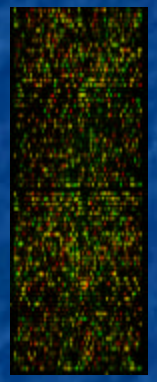
Genes, Circuits and Systems

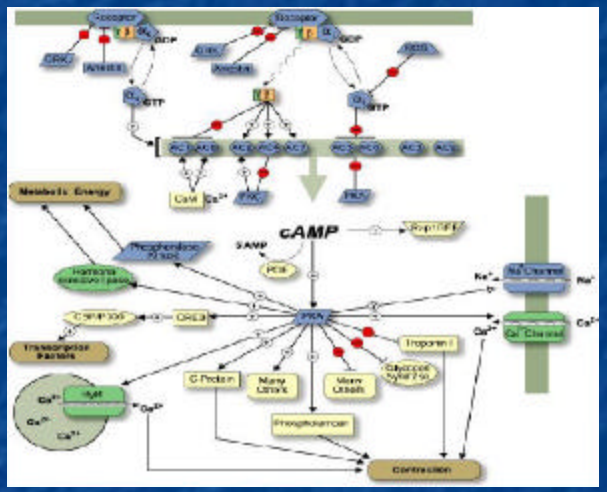
Circuits & System Workshop, 2005

Giovanni De Micheli
Centre Systèmes Intégrés

Gene Expression Analysis

DNA microarray:
Disruptive technology for uncovering gene function on a global scale





What we know:
A (nearly) complete list of the genes

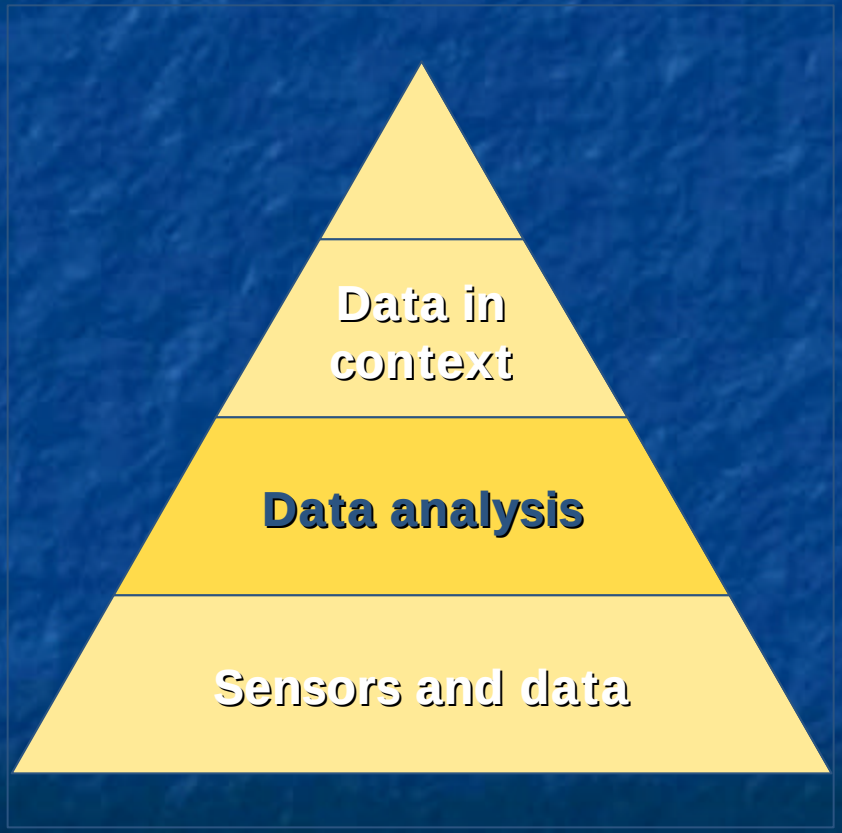
What we don't know:
Gene functions and interactions

The global picture

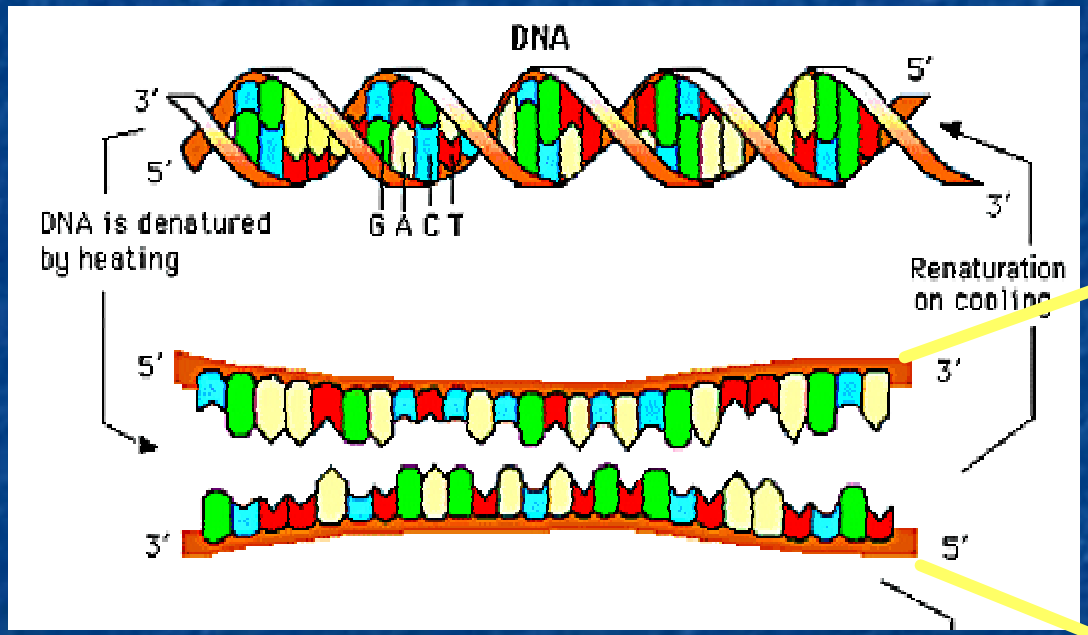
- Measuring biological samples
  - Biosensors --> Circuits
- Analyzing biological information
  - Software tools --> Signal Processing
- Modeling and controlling interactions
  - Laboratory on a chip --> Systems

Bio-discovery uses technologies and tools similar to those used with traditional circuits and systems

Sensors and data



DNA/RNA Hybridization

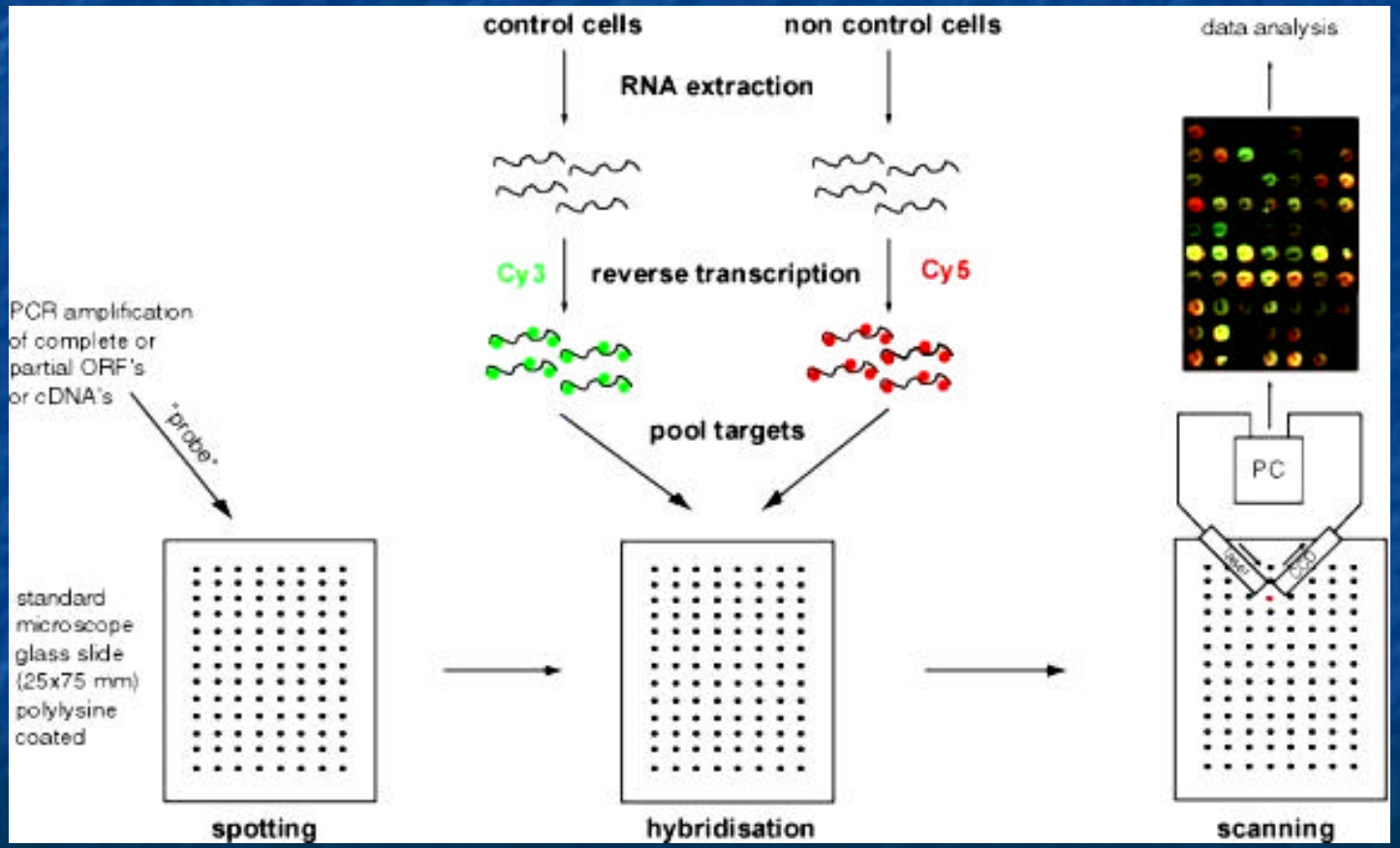


Test (red)

blueprint

The micro-array operation

a b c



Three phases

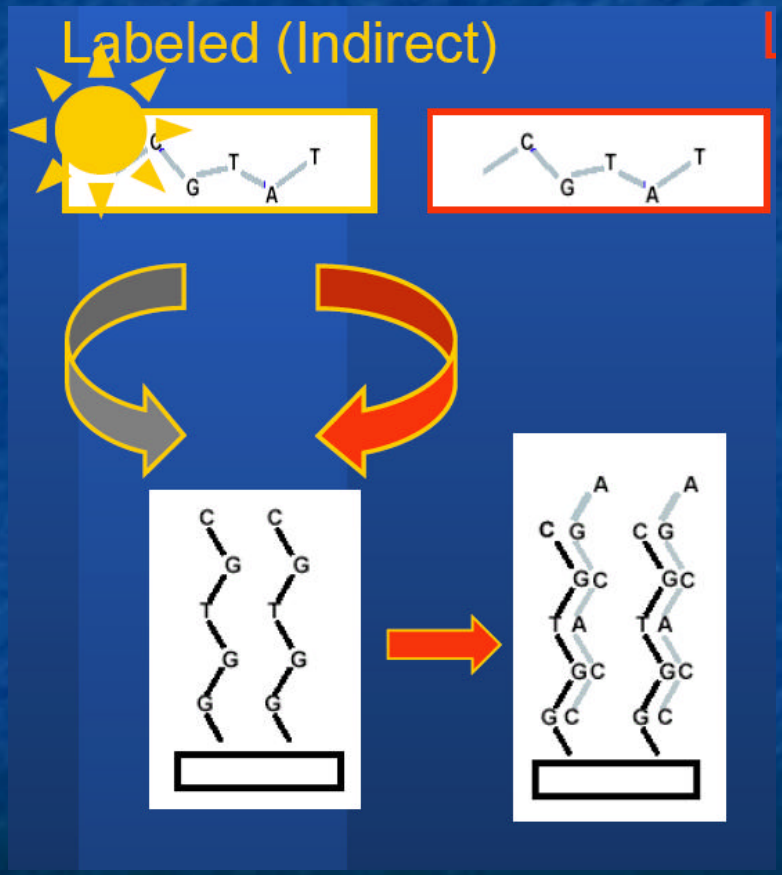
- Micro-array design and fabrication
- Micro-array hybridization and readout
  - After sample preparation
- Micro-array data analysis

