODEL
AND PREDICT
RESPONSES OF
CELLS TO
ENVIRONMENT**

WORKING
CELL

Many protein
machines interact
through complex,
interconnected
pathways. Analyzing
these dynamic processes
will lead to a model of a
living cell.

goal
**CHARACTERIZE GENE
REGULATORY NETWORKS**

URL DOEGenomesToLife.org

The Genomics paradox

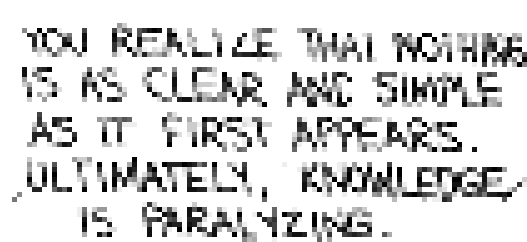
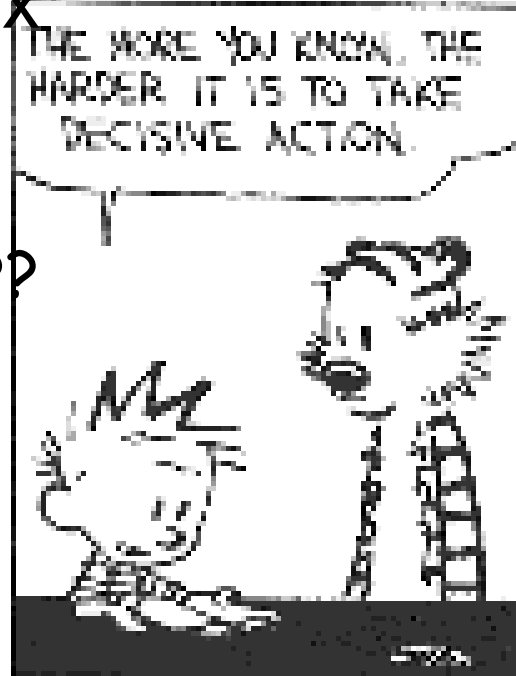
The more you know
the harder is
to take decisive action ???

A man of action
cannot afford this risk

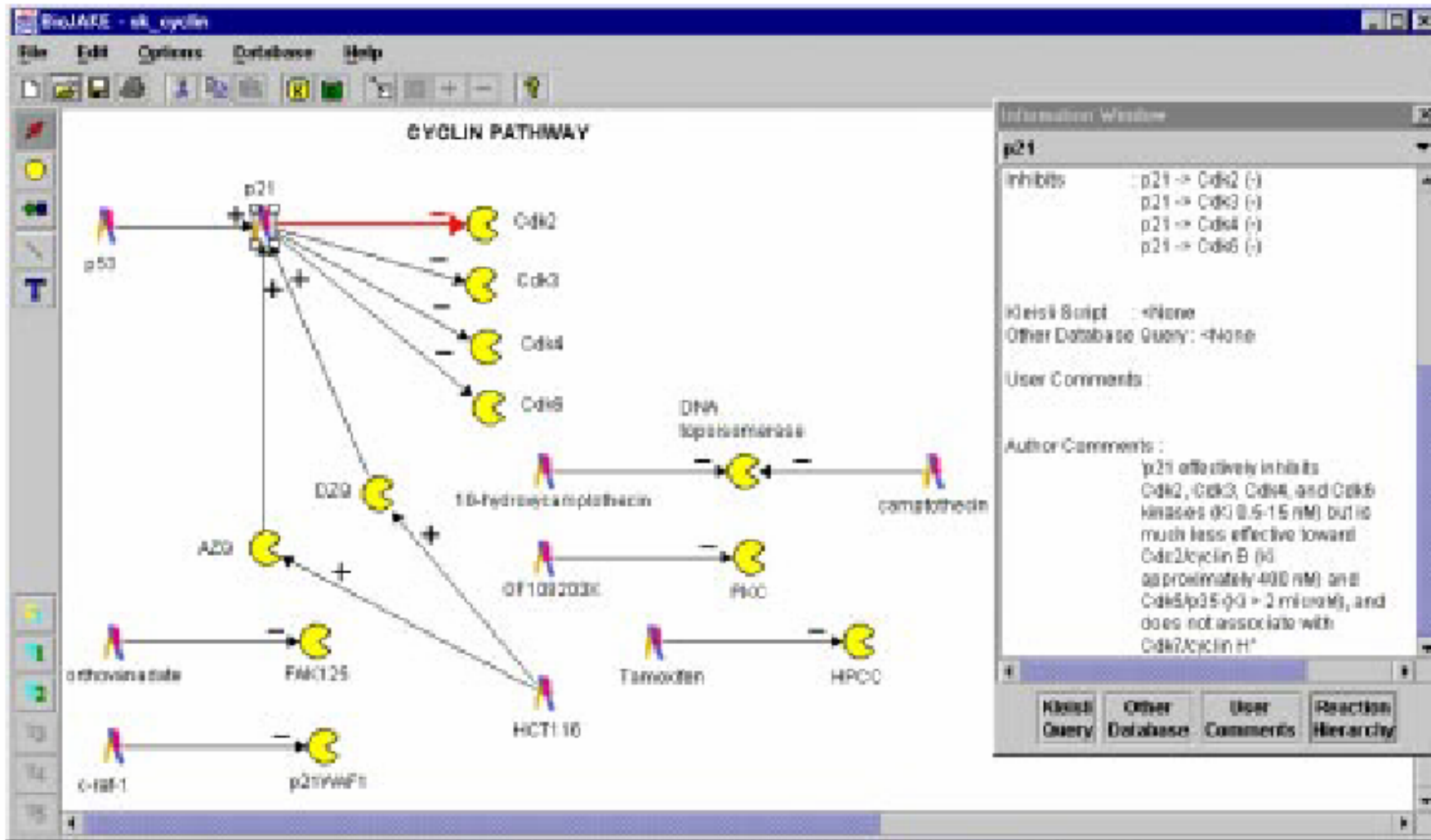
Either go back to a
reductionist approach

OR....

apply tools enabling to
handle complexity



Challenge 1- Connect the data we produce to what we already know



A protein-protein interaction pathway map automatically constructed from a user query of effective human cyclin inhibitor".

Integration of whole genome gene expression data with biological function

REPORTS

Functional Characterization of the *S. cerevisiae* Genome by Gene Deletion and Parallel Analysis

Elizabeth A. Winzeler,^{1*} Daniel D. Shoemaker,^{2*} Anna Astromoff,^{1*} Hong Liang,^{1*} Keith Anderson,¹ Bruno Andre,³ Rhonda Bangham,⁴ Rocio Benito,⁵ Jef D. Boeke,⁶ Howard Bussey,⁷ Angela M. Chu,¹ Carla Connelly,⁶ Karen Davis,¹ Fred Dietrich,⁸ Sally Whelen Dow,² Mohamed El Bakkoury,⁹ Françoise Foury,¹⁰ Stephen H. Friend,² Erik Gentalen,¹¹ Guri Giaever,¹ Johannes H. Hegemann,¹² Ted Jones,¹ Michael Laub,¹ Hong Liao,⁴ Nicole Liebundguth,⁸ David J. Lockhart,¹¹ Anca Lucau-Danila,¹⁰ Marc Lussier,⁷ Nasiha M'Rabet,³ Patrice Menard,⁷ Michael Mittmann,¹¹ Chai Pai,¹ Corinne Rebischung,⁸ Jose L. Revuelta,⁵ Linda Riles,¹³ Christopher J. Roberts,² Petra Ross-MacDonald,⁴ Bart Scherens,⁹ Michael Snyder,⁴ Sharon Sookhai-Mahadeo,⁶ Reginald K. Storms,⁷ Steeve Véronneau,⁷ Marleen Voet,¹⁴ Guido Volckaert,¹⁴ Teresa R. Ward,² Robert Wysocki,¹⁰ Grace S. Yen,¹ Kexin Yu,⁶ Katja Zimmermann,¹² Peter Philippsen,⁸ Mark Johnston,¹³ Ronald W. Davis^{1†}

Large-scale analysis of the yeast genome by transposon tagging and gene disruption

Petra Ross-Macdonald^{*}, Paulo S. R. Coelho^{*}, Terry Roemer^{*}, Seema Agarwal^{*}, Anuj Kumar^{*}, Ronald Jansen[†], Kei-Hoi Cheung[‡], Amy Sheehan^{*}, Dawn Symoniatis^{*}, Lara Umansky^{*}, Matthew Heidtman^{*}, F. Kenneth Nelson^{*}, Hiroshi Iwasaki^{*}, Karl Hager[§], Mark Gerstein[†], Perry Miller[‡], G. Shirleen Roeder^{*} & Michael Snyder^{*†}

NATURE | VOL 402 | 25 NOVEMBER 1999 | www.nature.com

© 1999 Macmillan Magazines Ltd

41

letters to nature

Functional genomic analysis of cell division in *C. elegans* using RNAi of genes on chromosome III article

Pierre Gönczy^{*†‡}, Christophe Echeverri^{*†‡}, Karen Oegema^{*†}, Alan Coulson[§], Steven J. M. Jones^{||}, Richard R. Copley[†], John Duperon[†], Jeff Oegema^{*}, Michael Brehm^{*‡}, Etienne Cassin^{*}, Eva Hannak^{*}, Matthew Kirkham^{*}, Silke Pichler^{*†}, Kathrin Flohrs^{*}, Anoesjka Goessen^{*}, Sebastian Leidel^{*}, Anne-Marie Alleaume^{*‡}, Cécilie Martin^{*‡}, Nurhan Özlü^{*}, Peer Bork[†] & Anthony A. Hyman^{*†}

NATURE | VOL 408 | 16 NOVEMBER 2000 | www.nature.com

© 2000 Macmillan Magazines Ltd

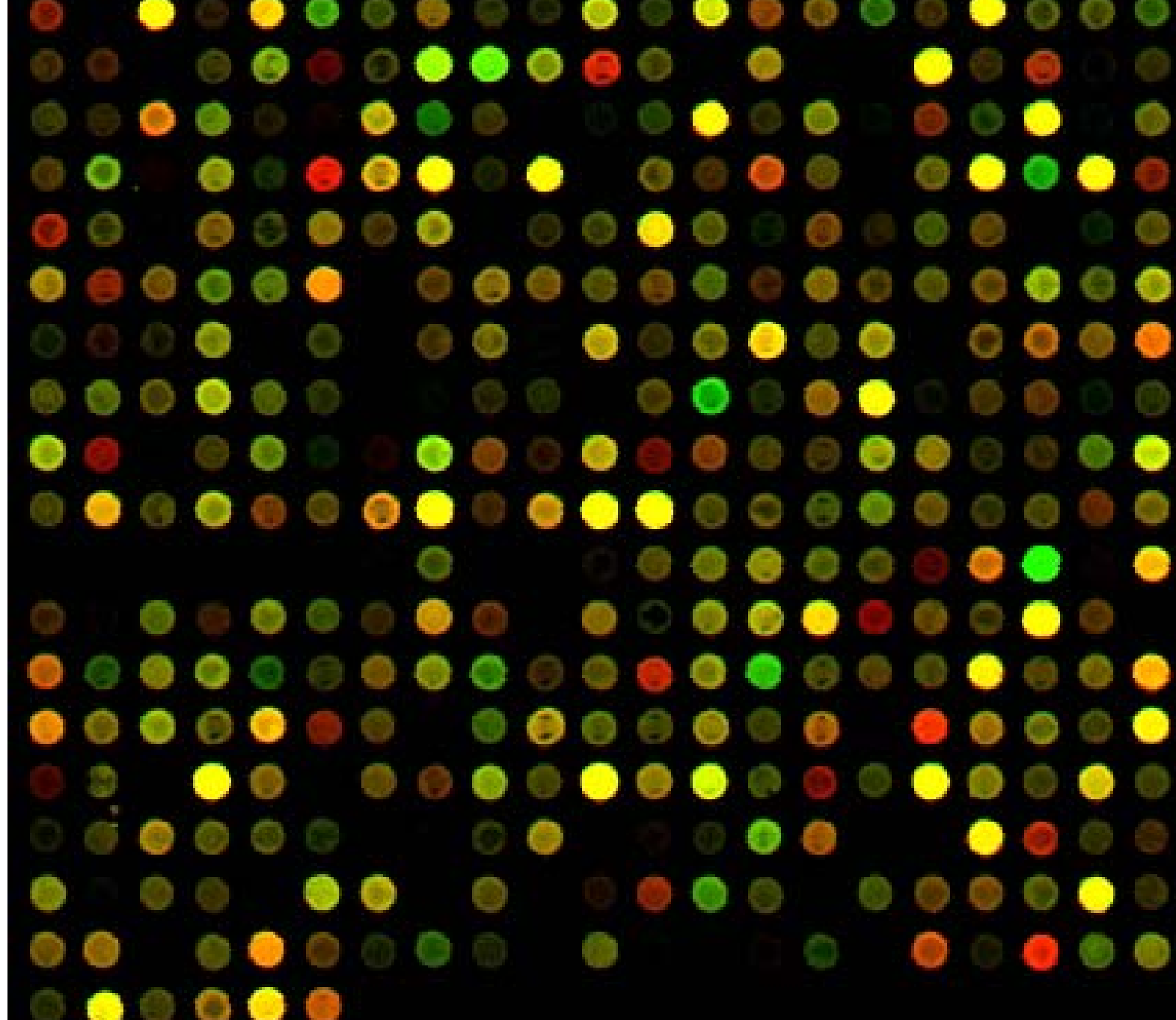
33

Functional genomic analysis of *C. elegans* chromosome I by systematic RNA interference article