- This technologies exists since the mid-nineties
  - Ubiquitous in top-level research
  - Still virtually unused in clinical practice
  - Plenty of opportunities for refinement

Established micro-array fabrication technologies

- DNA spotting on glass substrates
  - Relatively inexpensive and reproducible
  - Pat Brown (Stanford)
- Mask-based (semiconductor-like) fabrication
  - Oligonucleotides (25mers)
  - High-density, high cost
  - Affymetrix (Mountain view)

Micro-array readout

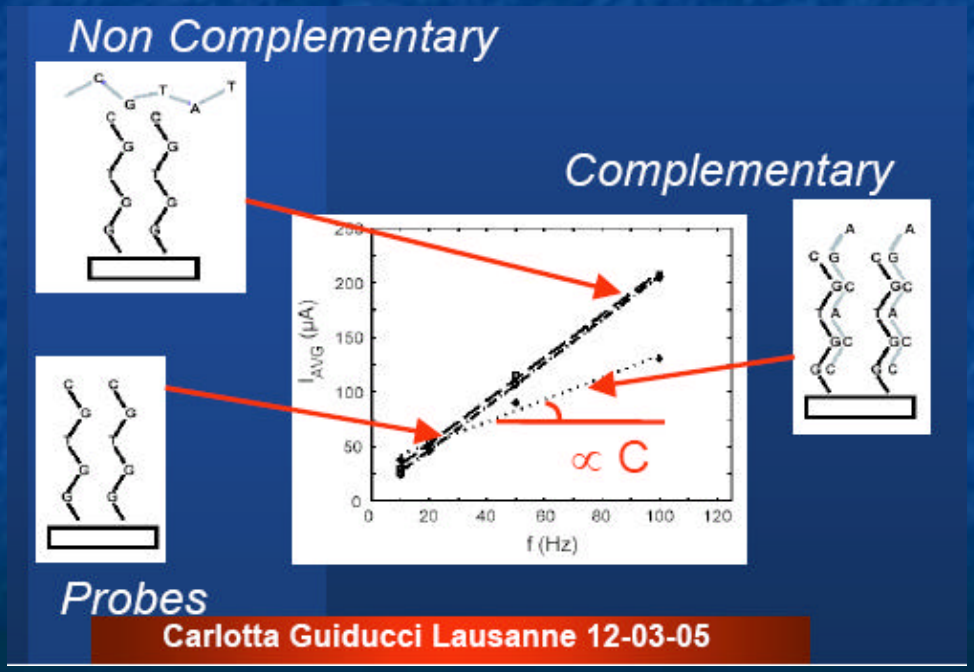


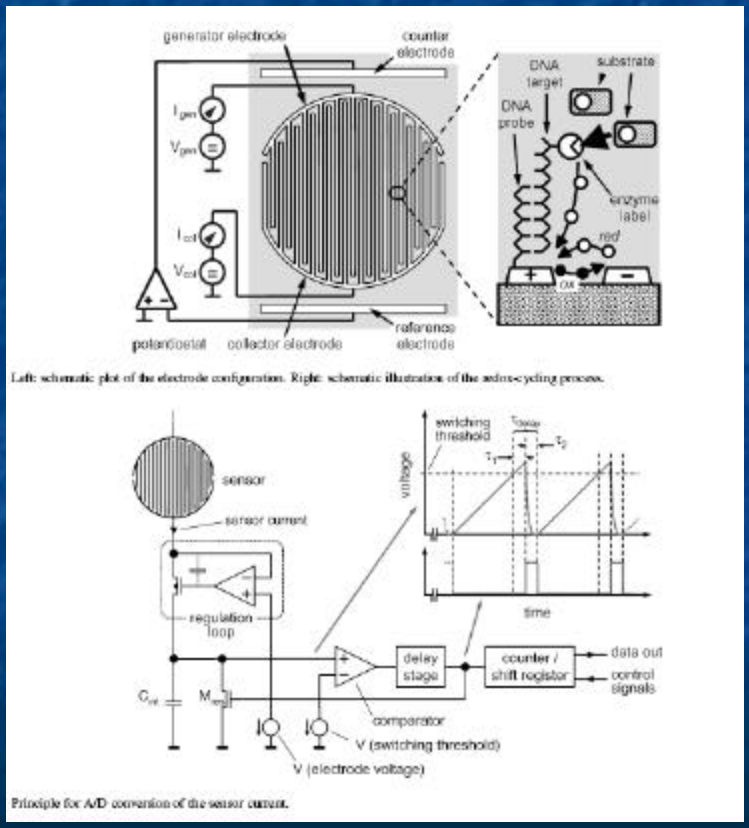
Non-labeled (direct)

- Electrical techniques
  - Changes in the capacitive behavior of an electrode/ solution bio-sensing interface
- Integrated optical sensors
  - Molecular light absorbance
    - Visible, UV
  - Optical arrays

Electrical measurements

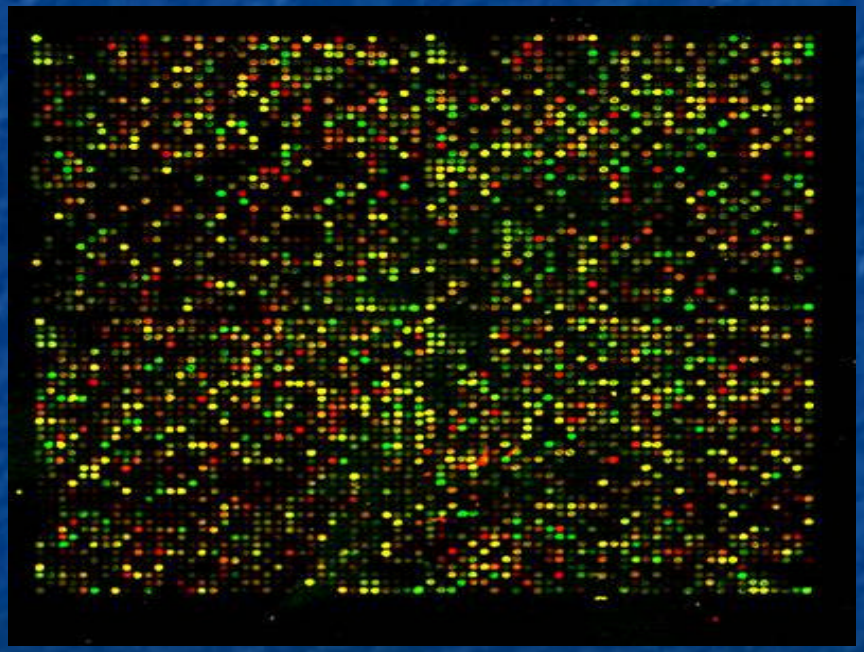
- Dedicated on chip circuitry for capacitance measurements





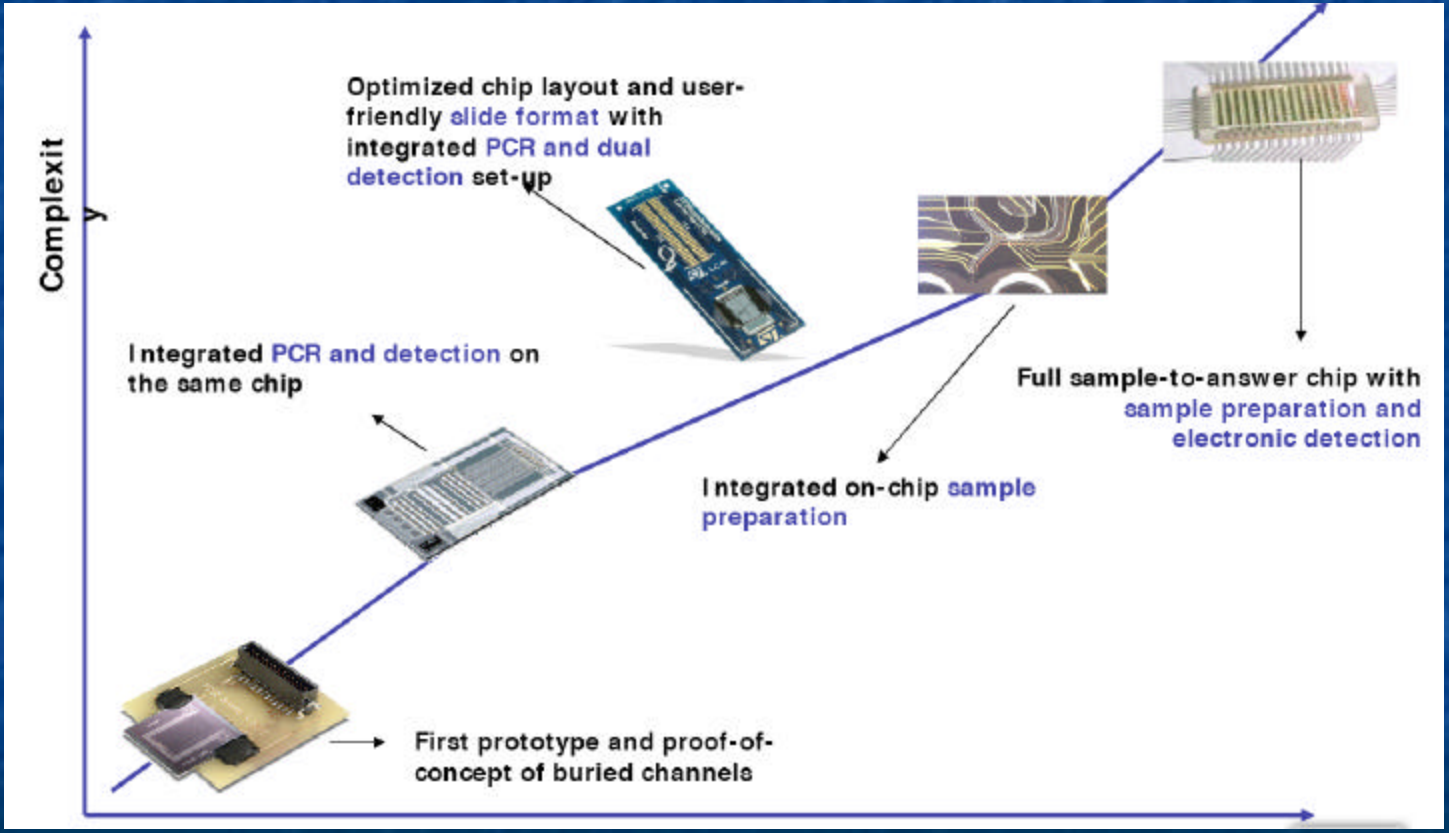
 [Schienle - Infineon 2004]