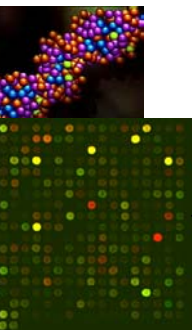
Andrew G. Fraser^{*†‡}, Ravi S. Kamath^{*†‡}, Peder Zipperlen^{*†‡}, Maruxa Martinez-Campos^{*}, Marc Sohrmann[‡] & Julie Ahringer^{*}

NATURE | VOL 408 | 16 NOVEMBER 2000 | www.nature.com

© 2000 Macmillan Magazines Ltd



Genomics data



Methods that use genomic information to infer correlations between genes and clusters of genes

Identification of similar core clusters of genes that are most strongly affected by coordinate regulation

Hierarchical Cluster

Non Hierarchical Cluster

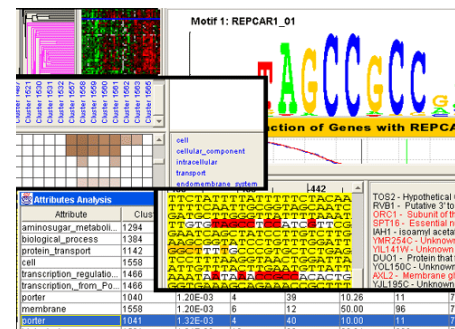
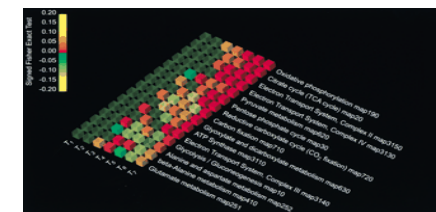
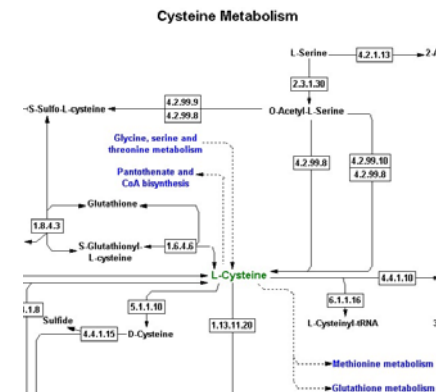
Methods that integrate functional genomics result in metabolic or cellular pathways

Methods that automatically display functional genomics results on metabolic or cellular charts

Methods that test for statistical significance of enrichment of genes belonging to pathway or network

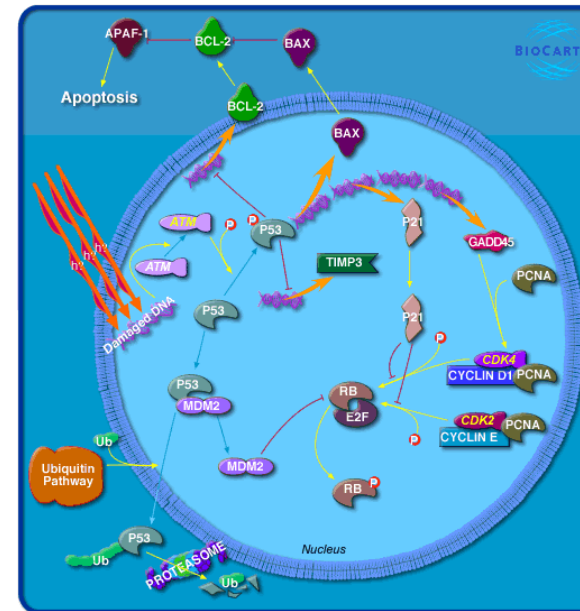
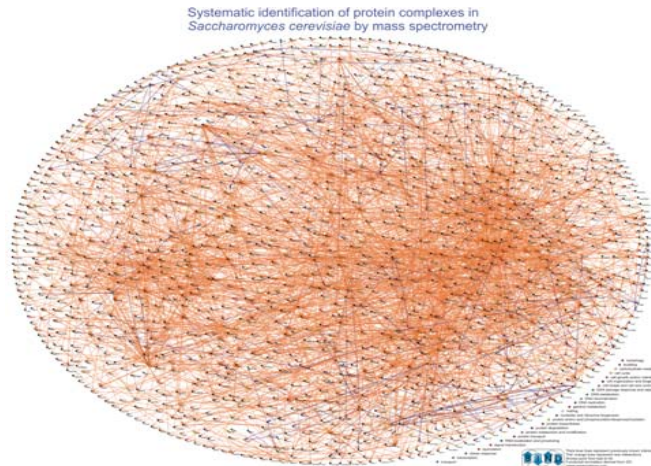
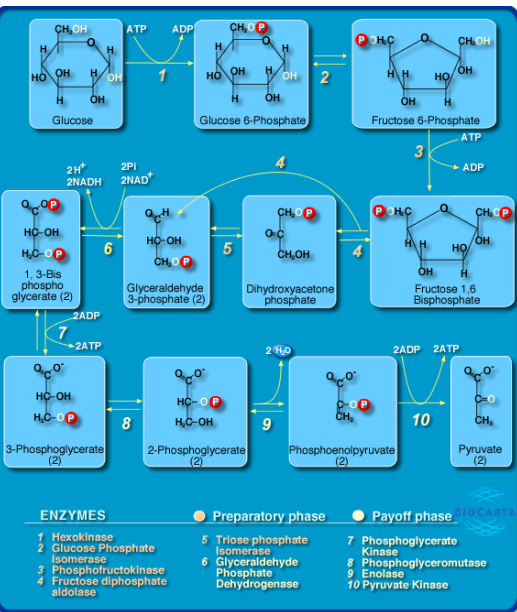
Methods that use functional genomics results, existing biological information, and clustering to reconstruct a novel biological network

Integration of the results from pathway-based analyses with those of clustering algorithms to provide information on genes of unknown function



What is a pathway?

Depends on who you ask



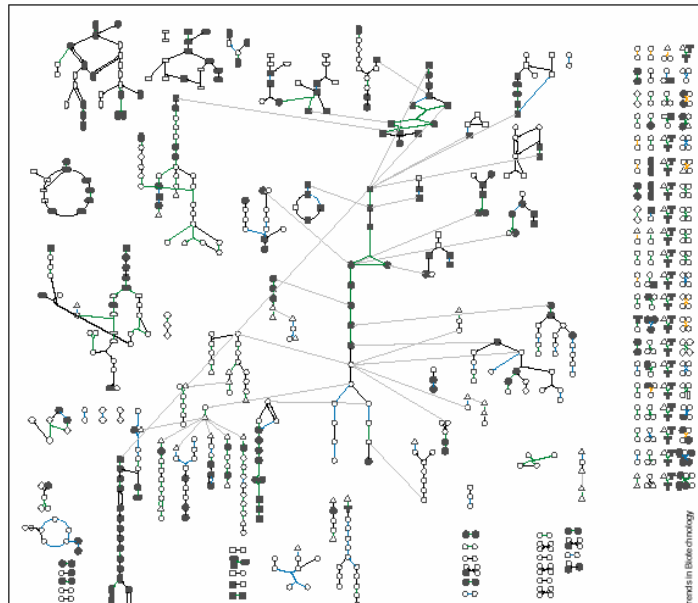
- Pathways are a way to describe biological networks.
- One of the main differences of a network and a pathway is that a pathway include a directionality, lines become arrows, with a given direction.

Open Goals and questions

- 1-How can we infer the actual regulatory networks from heterogeneous sources of data?
- 2-What are the limits of the existing probabilistic models in assessing pathway alteration?
- 3- How the genes and the pathways should be annotated in the different organisms?
- 4- How can we properly reconstruct novel pathways based on specific information in genomics?
- 5 Which are the best graphical models to visualize pathways?

Conceptualizing Biology: Ontologies, Databases and Visualization

- Ontologies: How to formalize our knowledge?
- Databases: How to store pathway information
- Statistics: How to determine significance?
- Visualization: How to display it?



Pathway Standards

Ontologies, Databases and Visualization

Ontologies: How to formalize our knowledge in order to have the most direct correlation between a gene and a function in a pathway? **Biopax!!!!** Systems Biology Markup Language SBML.

Databases: How to store and access data? How to annotate biological function data? **Reactome!!!!**

Statistics: Fisher exact test????

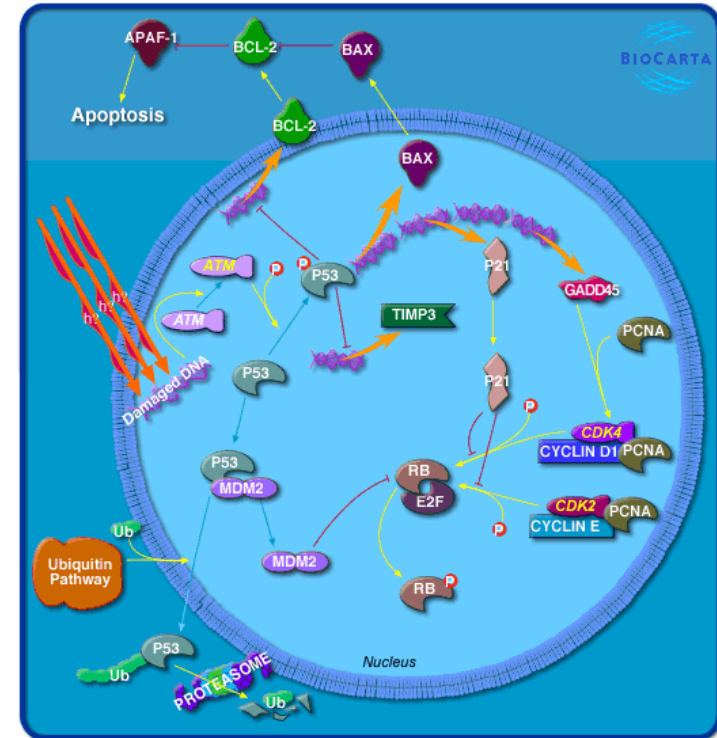
Visualization: How to display biological function?????
KEGG Maps or Biocarta.

For now we are working on the annotation.

Pathway
Databases

Metabolic
Protein Interaction
Signal Transduction
Gene Regulatory

WIT
BioCyc
Reactome
aMAZE
KEGG
BIND
DIP
HPRD
MINT
IntAct
PSI format
CSNDB
TRANSPATH
TRANSFAC
PubGene
GeneWays



Universal Pathway database

Visualizing expression on Pathways

Cytoscape/pathway miner are useful tools for the integration of genomic data onto gene networks.

MetaCyC and EcoCyC display expression data of some experiments on their maps.

Genemapp, UCSF, map genomics data on metabolic or cellular charts. Proposing a new MAPP format, MAPPs.

Gene-Go, integrates expression data with data from different sources on pathways.

Kegg also allows to map data on metabolic MAPS.

Gene-Merge, (Castillo Davis and Hartl 2003) provides statistical rank scores for over-representation of particular functions or categories.

Gostat is another recent program using GO terms.

Gene Ontology

Gene Ontology Consortium - Microsoft Internet Explorer

File Modifica Visualizza Preferiti Strumenti ?

Indietro Cerca Preferiti Multimedia

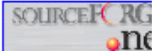
Indirizzo <http://www.geneontology.org/> Vai Collegamenti

GENE ONTOLOGY™ CONSORTIUM

- [What's New?](#)
- [Current Ontologies](#)
- [Current Annotations](#)
- [GO Tools](#)
- [DAG-Edit](#)
- [GO Database](#)
- [Indices of other Classification systems](#)
- [Documentation](#)
- [FTP Archive](#)
- [CVS Repository](#)
- [Software](#)
- [Minutes](#)
- [Global Open Biological Ontologies](#)
- [Publications by GO](#)
- [Publications on the GO](#)

[The GO Consortium](#) [Contact GO](#) [Cite GO](#) [GOBO](#) [Career Opportunities](#) [Acknowledgments](#)

The goal of the Gene Ontology™ Consortium is to produce a dynamic controlled vocabulary that can be applied to organisms even as knowledge of gene and protein roles in cells is accumulating and changing.

To submit specific suggestions for new GO terms to the Consortium, please use the 'Submit New' option on the web form at [SourceForge.net](#). 

Note: Help on the use of the SourceForge GO term submission page is available at [GO curator requests.html](#).

Please send any comments or questions by email to: go@geneontology.org. More details on how to contact GO and how to follow the progress of the project are available from the [GO Contacts](#) page.

Search our site via Google:

Search for GO terms and associated genes:

Select a Browser:

☒  ☐  ☐  ☐ 

Internet

LEU4

GO

metabolism

AA metabolism

leucine
metabolism

synthase



Is like an onion

mitochondrial
metabolism

Analyze expression profiles with Pathway Processor



Pathway Analyzer

[http://www.cgr.harvard.edu/
cavalieri/pp.html](http://www.cgr.harvard.edu/cavalieri/pp.html)

Paul Grosu
CGR Computational
Biology Group

Uses only 96 well established pathways in a well established model microorganism, yeast

Scores biochemical pathways according to the probability that as many or more genes in a pathway would be significantly altered in a given experiment by chance alone.