Optical readout



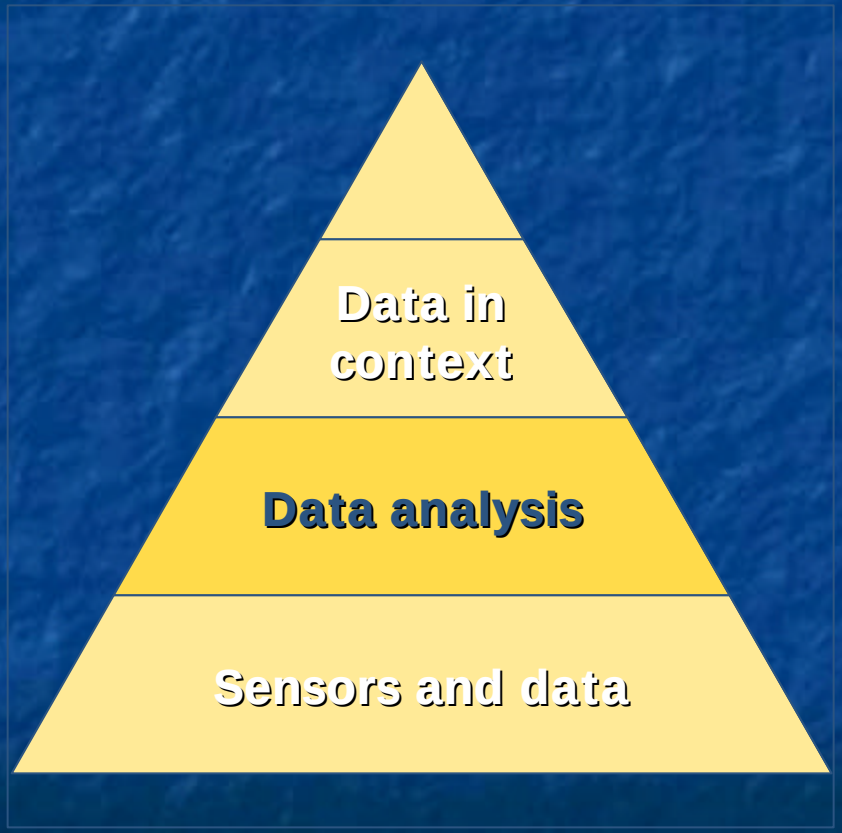
- Laser scanning:
  - Red: sample
  - Green: control
- Measure: $\log\left(\frac{Intensity_{red}}{Intensity_{green}}\right)$
- Data organized in matrices:
  - Rows: genes
  - Column: experiments

Industrial Technology roadmap



[STM]

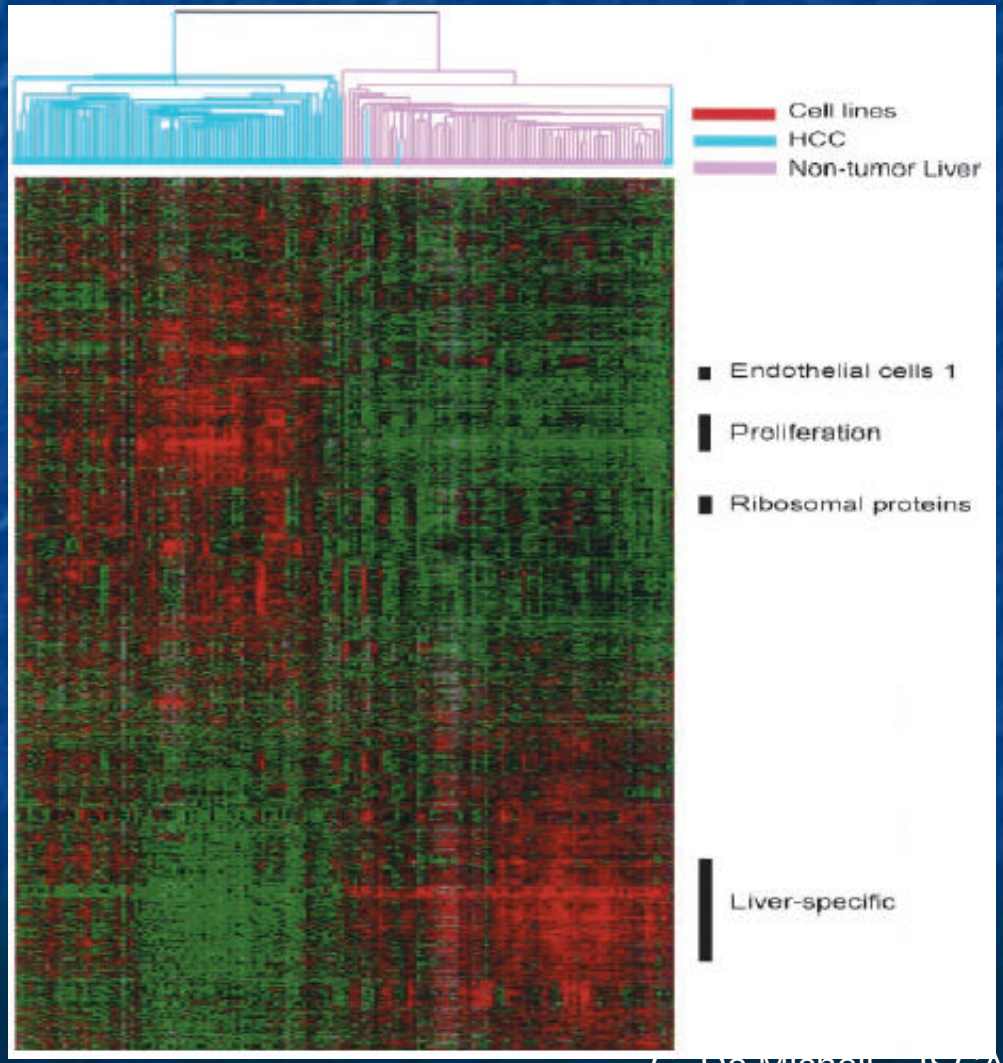
Data analysis



Objectives

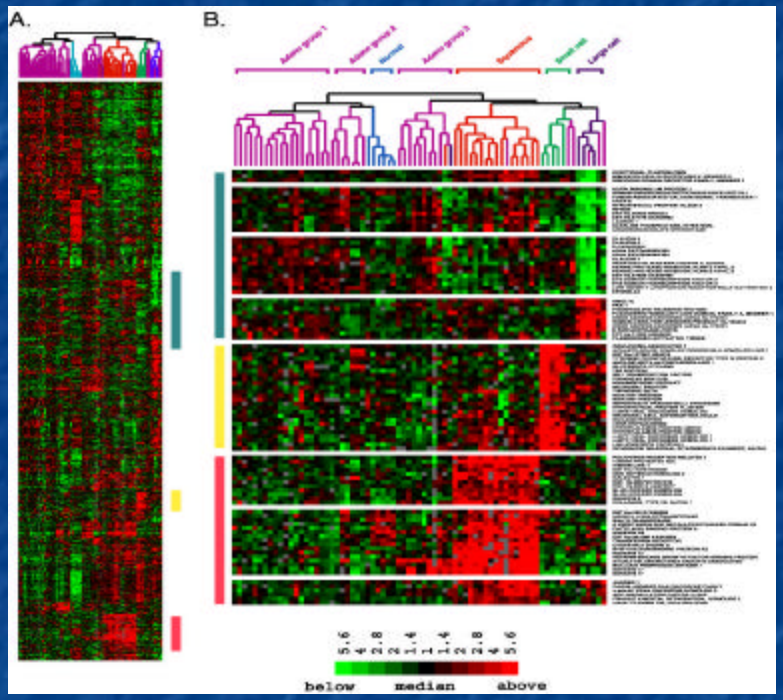
- Clinical analysis
  - Discover signatures of diseases
  - Discover effect of medications
- Bio-discovery
  - Gene regulatory networks
- Pharmacogenomics
  - Design drugs with direct impact on genetic features

Identifying Disease Genes



X. Chen & P.O. Brown et al
Molecular Biology of the Cell
Vol. 13, 1929-1939, June 2002

Disease Subtype Classification

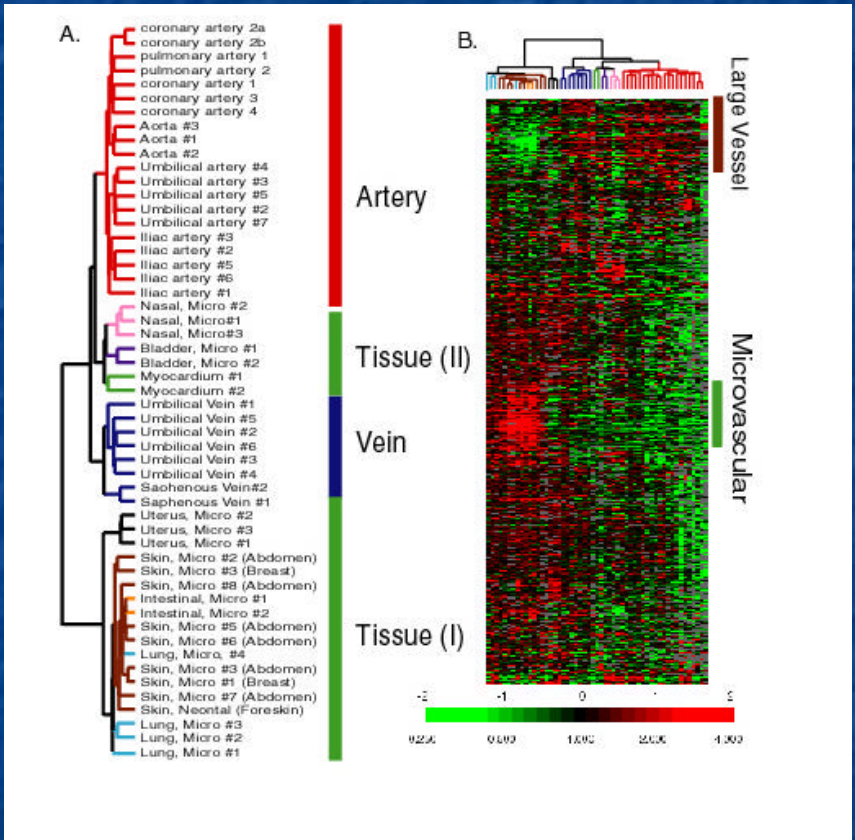


Diversity of gene expression in adenocarcinoma of the lung
Garber et Al. PNAS November 20, 2001 vol 98 n.24 13784-13789

Today

- Patient side:
  - More efficient treatment, i.e. avoid side effects of useless therapy
- Health Care System side:
  - Efficient use of information, time and resources

Understanding of Bio-mechanisms



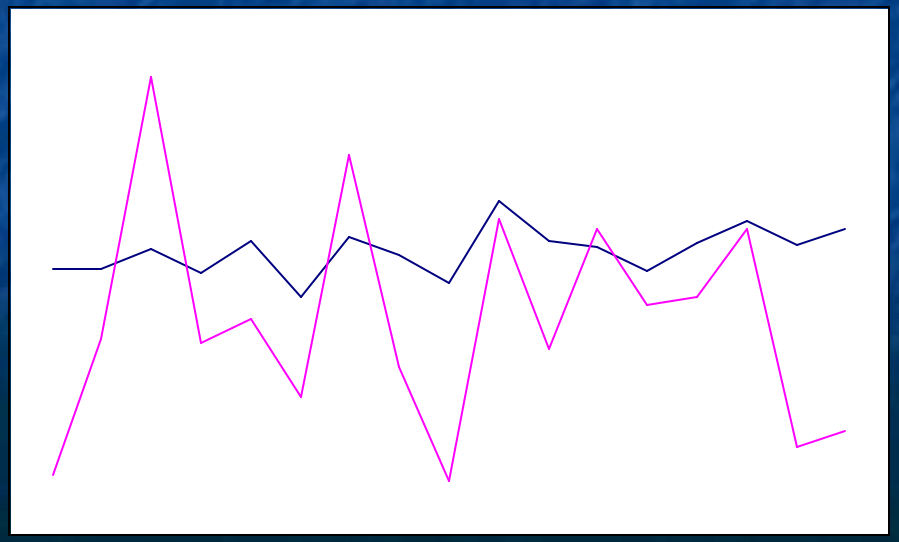
Today

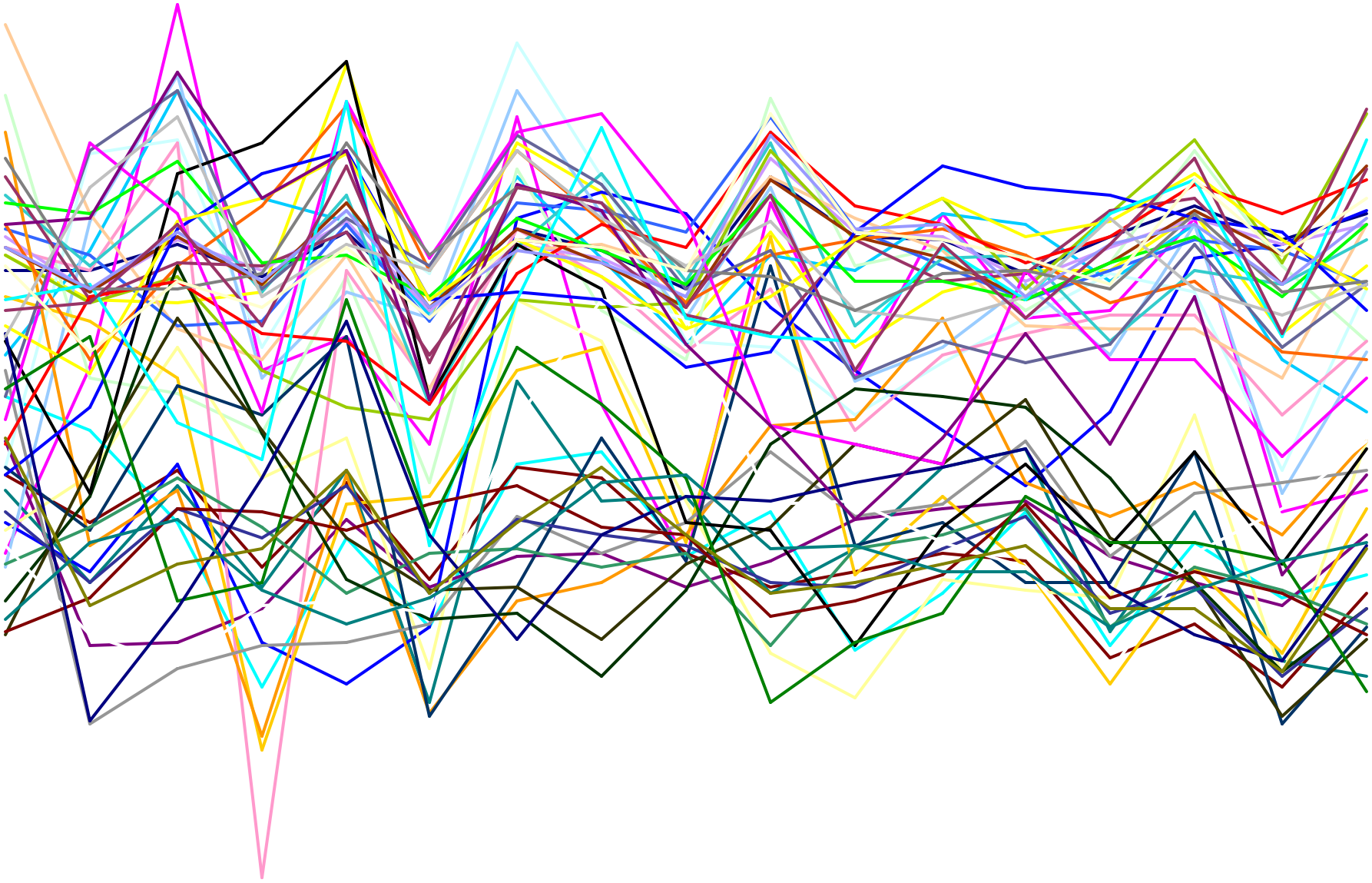
- Experiments are non-pathological varying environmental condition
- i.e. how gene expression of endothelial cells changes with blood pressure variation

Endothelial cell Diversity revealed by global expression profiling, Chi, Chang, Haraldsen, PNAS, September 16 2003, vol100, n.19, 10623-10628

Data analysis

- To find common patterns in gene expression levels
  - Upregulated/downregulated genes
- To group similar patterns

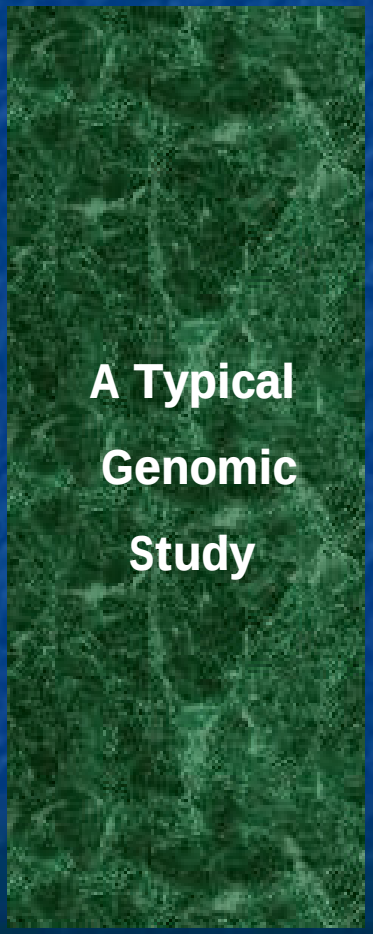




The curse of dimensionality

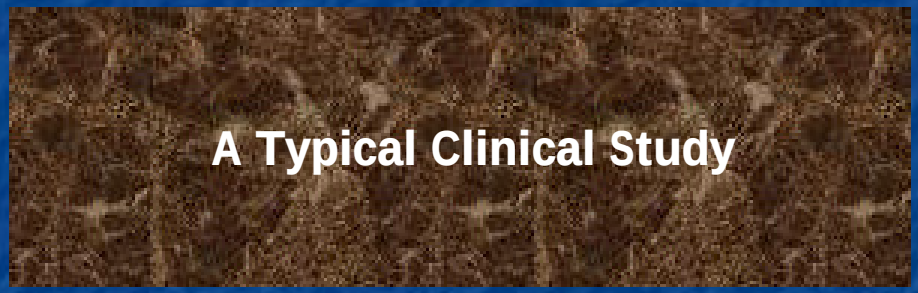
Cases (10's – 100's)

Variables (10,000's – 100,000's)



Variables (10's – 100's)

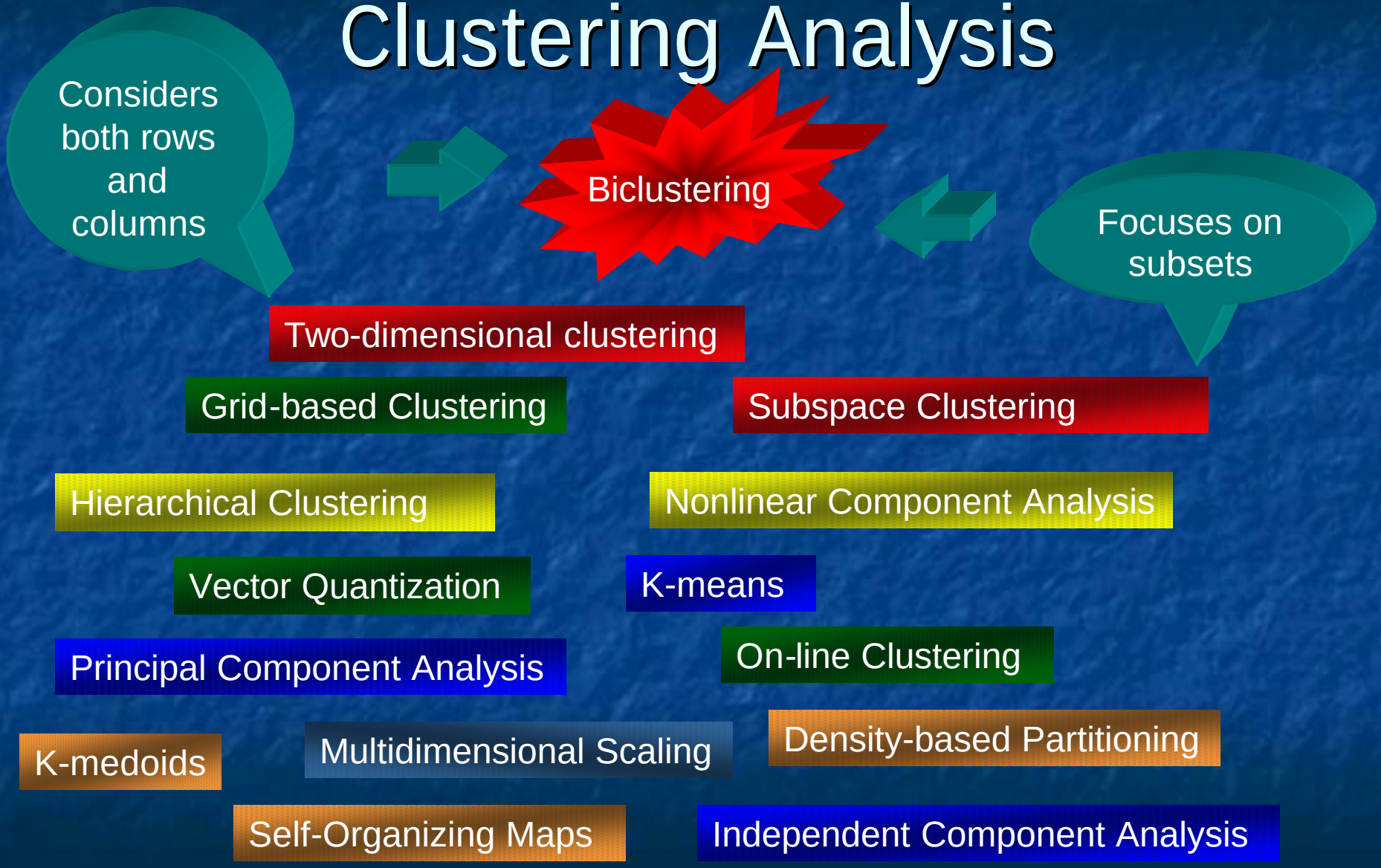
Cases (1,000's – 1,000,000's)



- Underdetermined system

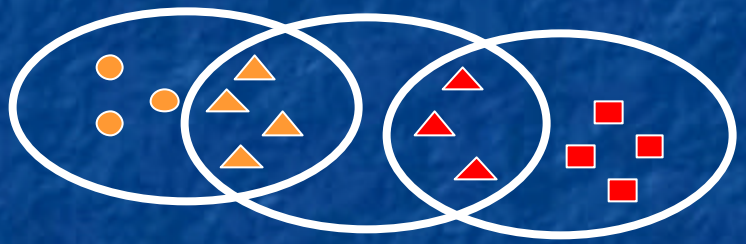
Kohane et al., *Microarrays for Integrative Genomics*, 2003

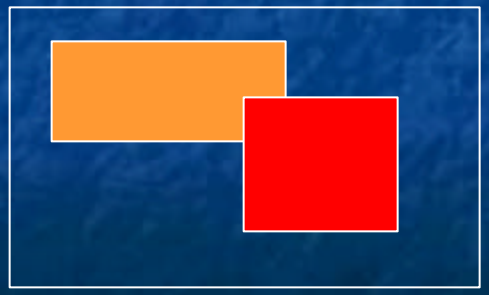
Clustering Analysis



Bi-clustering algorithms

- Several approaches
  - δ-biclustering [Cheng & Church 2000]
    - Heuristic
  - P-Clustering [Wang et al. 2002]
    - Heuristic
  - δ-Pclustering [Yoon et al. 2004]
    - Exact
- Advantages:
  - Capture relevant subspaces
  - Capture *coherence* and *fluctuation*
- Disadvantages:
  - Computationally hard