Use *P*-values to select pathways to examine more closely using Expression Mapper

The Fischer Exact Test measures the probability that a pathway is significantly enriched, given a threshold value. ¶

The general formula used is: ¶

¶

$$\text{Probability} = \frac{(\text{number of successes}) \times (\text{number of failures})}{(\text{number of anything happening})} ¶$$

Here we used the hypergeometric part of the Colt software program. We used two of the following two formulas: ¶

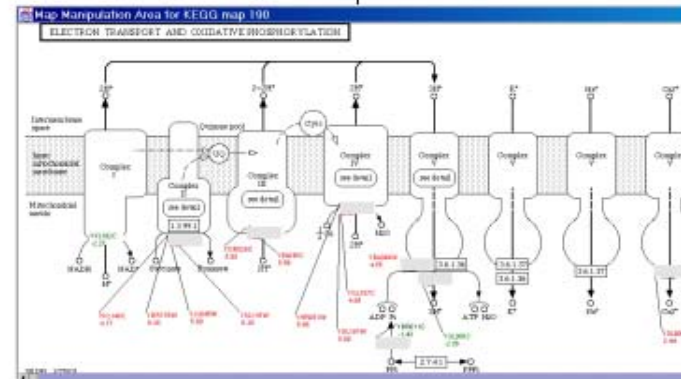
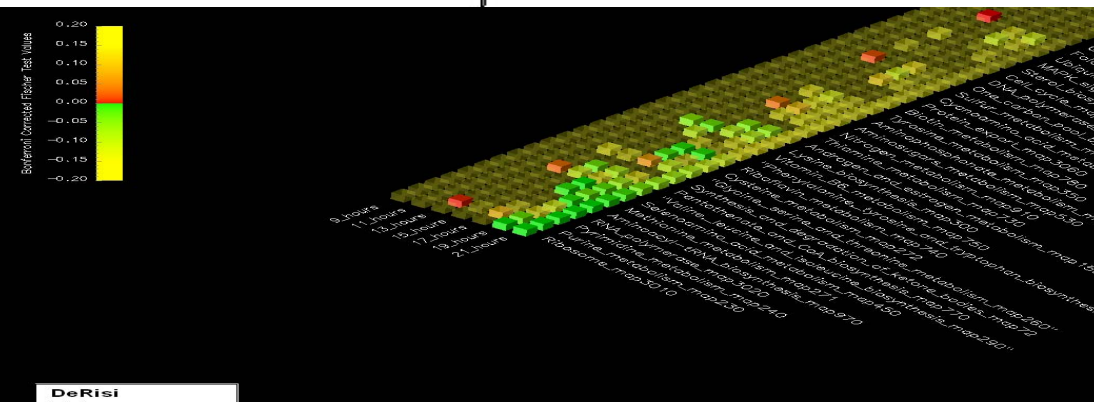
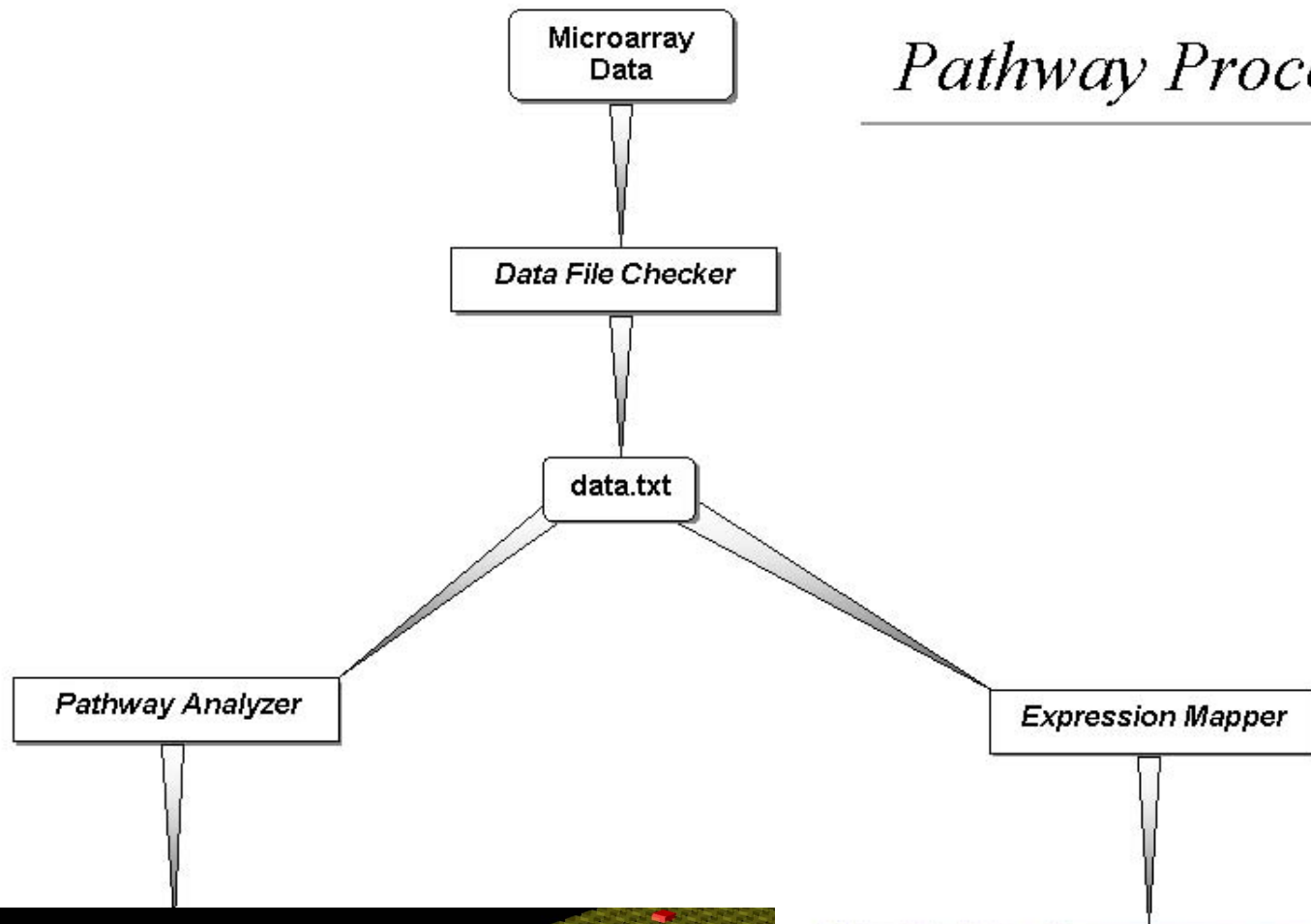
¶

$$P(k) = \frac{\text{Probability of } k \text{ ORFs passing cutoff in pathway}}{\text{ORFs in pathway}} = \frac{\binom{\text{ORFs passing cutoff in genome}}{k} \binom{\text{ORFs in genome} - k}{\text{ORFs in pathway} - k}}{\binom{\text{ORFs in genome}}{\text{ORFs in pathway}}} ¶$$

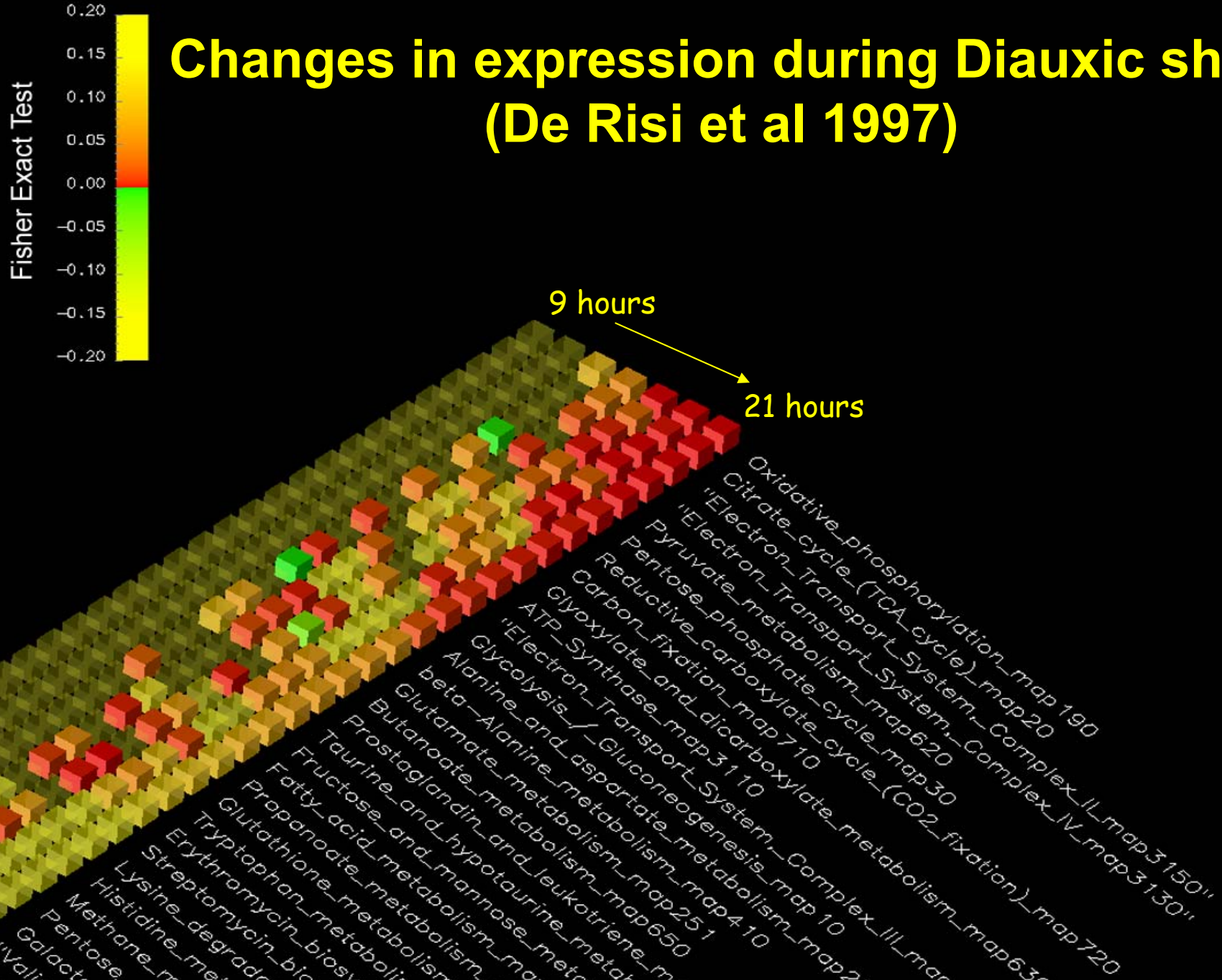
¶

$$\text{Fischer Exact Test for pathway} = \sum_{j=k}^{\text{ORFs in pathway}} P(j) ¶$$

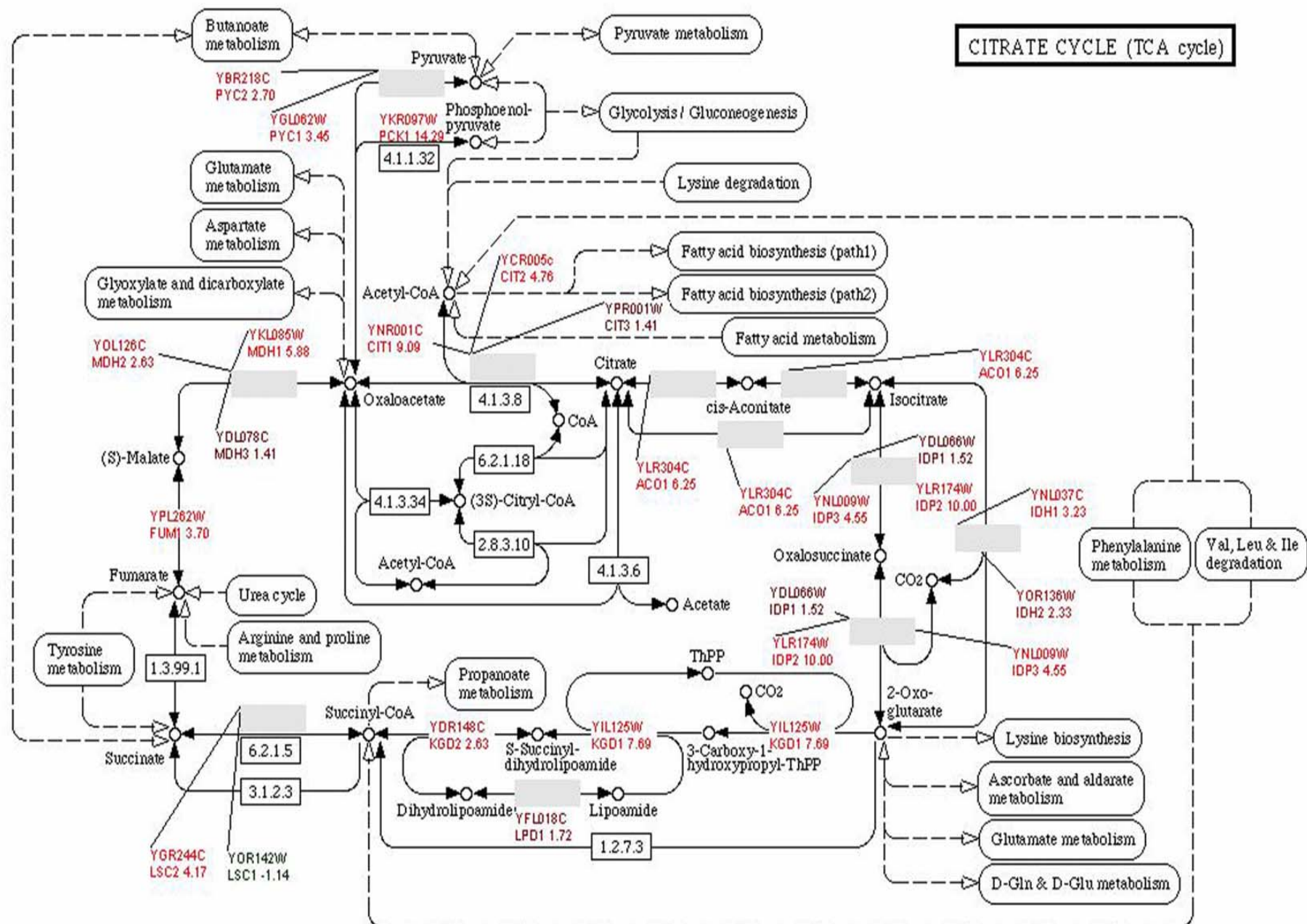
Pathway Processor



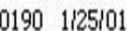
Changes in expression during Diauxic shift (De Risi et al 1997)

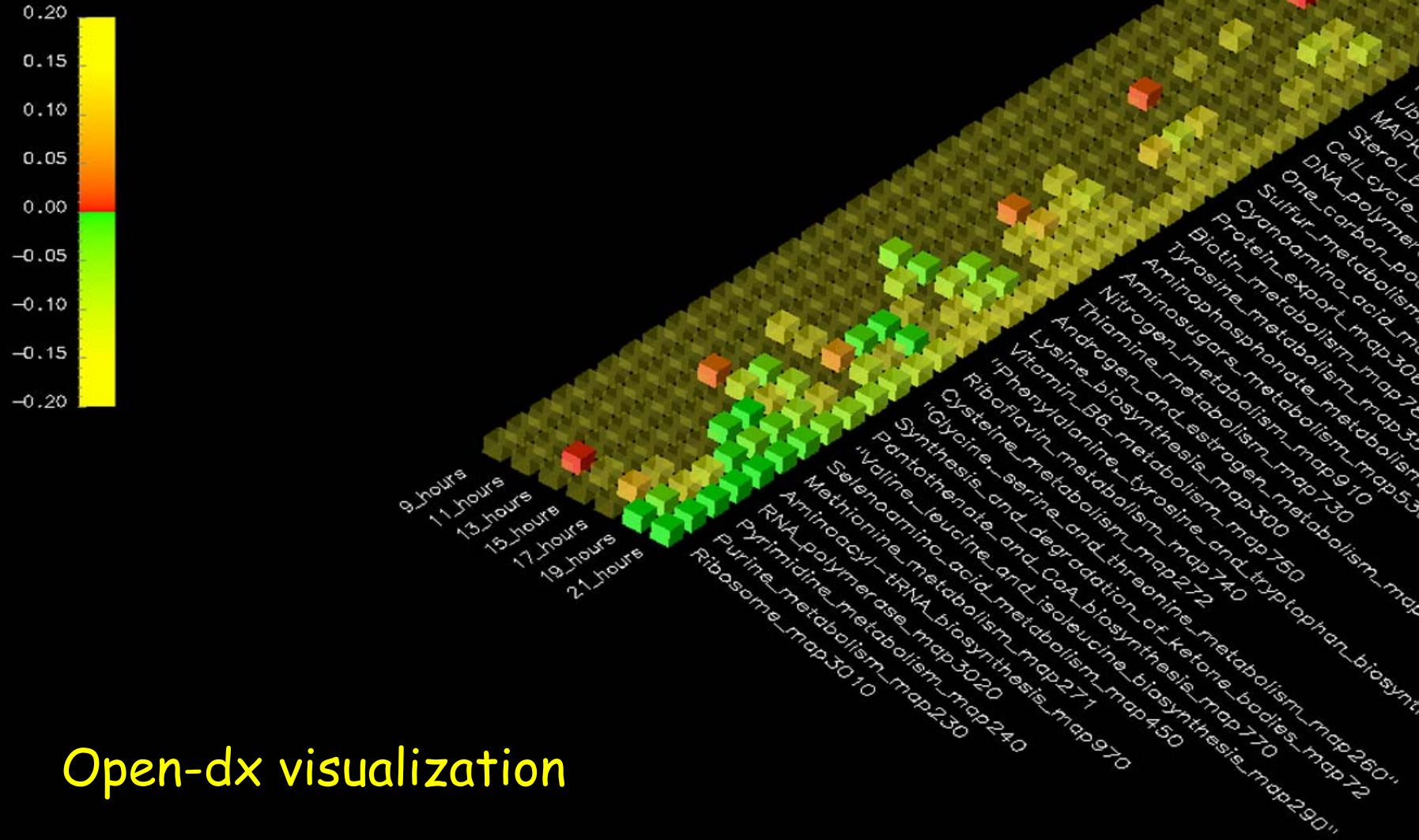


CITRATE CYCLE (TCA cycle)



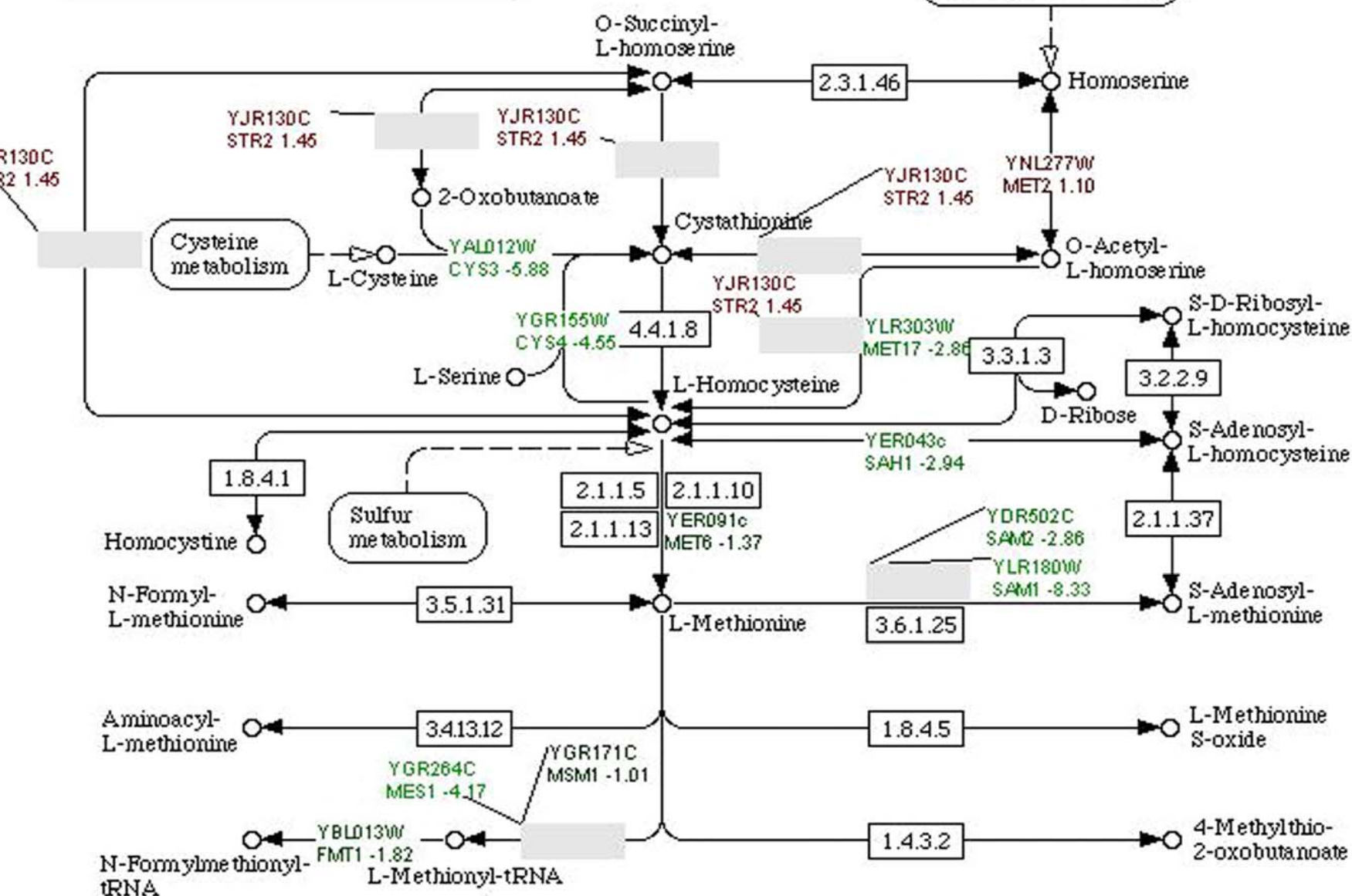
ELECTRON TRANSPORT AND OXIDATIVE PHOSPHORYLATION