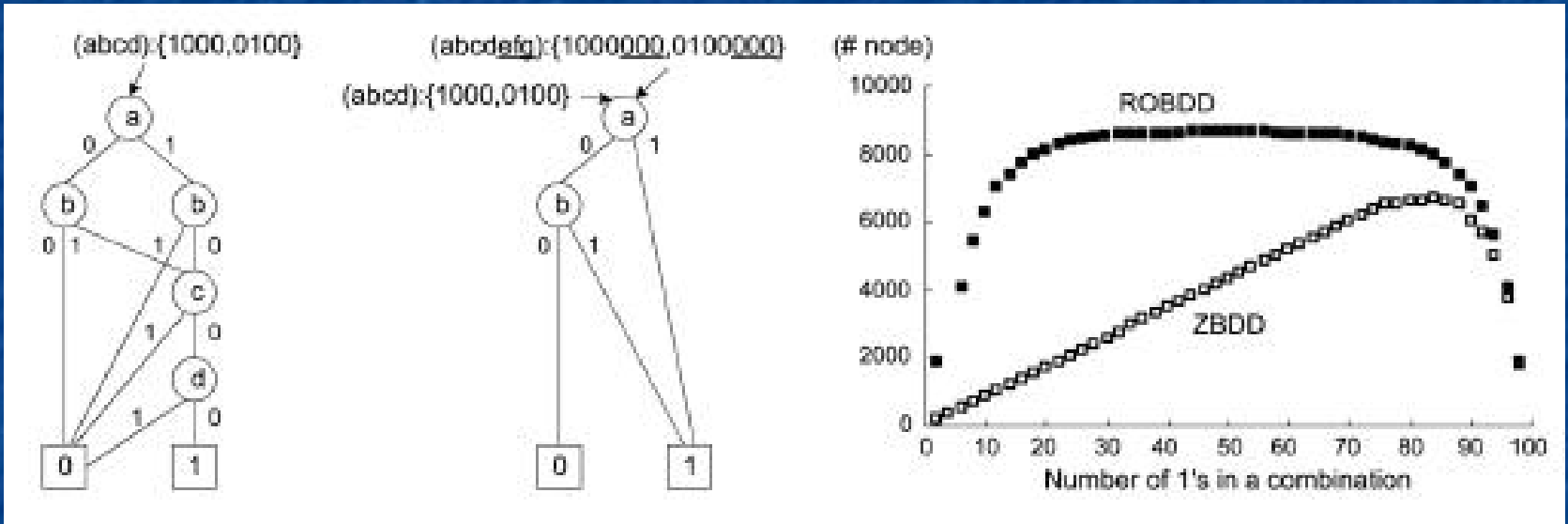




Computational Challenge

- Find all relevant maximal biclusters
  - Rank them according to some metric
    - E.g. Mean score residue (MSR)
- Exploit methods for large scale data handling
  - BDDs and ZDDs
- Use implicit representation to avoid full set representation and to support operations

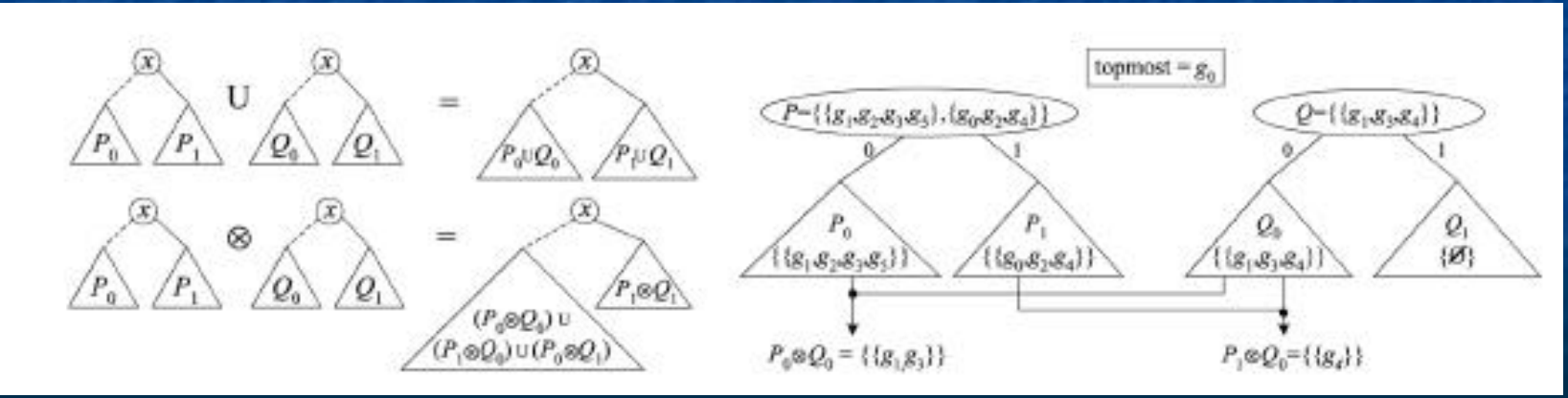
Zero-suppressed BDDs
for set representations



[Minato 86]

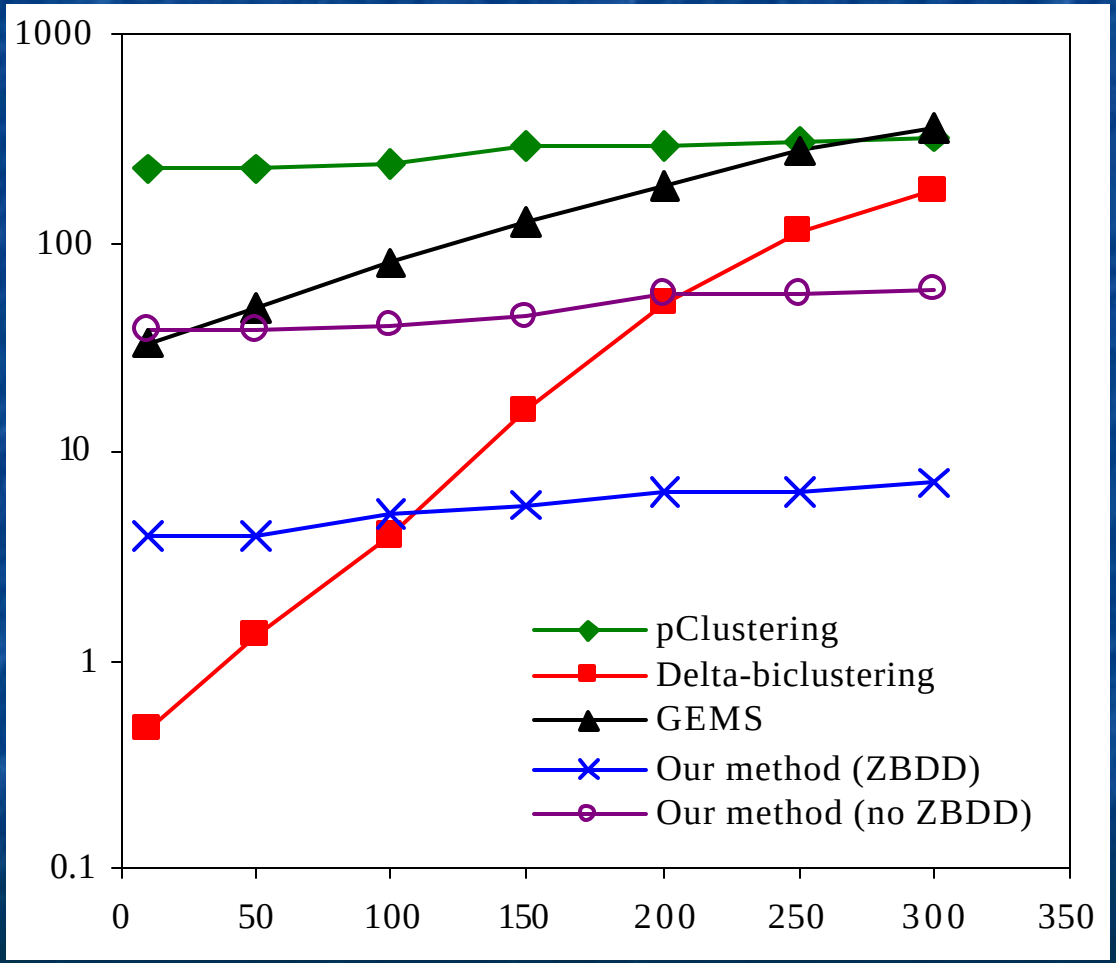
Set operations with ZDDS

- ZDDs support well recursive operations
  - Like union and intersection
- Recursive algorithms are w.c. exponential
  - But linear with representation size
  - Representation size grows mildly with problem size



Gene Expression Data

Time (sec.)



biclusters found

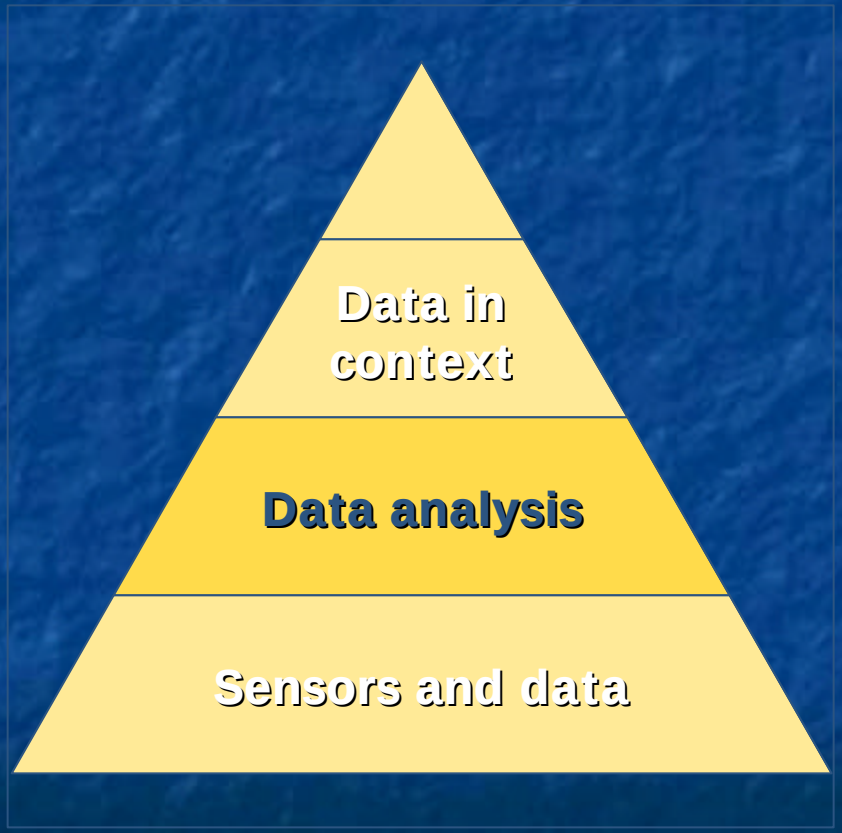
Yeast cell cycle data
(Tavazoie et al., 1999)

[Yoon 2004]

Keypoints

- Clustering is a key difficult problem to solve
- Methods based on implicit set representations (e.g. with ZDDs) can cope with large data set
- For all practical purposes, micro-array data can be efficiently processed today
- Outstanding issues
  - Biological validation of data
  - Statistical analysis of results