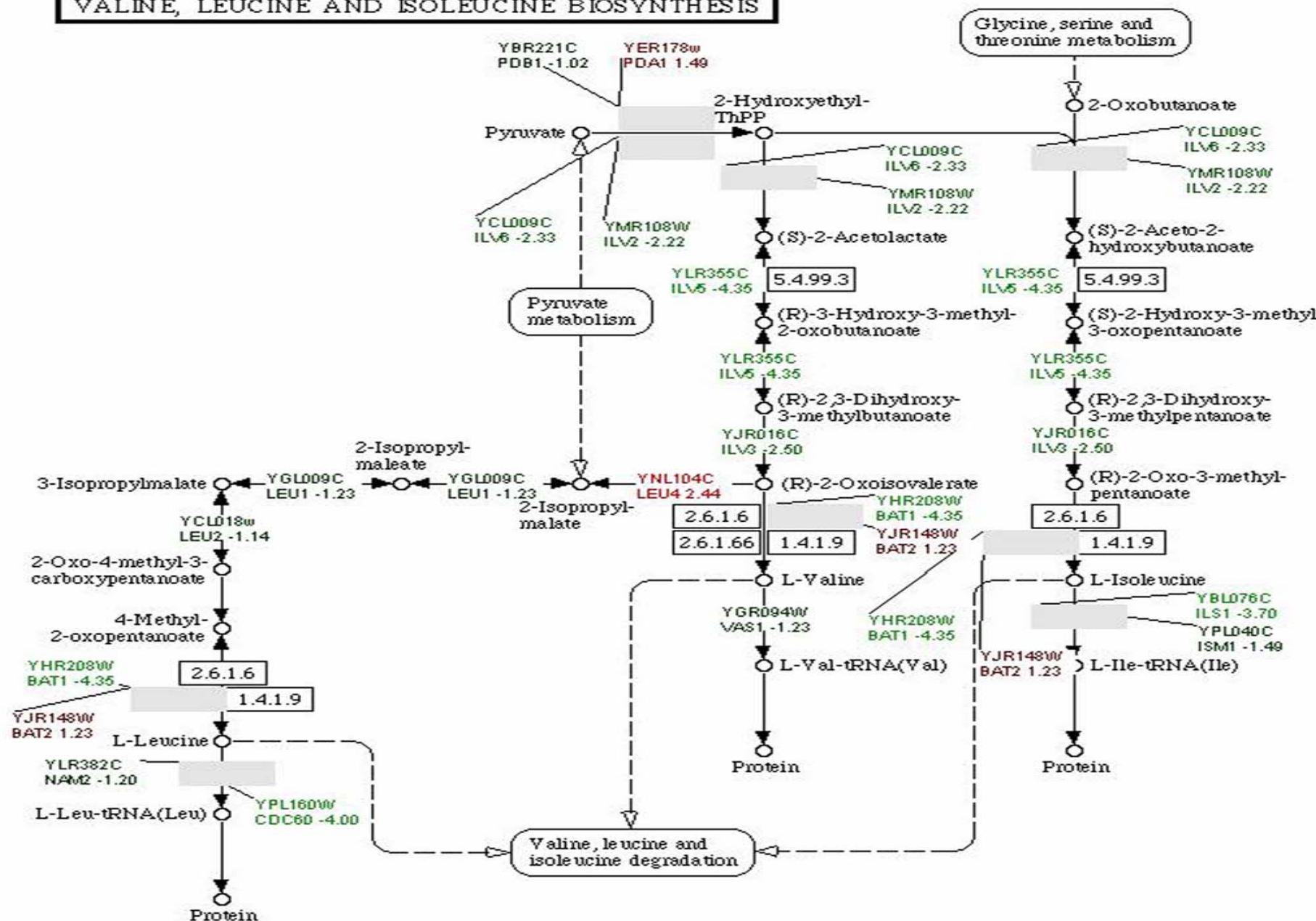


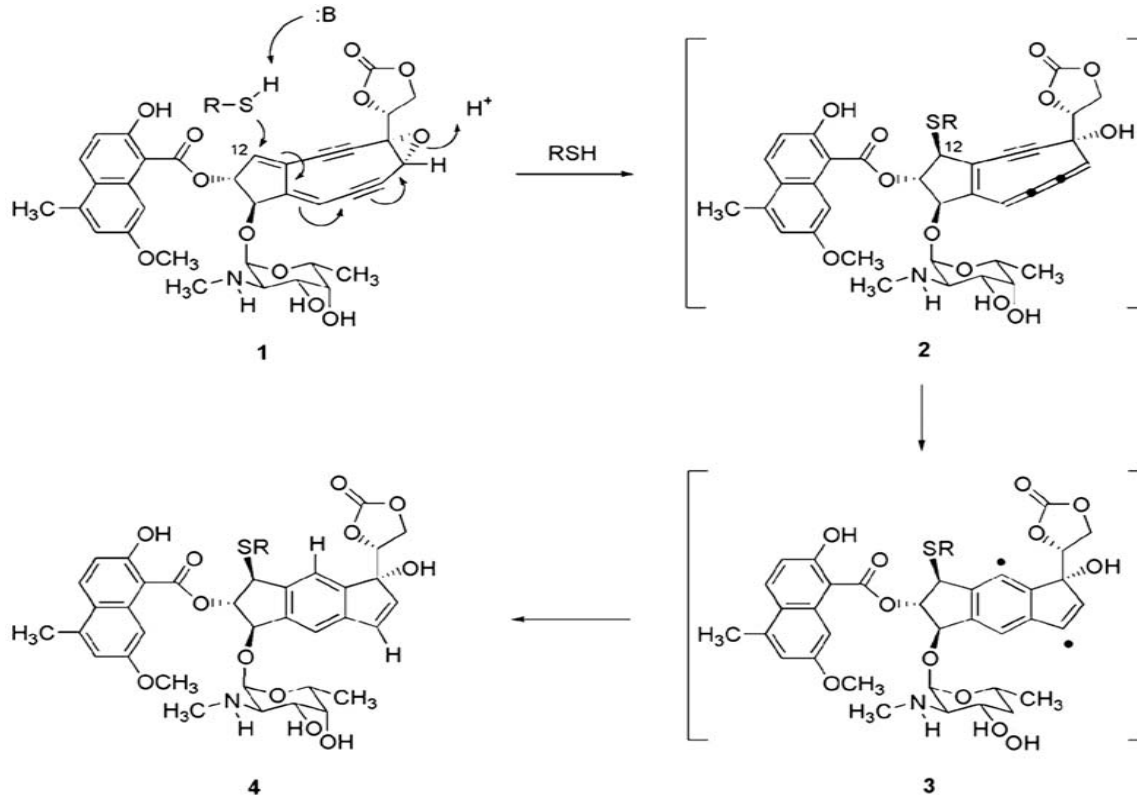
Open-dx visualization

METHIONINE METABOLISM



VALINE, LEUCINE AND ISOLEUCINE BIOSYNTHESIS





Nucleophilic activation of NCS due to thiols (R = CH₂CO₂CH₃) or glutathione (reducing environment).

NCS is a chromoprotein composed of a 11kDa protein (Apo-NCS) complexed with a highly reactive chromophore with antitumoral activity

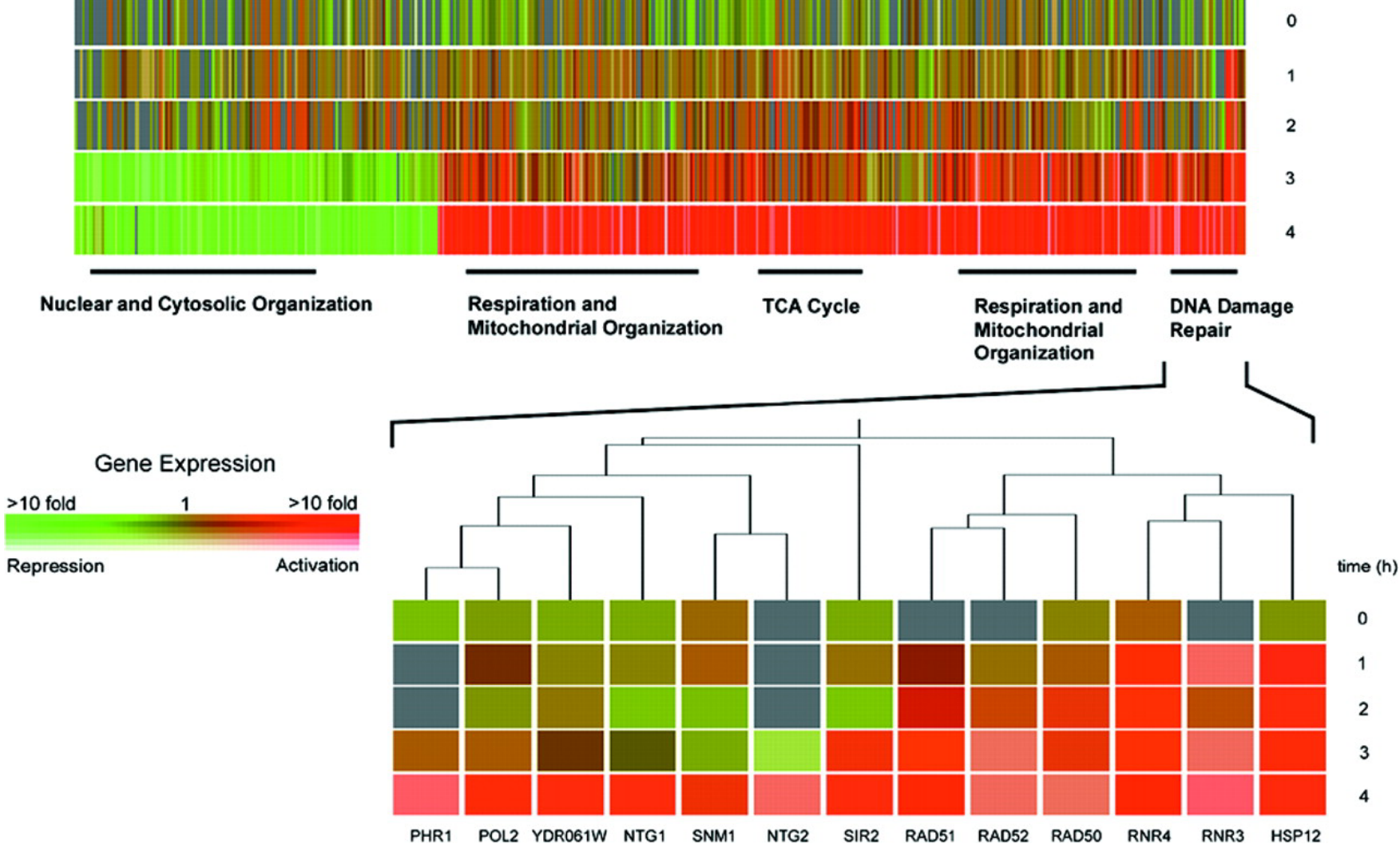
It is used in bladder cancer, breast, liver.

The reducing environment breaks the complex activating the NCS

NCS causes single and double strand break.

Genome wide analysis of the response to a protein-small molecule DNA-damaging agent

- Neocarzinostatin, NCS is a protein-small molecule complex that exhibits potent antiproliferative activity in mammalian cells but has little effect on the growth of the unicellular eukaryotic organism, *Saccharomyces cerevisiae*. The basis of this resistance is not known.
- The determination of the mechanism of action of any small molecule in vivo is challenging under normal circumstances, but is especially complicated when the molecule is highly reactive, such as NCS.



1,288 genes exhibited 2 fold modified mRNA transcript levels in time course of *S. cerevisiae* treated with (50 µg/ml) NCS at 1-, 2-, 3-, and 4-h time points

180 (24%) of the 780 genes overexpressed contain the **stress-responsive element (STRE)** 5' CCCCT in their respective promoters

RECT1 and **RAD9** were overexpressed 2.4- and 2.6-fold

IR2 3,4p complex is proposed to undergo Rad9p and Mec1p dependent relocation from silenced chromatin to sites of DNA damage, where it interacts with the Ku heterodimeric complex.

SP12, was overexpressed >2-fold after 1 h.

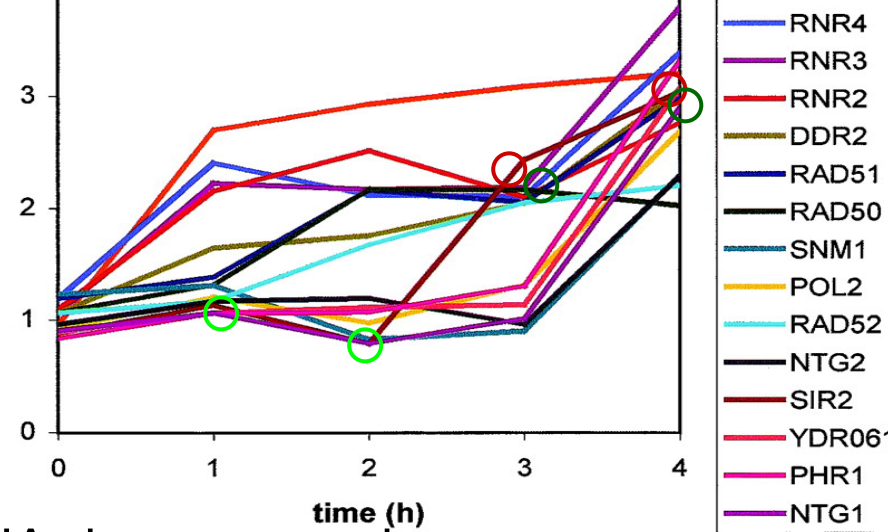
DR2 and **DDR48**, DNA damage stress response genes were not overexpressed until 3 h.

Glutathione S-transferase

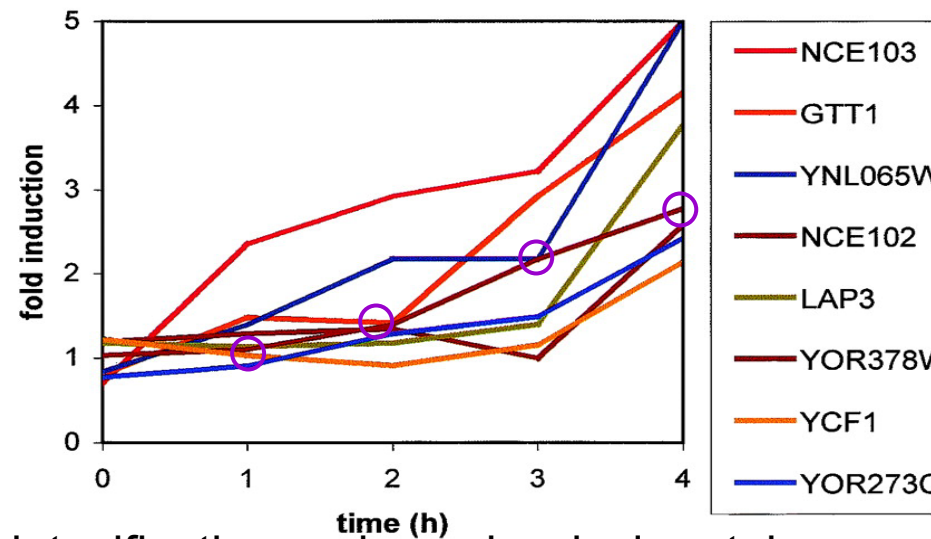
GTT1, was + 4.1-fold at 4 h.

CE1, homologous to the human **hMBP1**, was +2.1 fold at 4h

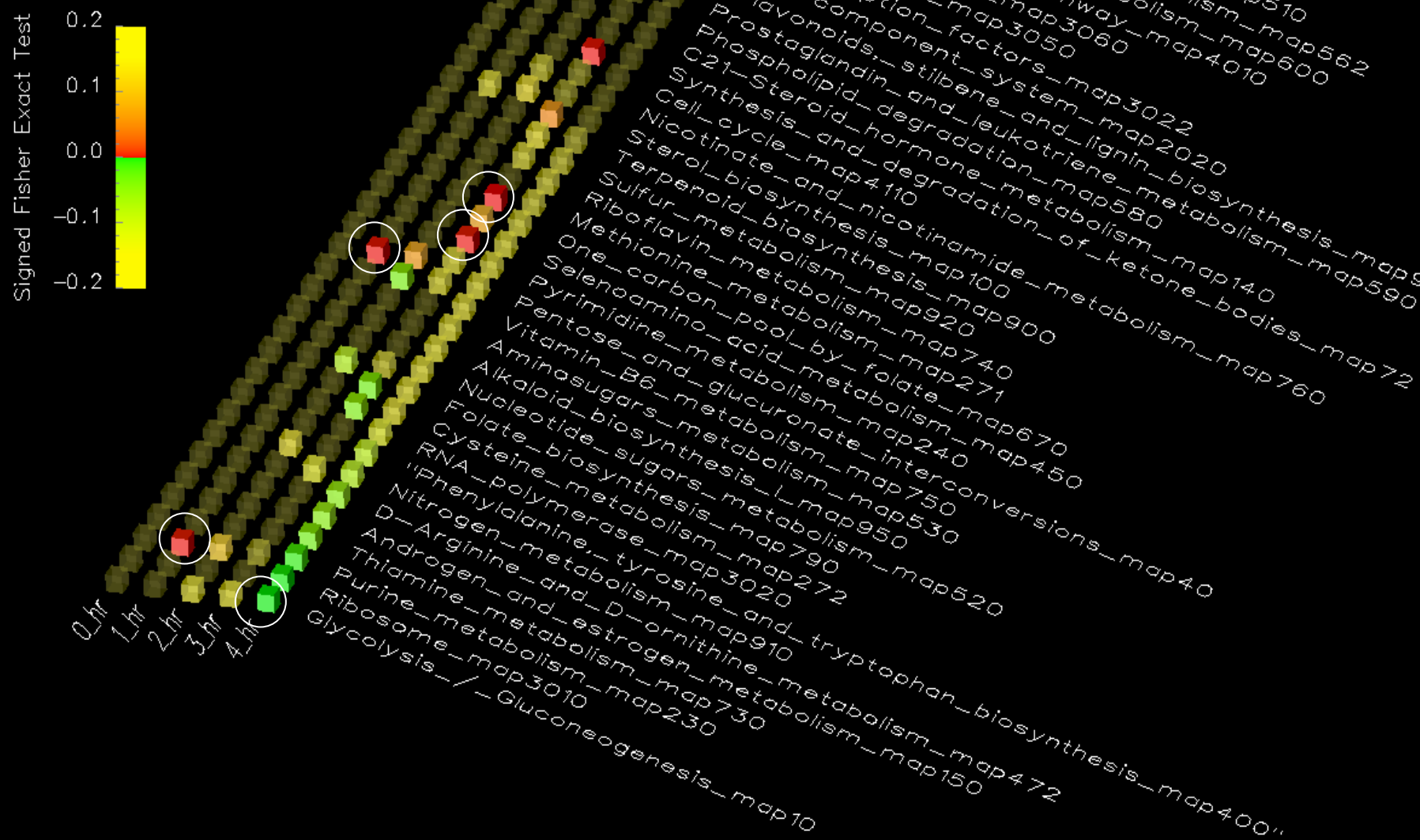
DR proteins **YOR273C**, **YOR378W**, and **YNL065W** showed a time-dependent increase in expression levels over the course of the treatment.



A DNA-damage repair

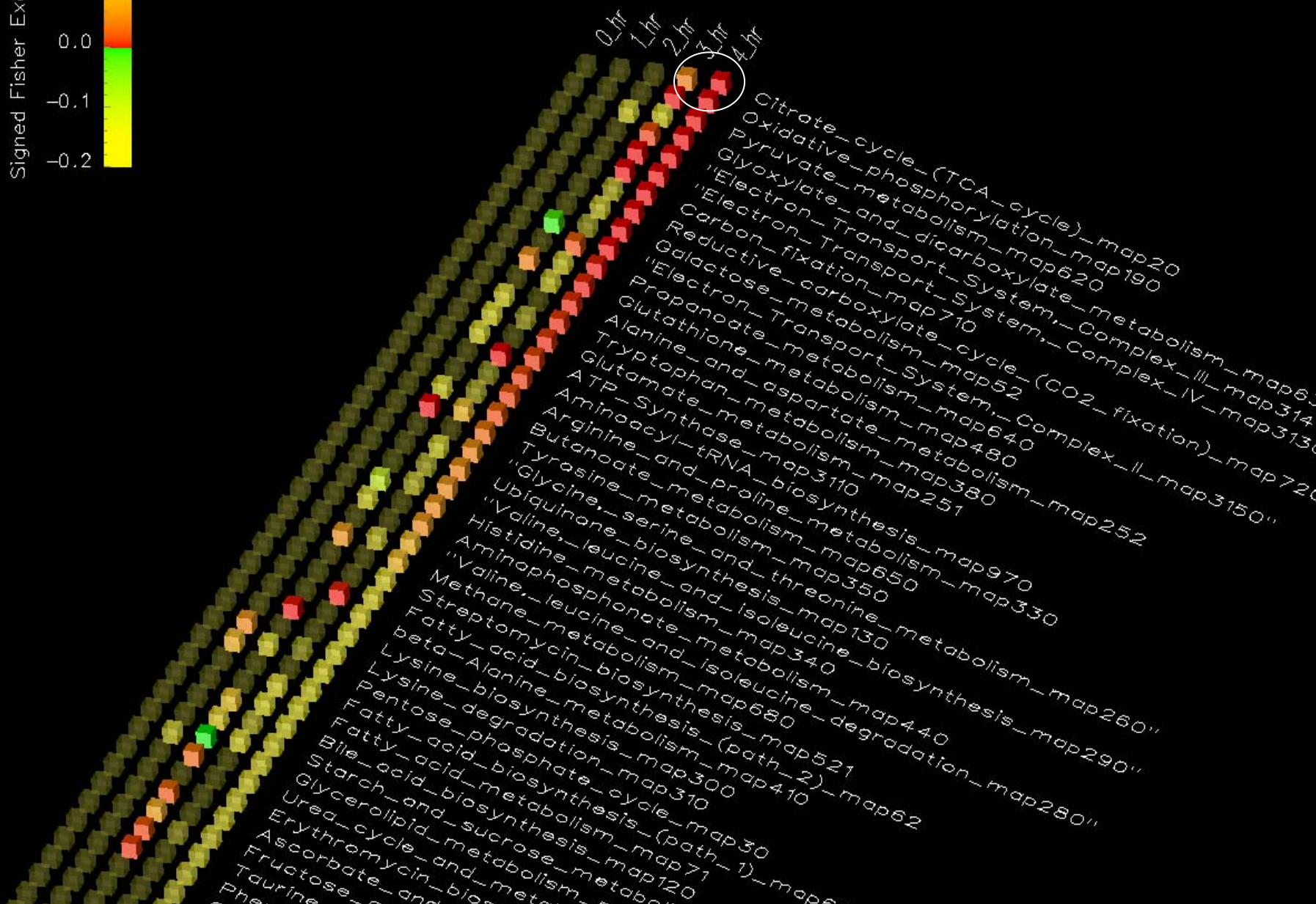


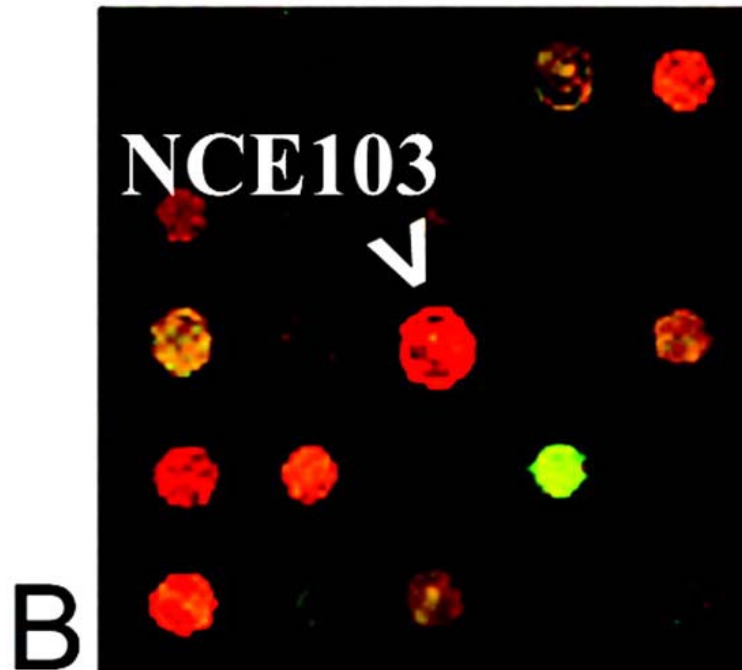
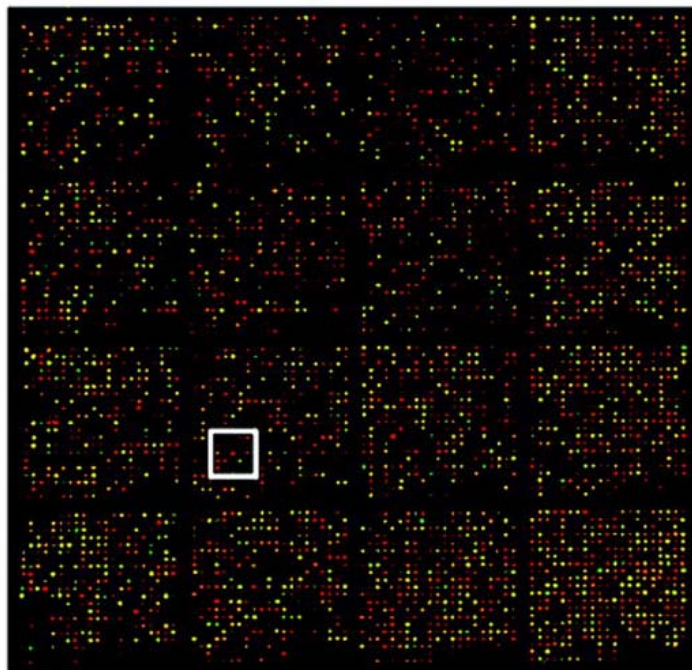
B detoxification and nonclassical protein export



- 1h-** Activation of **nucleotides** metabolism (DNA Repair) repressed at 4 h.
- 3h-** Activation of **methionine e selenoaminoacid** metabolism (detoxification).
- 4h-** Inactivation of **Glicolysis** and **activation of respiratory** metabolism.

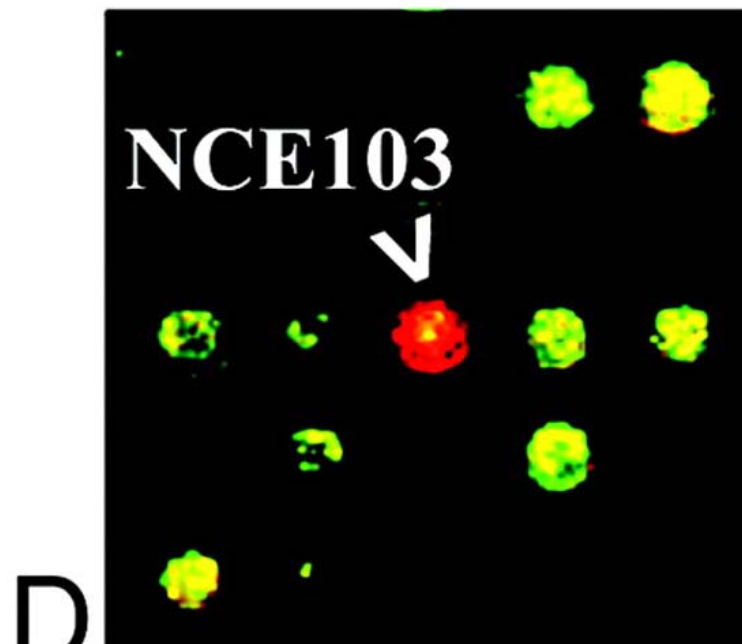
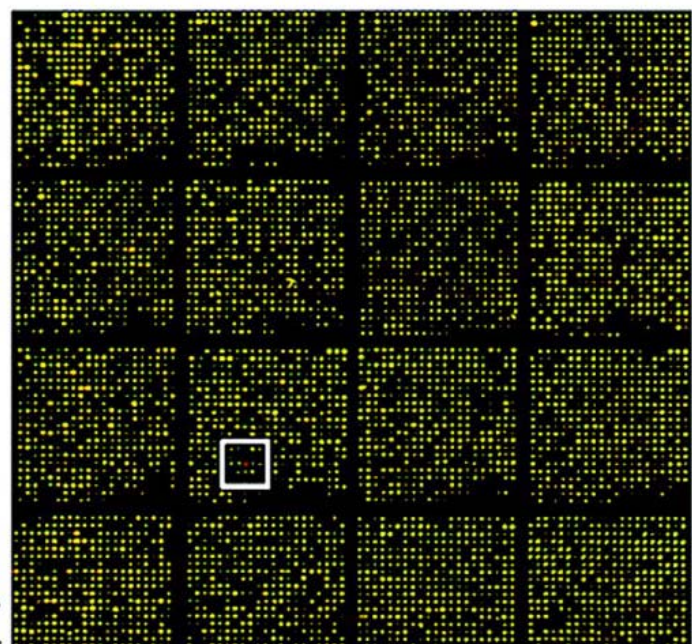
4h- Activation of TCA cycle and mitochondrial metabolism.