Data and its context ontologies

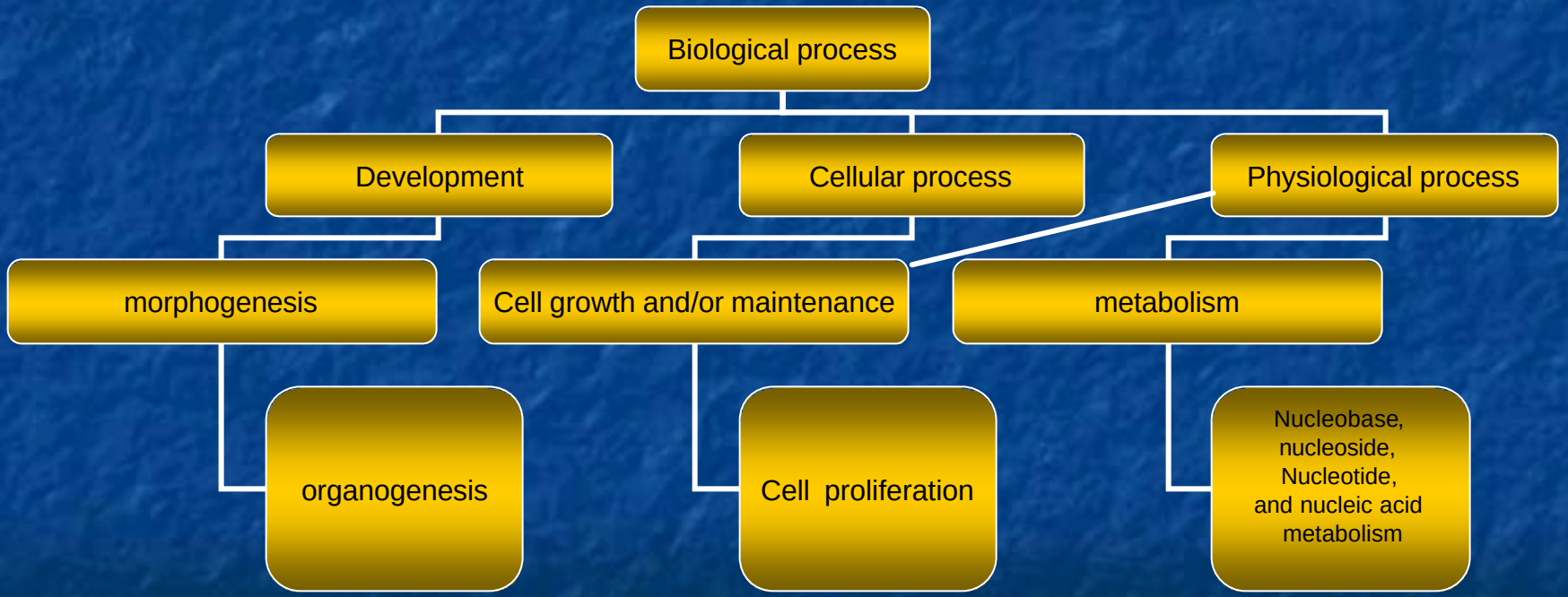


GO – The 3 Ontologies

- Every gene is classified from three points of view:
  - Biological process
  - Molecular function
  - Cellular component
- Directed acyclic graphs
  - Each node is uniquely labeled
    - Ex: GO:000xxxx
  - Flat description files

GO - Direct Acyclic Graph

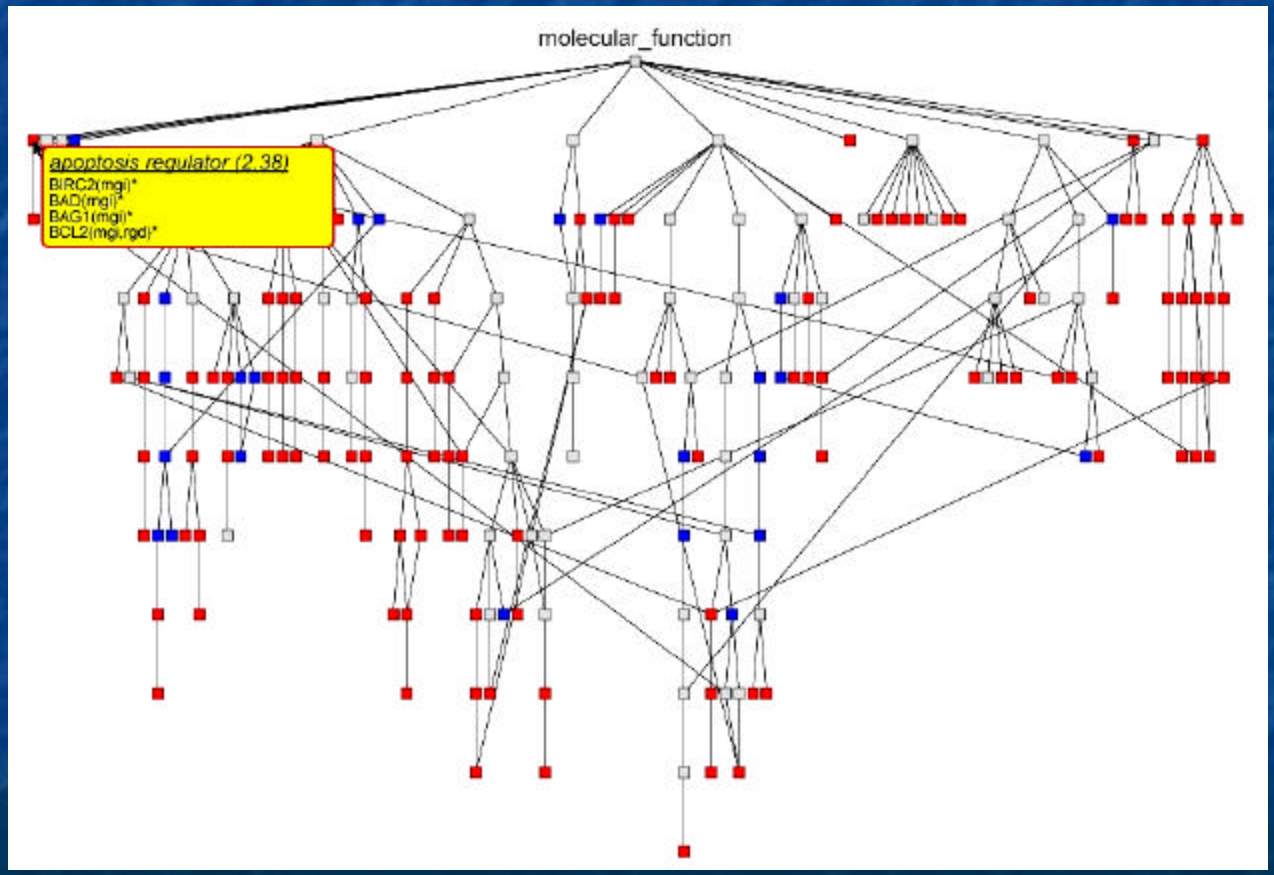
Biological Process Ontology



GO - Flat Files

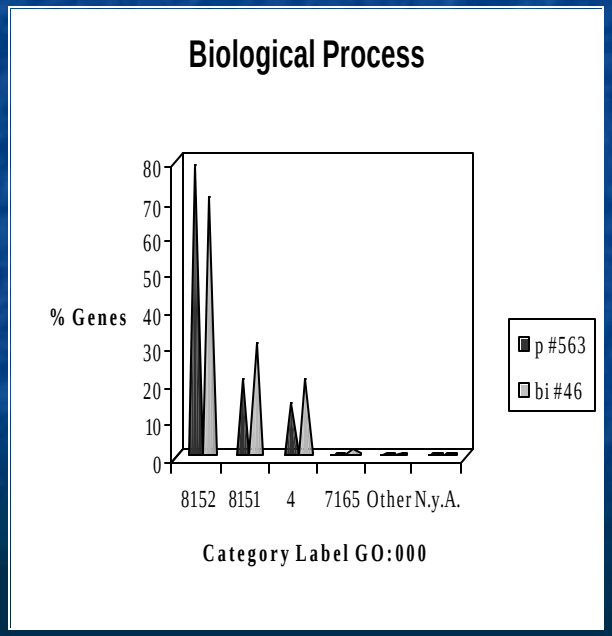
GO:0000067$Gene_Ontology ; GO:0003673
@part_of:biological_process ; GO:0008150
@is_a:cellular process ; **GO:0009987**
@is_a:cell growth and/or maintenance ; GO:0008151 ; synonym:cell physiology @is_a: physiological process ; GO:0007582
@is_a:cell proliferation ; GO:0008283
@part_of:cell cycle ; GO:0007049 ; synonym:cell-division cycle
@part_of:DNA replication and chromosome cycle ; **GO:0000067**
@is_a:physiological process ; **GO:0007582**
@is_a:cell growth and/or maintenance ; GO:0008151 ; synonym:cell physiology @is_a: cellular process ; GO:0009987
@is_a:cell proliferation ; GO:0008283
@part_of:cell cycle ; GO:0007049 ; synonym:cell-division cycle
@part_of:DNA replication and chromosome cycle ; **GO:0000067**
@part_of:cellular_component ; GO:0005575
@part_of:molecular_function ; GO:0003674

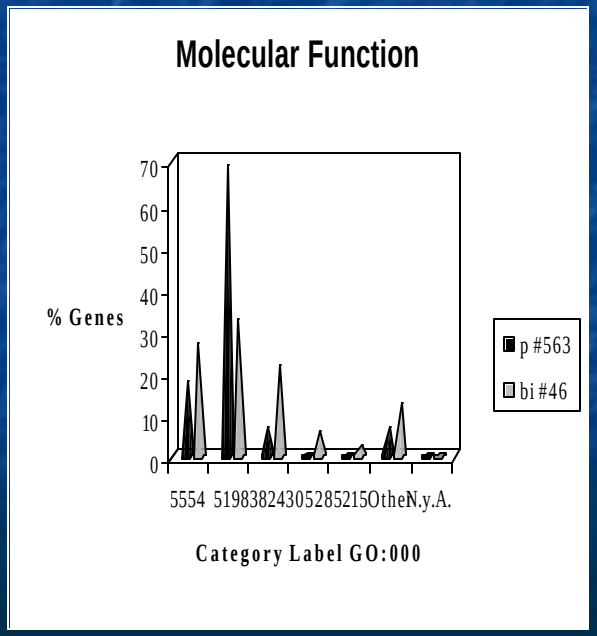
GoMiner - Positioning in Ontology

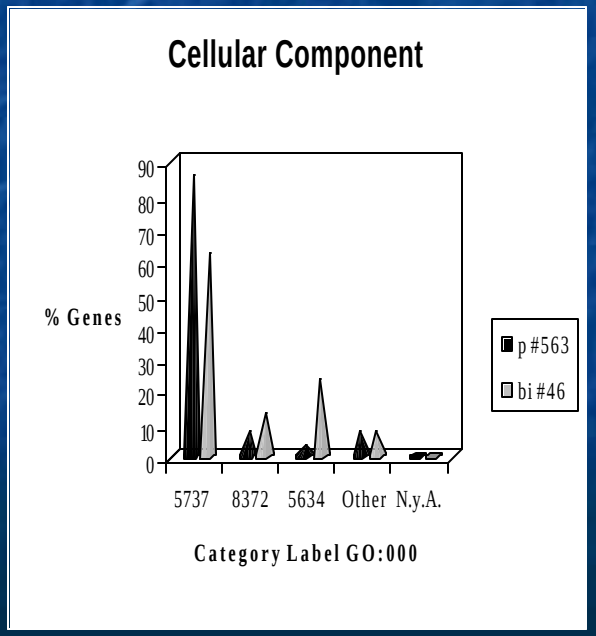


Cluster validation by GO

- A cluster with genes in a single category
  - Clear biochemical meaning
  - Precise positioning in GO