A, B Treatment
with NCS-
Protein
complex

Hundreds of
genes are
activated



C, D
Treatment
with apo-
protein

NCE103, NCS
specific
transporter is
the only gene
that is activated

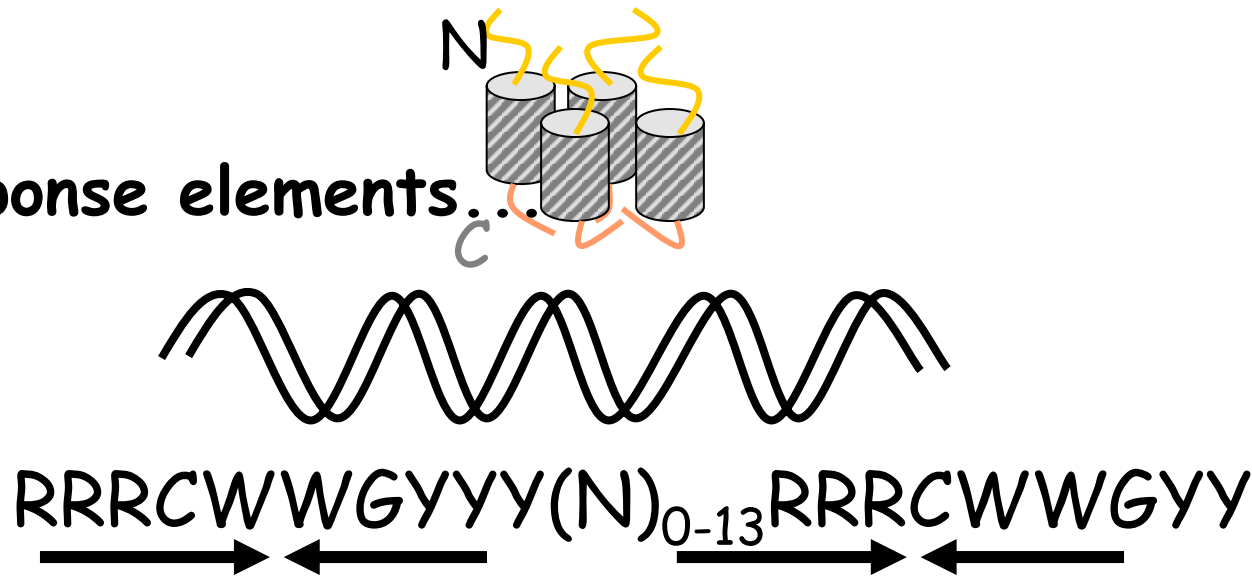
Motif	Cluster	Z-Score	PValue	Hits	#Genes	%	Genome Hits	Genome Genes	%
STRE_B	946	4.018548	1.65E-23	104	125	83.2	294	669	43.95
HAP234_01	1297	3.706969	3.82E-14	18	23	78.27	77	669	11.51
ROX1_Q6	1256	3.70189	0.009467	9	22	40.91	127	669	18.99
HAP1_B	1355	3.70045	0.00142	51	193	26.43	130	669	19.44
PHO4_01	999	3.700627	2.89E-08	11	14	78.58	88	669	13.16
MIG1_01	909	3.706847	5.17E-04	21	63	33.34	115	669	17.19
STE12_Q4	1062	3.948236	4.81E-04	32	83	38.56	158	669	23.62
LEU3_B	818	3.701191	6.04E-04	7	7	100	234	669	34.98
ABF1_01	1060	3.702356	6.09E-04	23	80	28.75	104	669	15.55

Tests for enrichment of functional categories and for transcription factors binding sites
reconstructs regulatory networks.

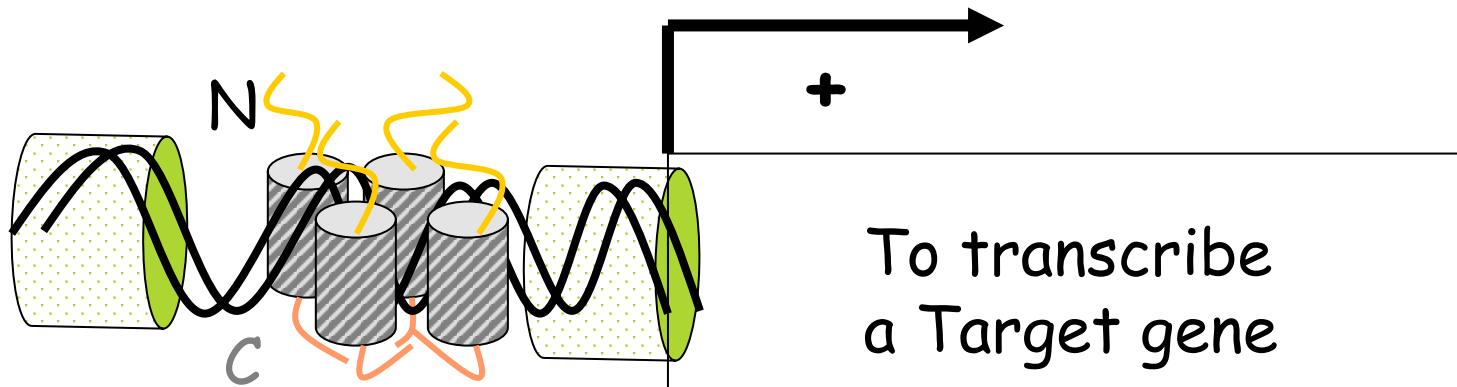
Enrichment for Tf's involved in respiratory metabolism and stress response.
Hap1-Hap4

Transcription factors bind...

...to different response elements...



And act together with co-activators/adaptors/co-repressors:
ADA3, ASPP, p53-BP1; p53-BP2, p33ING1, WRN, BRCA1, TFIID,
TFIIH, SIR2, CBP/p300, MDM2, MDMX





Interaction of Genes with REPCAR

endomembrane system	L442	
---------------------	------	--

TTCTATTTTATTTTTCTACAA
TTTTCAATTGCGGTAGCAATC
GATGCTTGGGTTATTTTAAAT
TTGTGTAGCCTCCATCGTTTCG
GAATCAGCTATCCTTGTCCTTG
AAGCGGTATCCTGTTTGGATT
GGCTTTTGGCCCGTGCTCTGAG
TCCTTTTAAGGTAAC TGGATTA
ATTGTTTACTTGAATGTTATT
AAATAATAAACCAGCCACACTG
GGTGAAAGCAGAAACCCTTT

TOS2 - Hypothetical ORF
RVB1 - Putative 3' to 5' UTR
ORC1 - Subunit of the origin recognition complex
SPT16 - Essential nucleosome assembly factor
IAH1 - isoamyl acetate hydrolase
YMR254C - Unknown function
YIL141W - Unknown function
DUO1 - Protein that forms a complex with the DNA replication fork
YOL150C - Unknown function
AXL2 - Membrane glycoprotein
YIL195C - Unknown function

1.20E-03	4	39	10.26	11	795
1.20E-03	6	12	50.00	96	795
1.32E-03	4	40	10.00	11	795
1.34E-03	19	30	63.34	680	795
1.44E-03	4	41	9.76	11	795

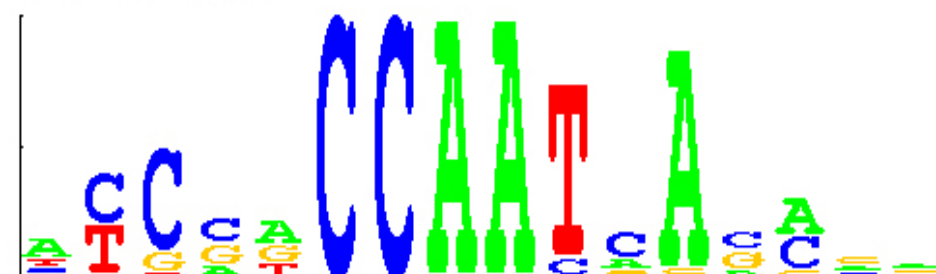
NCS treatment in aerobically and
anaerobically growing cells
treated with 50 and 100 $\mu\text{g/ml}$ NCS
Anaerobic-----Aerobic

**30% of the genes varying
in expression contains
Hap1-Hap4-Rox1
signatures in their
promoters**

ran Segal, Aviv Regev, <http://genexpress.stanford.edu/>



Motif 1: HAP234_01



Pathway processor enables to interpret gene expression of drug treatment in the context of cell biology

- **DNA** Repair one of the early responses to NCS treatment.
- **Switch from anaerobic to aerobic metabolism** might represent a long term adaptive response to NCS. Respiration causes a change in pH and a passage from a reductive to an oxidative environment and as a consequence a small number of free thiolic groups and is likely to prevent the nucleophilic activation of NCS.
- **Active transport of NCS** outside the cell.

The future of the Pathway processor Project

- Up-load the gene lists from a library of pathways, including **human**, **mouse** and **rat**.
- Map gene expression data on **your favorite maps** annotating the pathways with a universal standard.
- Calculate **connectivity** between Pathways
How frequently the genes in a given pathway are represented in other pathways is an estimate of the degree of connection between the different pathways.
Flux balance of metabolites can be used to estimate interdependence of pathways???

Statistics

- What is the statistics most useful for multiple hypothesis testing if the pathways are clearly interconnected.
- Is it correct to analyze pathways that are not properly annotated or defined???
- Is it correct to analyze entities that are partially or totally overlapping???
- Pathway databases should contain information on pathway architecture, hierarchy.

WE NEED GOOD STATISTICS BECAUSE

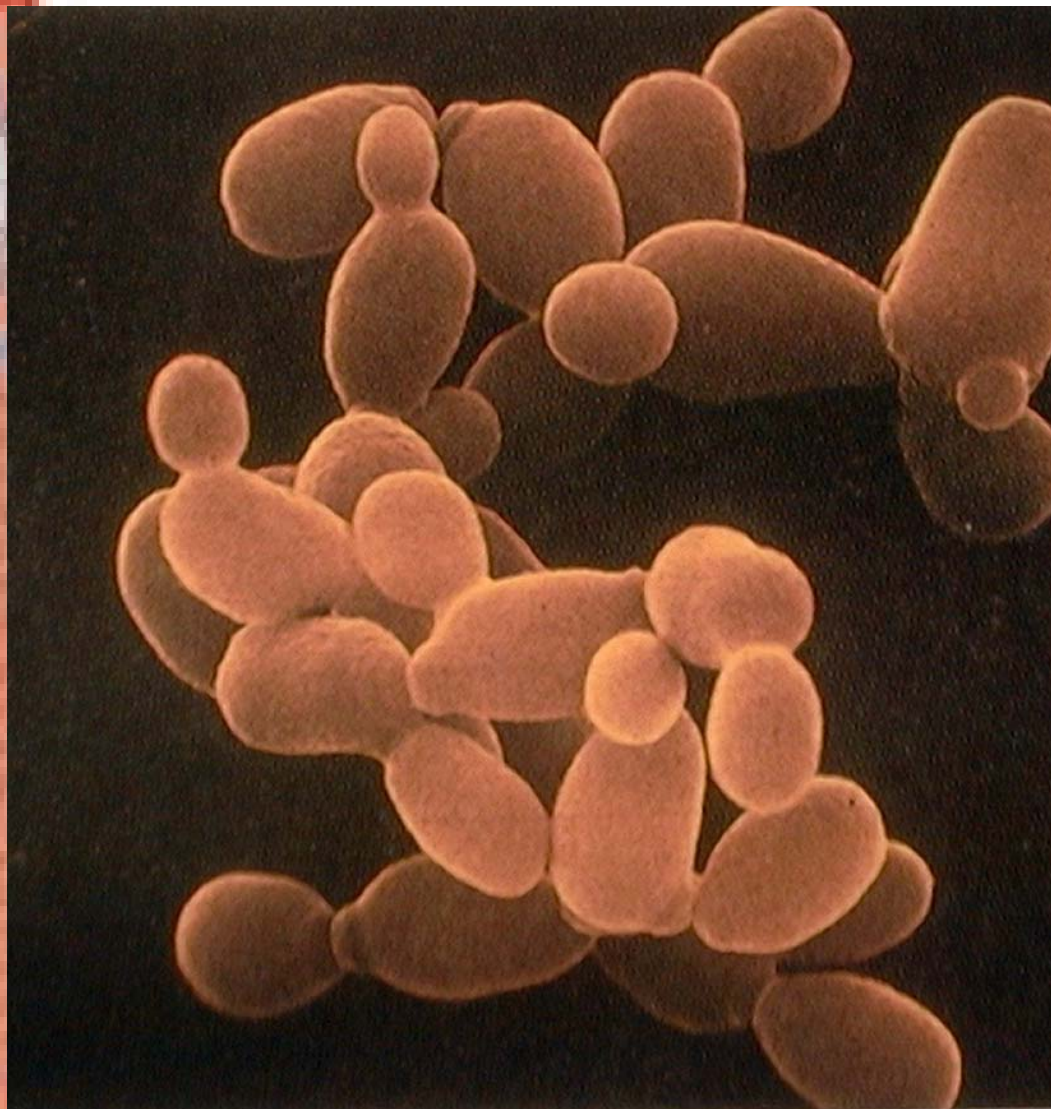
Magical induction won't turn the data flood into knowledge





QUICK YEAST

For Bread Machines
& Baking
- Just Add To Flour -







Saccharomyces Genome Database

Quick Search:

[Site Map](#) | [Help](#) | [Full Search](#) | [Home](#)

[Community Info](#)

[Submit Data](#)

[BLAST](#)

[Primers](#)

[PatMatch](#)

[Gene/Seq Resources](#)

[Virtual Library](#)

[Contact SGD](#)

Search Options

[Search SGD web pages](#), [Global Gene Hunter](#),
[Search Literature](#), [and more](#).

Help Resources

[Getting Started](#), [Sitemap](#), [FAQ](#), [and more](#).

Analysis & Tools

[BLAST](#), [Gene/Seq Resources](#), [Maps](#),
[and more](#).

Homology & Comparisons

[DB Homologs](#), [Protein Motifs](#), [Homologs](#),
[and more](#).

Function & Expression

[Protein Info](#), [Function Junction](#),
[Expression Connection](#), [and more](#).

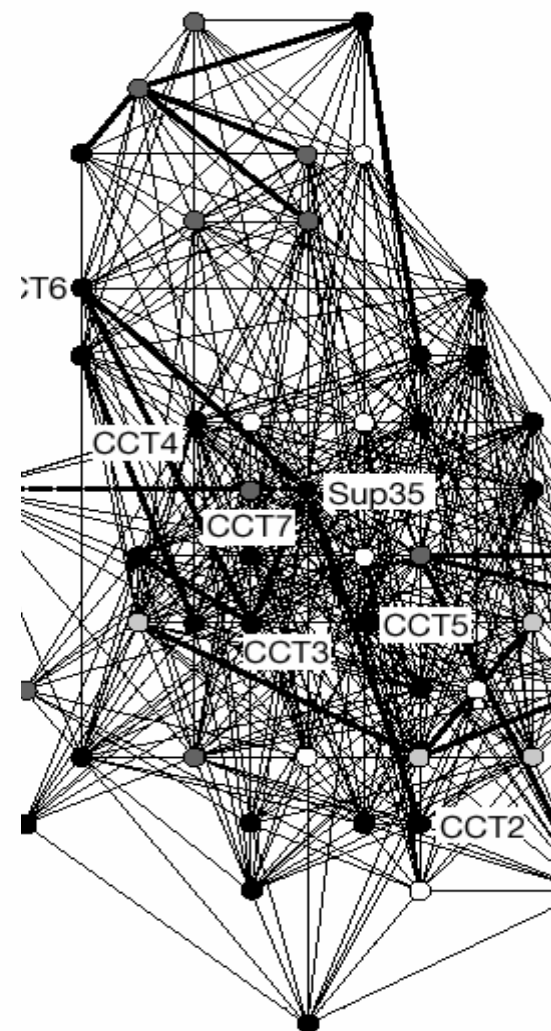
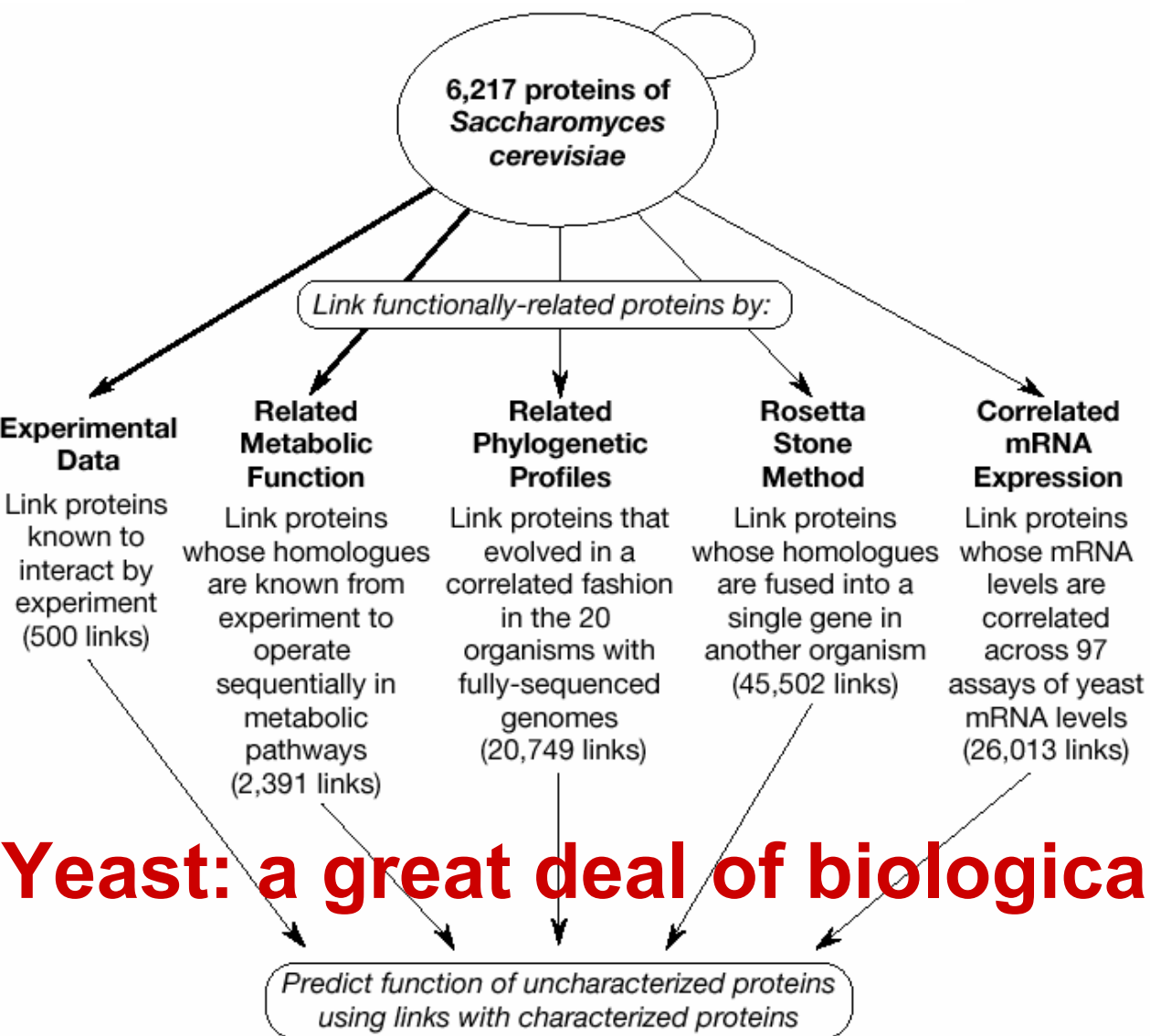
SGD™ is a scientific database of the molecular biology and genetics of the yeast *Saccharomyces cerevisiae*, which is commonly known as baker's or budding yeast.

New and Noteworthy

• New Expression Connection at SGD

SGD has just released a new version of [Expression Connection](#). The new version allows more flexible searches of microarray expression data, presents tabulated results in succinct form, and displays new kinds of information including the Pearson Correlation between queried genes and a line graph of the expression of all queried genes. In addition, a new "Stats" page provides a global view of the data in each dataset by indicating the number of genes whose expression is increased, decreased, and etc. To help users determine commonalities among gene clusters in Expression Connection results, GO Term Mapper and GO Term Finder tools are available from every results page. (Posted April 03, 2003)

Yeast: a great deal of biological function

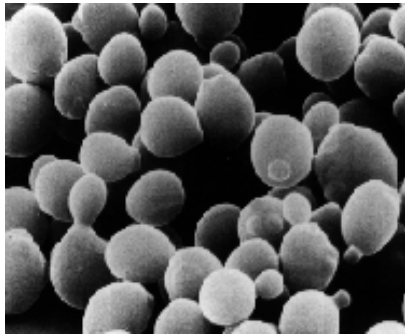


Yeast: a great deal of biological function

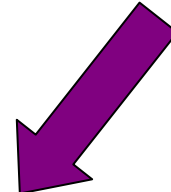
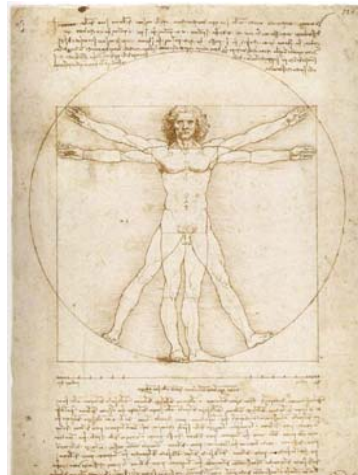
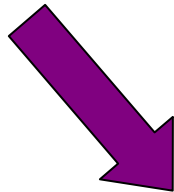


The importance of the models

As in Biology nothing makes sense unless interpreted through evolution, in genomics nothing makes sense unless analyzed in a comparative fashion.

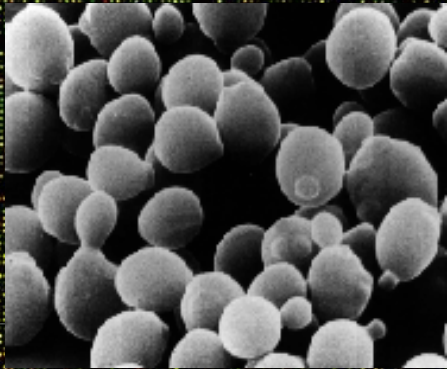


Model systems



From models

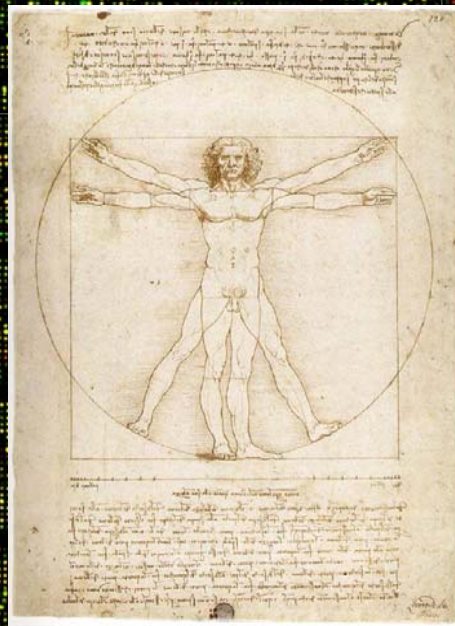
Yeast



Mouse-Rat

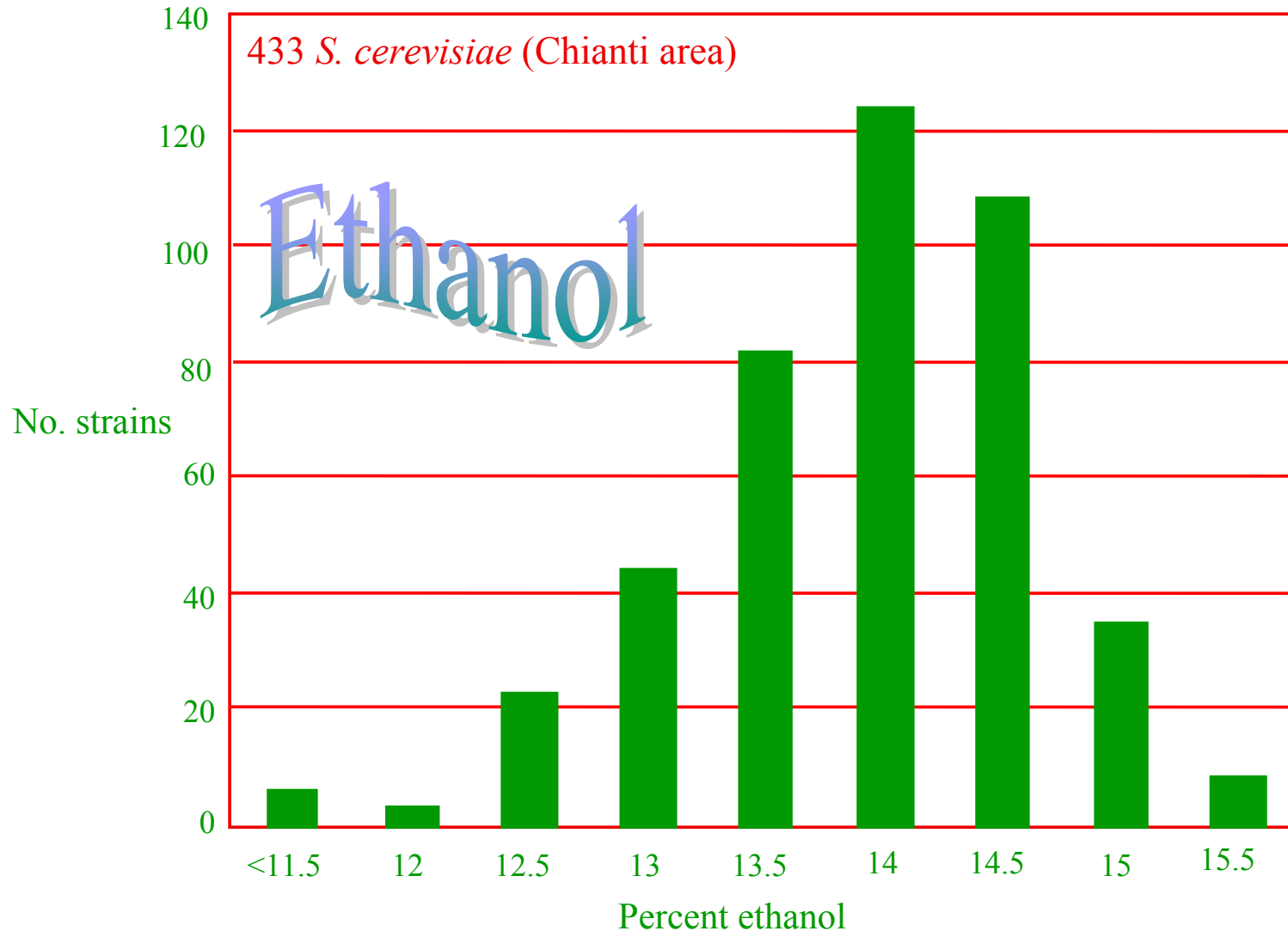


**to
Human**





Phenotypic variation



Phase-contrast (A) and fluorescence (B) microscope images (magnification $\times 1,000$) showing budding yeast after incubation with NCS-FI (NCS bearing a 350-nm fluorescence label). The fluorescence image shows uniform intracellular distribution of the labeled drug, providing evidence that the NCS protein component is internalized in yeast. Fluorescence imaging was conducted with excitation between 340 and 380 nm by using a 400-nm emission filter.

A



B