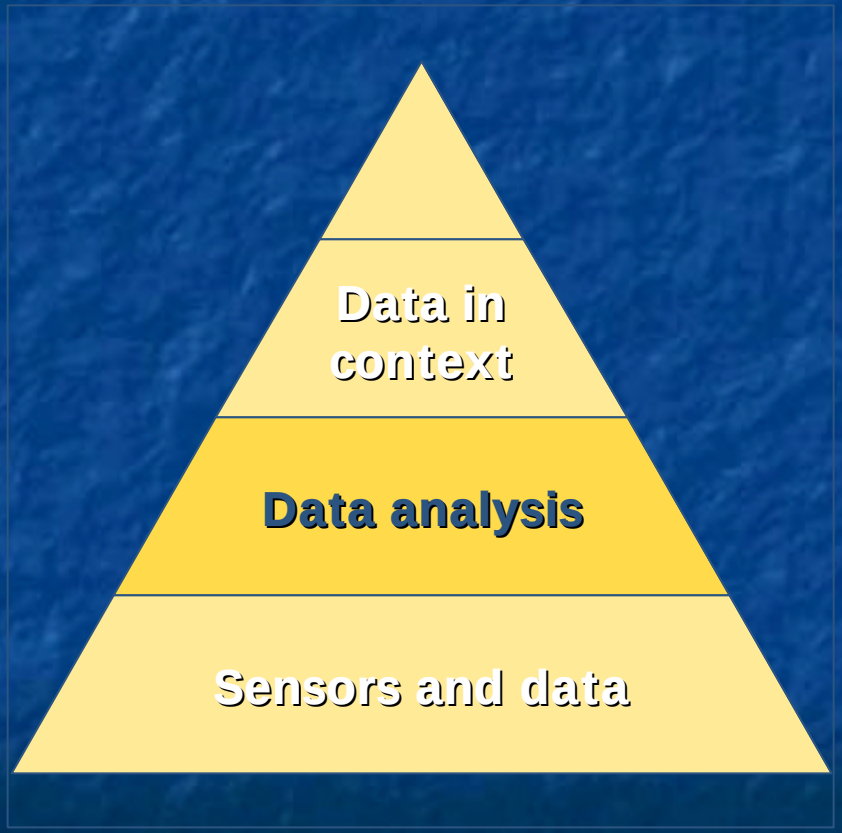






[Nardini et al. 2004]

Data and its context clinical traits



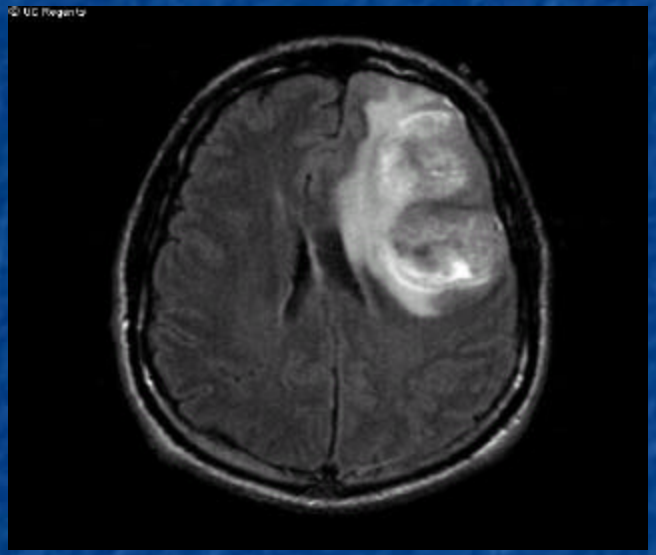
Goal

- Correlate clinical traits with genetic data
- Example:
  - Biopsy genetic analysis and imaging data
- Objective:
  - Better diagnosis of diseases and response to medication

CAD -- Computer Aided Diagnosis

Radiological Data

MRI



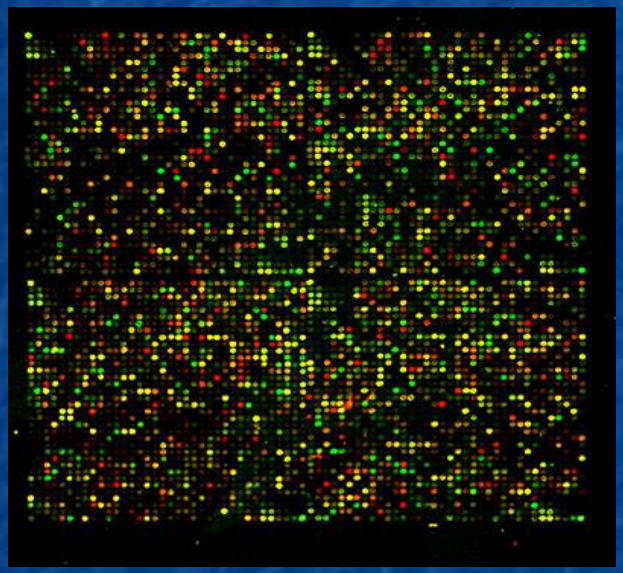
Coding into a categorical imaging scoring schema

PATIENT ID	shx013: 49A34	shv060: 45A9	shq077: 52A28	shx009: 4A34	shx014: 61A31	shq082: 99A6	shq083: 46A15	shx008: 41A31	shv062: 70P	shx015: 69A29
Imaginog score	0	1	0	1	1	1	1	1	1	0

[Nardini, RSNA 2004]

Genomic Data

Microarray



Coding into an expression value scoring schema

CLID	PATIENT ID	shx013: 49A34	shv060: 45A9	shq077: 52A28	shx009: 4A34	shx014: 61A31	shq082: 99A6	shq083: 46A15	shx008: 41A31
IMAGE:74	ISG20 || in	-1.02	-2.34	1.44	0.57	-0.13	0.12	0.34	-0.51
IMAGE:76	TNFSF13	-0.52	-4.06	-0.29	0.71	1.03	-0.67	0.22	-0.09
IMAGE:36	LOC93343	-0.25	-4.08	0.06	0.13	0.08	0.06	-0.08	-0.05
IMAGE:23	ITGA4 || in	-1.375	-1.605	0.155	-0.015	0.035	-0.035	0.505	-0.865

Radiogenomic Data

Patients (20)

Clinical Data (5 Binary traits)

Genomic Data (~2188)

CLID	PATIENT ID	shx013: 49A34	shv060: 45A9	shq077: 52A28	shx009: 4A34	shx014: 61A31	shq082: 99A6	shq083: 46A15	shx008: 41A31	shv062: 70P
Imaging data	Trait 1	0	1	0	1	1	1	1	1	1
	Trait 2	1	0	1	0	0	1	0	1	0
	Trait 3	1	1	1	0	1	1	1	1	1
	Trait 4	1	1	1	1	0	0	1	1	0
	Trait 5	1	1	1	0	0	1	1	0	1
IMAGE:74	ISG20 || in	-1.02	-2.34	1.44	0.57	-0.13	0.12	0.34	-0.51	0.47
IMAGE:76	TNFSF13	-0.52	-4.06	-0.29	0.71	1.03	-0.67	0.22	-0.09	0.1
IMAGE:36	LOC93343	-0.25	-4.08	0.06	0.13	0.08	0.06	-0.08	-0.05	-6.56E-09
IMAGE:23	ITGA4 || in	-1.375	-1.605	0.155	-0.015	0.035	-0.035	0.505	-0.865	1.325
IMAGE:78	SGCE || sa	-0.4	-3.13	0.62	1.91E-08	0.56		-0.42	-0.91	0.12
IMAGE:83	RPS4Y || *	-0.83	-1.3	0.65	-2.07	0.31	-2.16	0.88	-2.37	0.36
IMAGE:81	SMCY || S	-1.99	-2	0.21	-2.15	1.25	-2.7	0.38	-2.71	0.1
IMAGE:78	DBY || DE	-1.29	-0.99	0.34	-1.35	-0.03	-0.95	0.41	-1.01	0.42
IMAGE:30	EIF1AY ||	-1.06	-0.72	0.78	0.01	0.24	-0.91	0.48	-0.6	0.54

Co-clustering

- Unsupervised cluster search method in higher dimensional spaces
- Correlate data and provide qualitative information about pathology
- Results:
  - Method was applied to renal carcinoma with good correlation
- Outstanding issue:
  - Scoring of radiological data
  - Can this be automated?

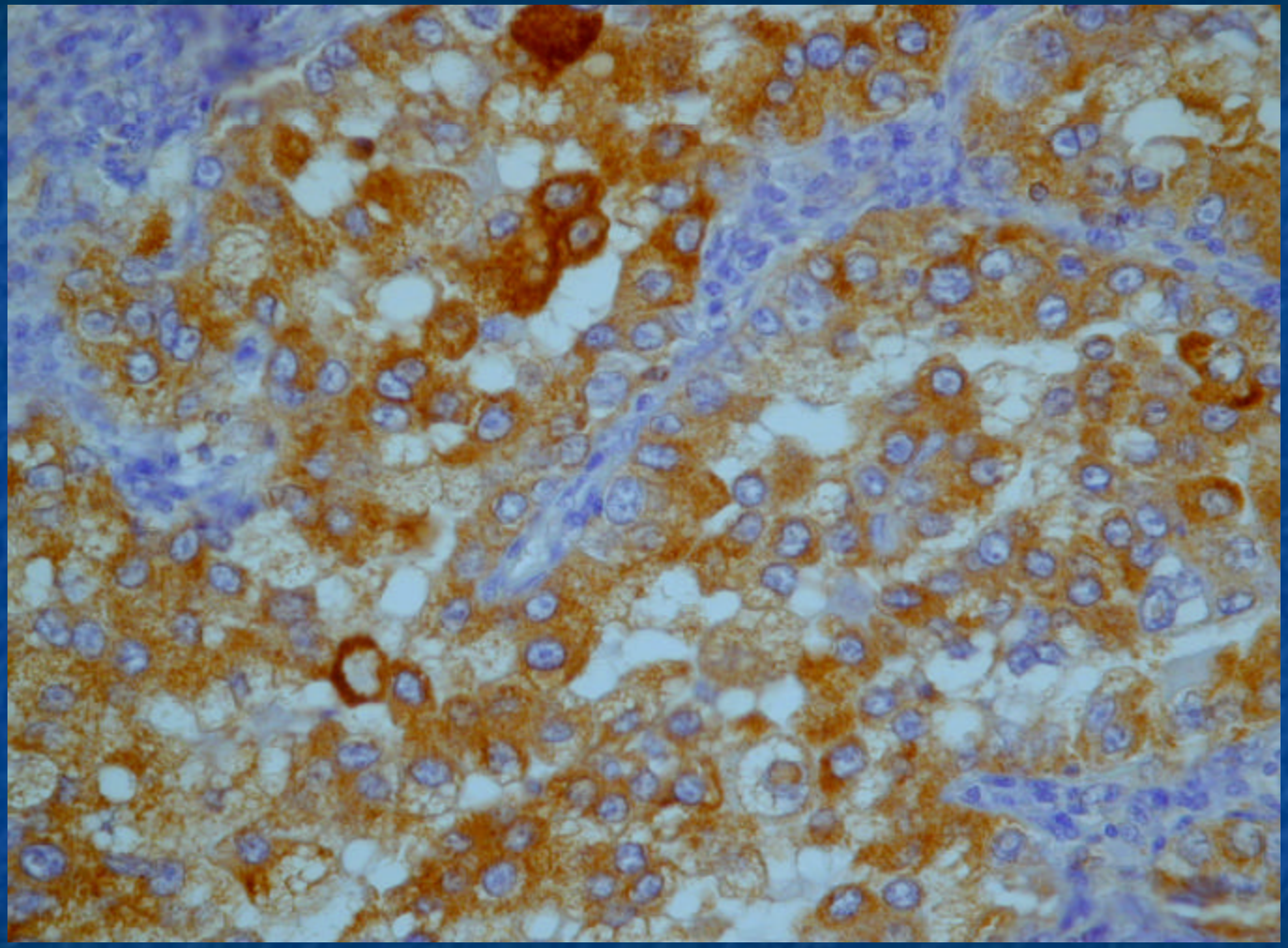
Immunohistochemical
Automated Quantification

- Acquiring quantitative and qualitative information from immuno-stains

ß

- Automated image processing methods to standardize IHC analysis

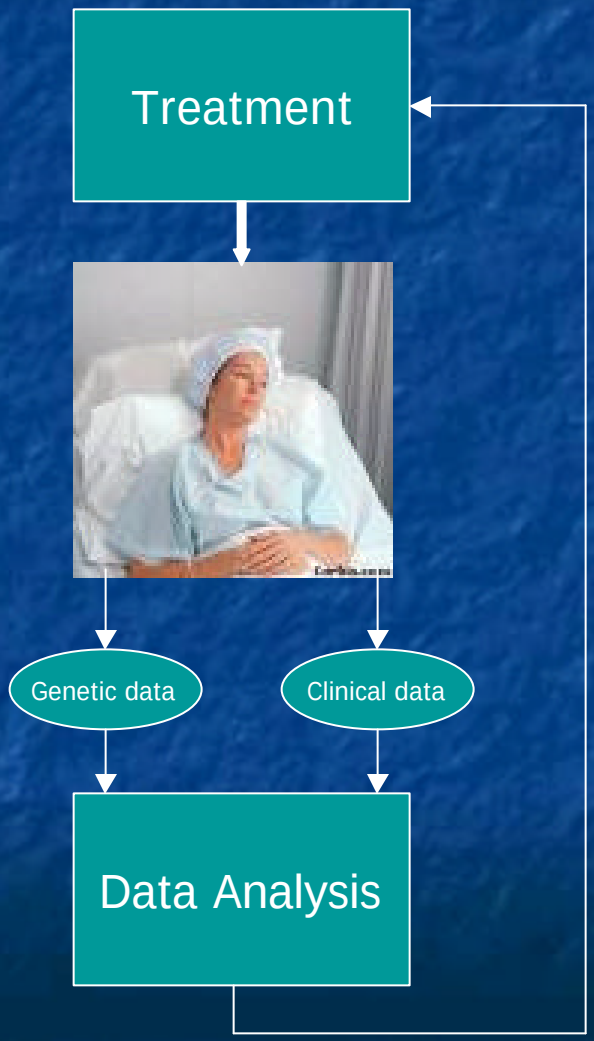
Example of IHC images



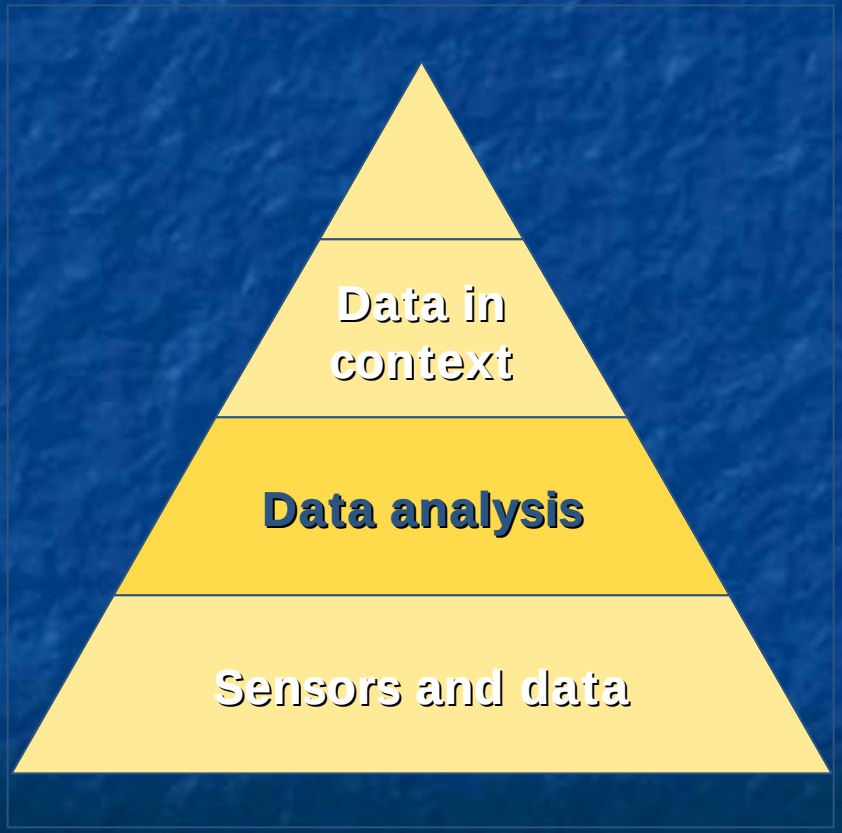
EGFR/erb-B receptors positivity (in carcinoma cells) as brown stain
Negative carcinoma cells and other cells in the sample as blue stain

System view

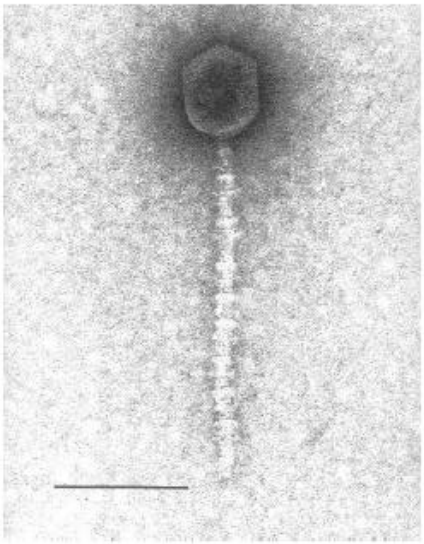
- A closed loop system
  - Identification
  - Optimum control

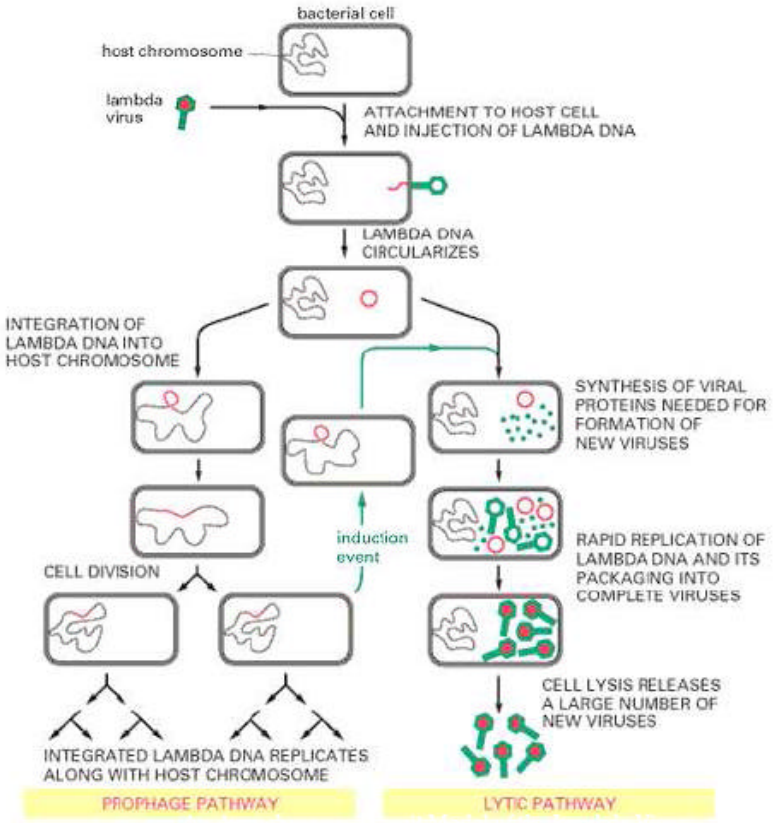


Microarray data and genetic pathways



Gene regulatory networks
Example: Phage λ

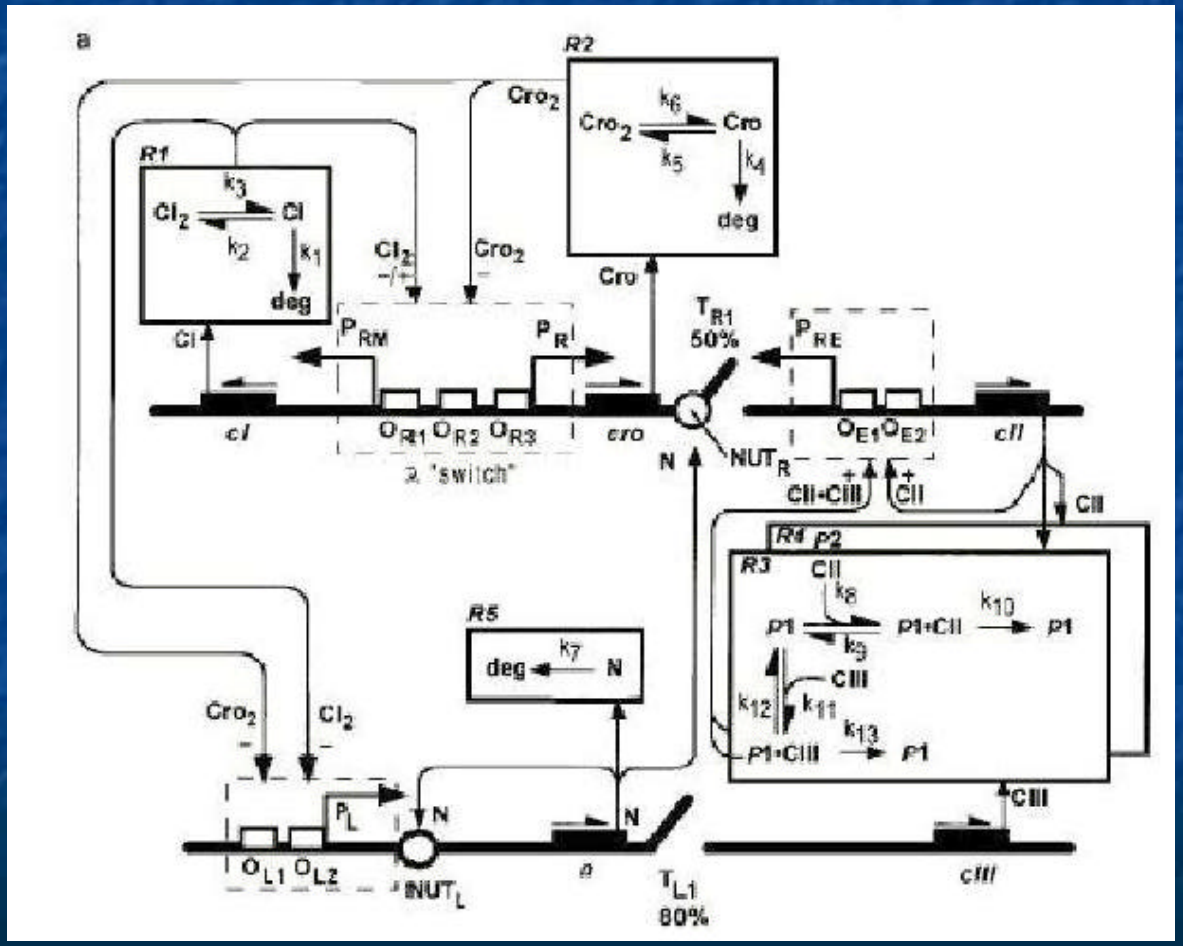




[Myers 2003]

Gene regulatory networks

Example: Phage λ decision circuit

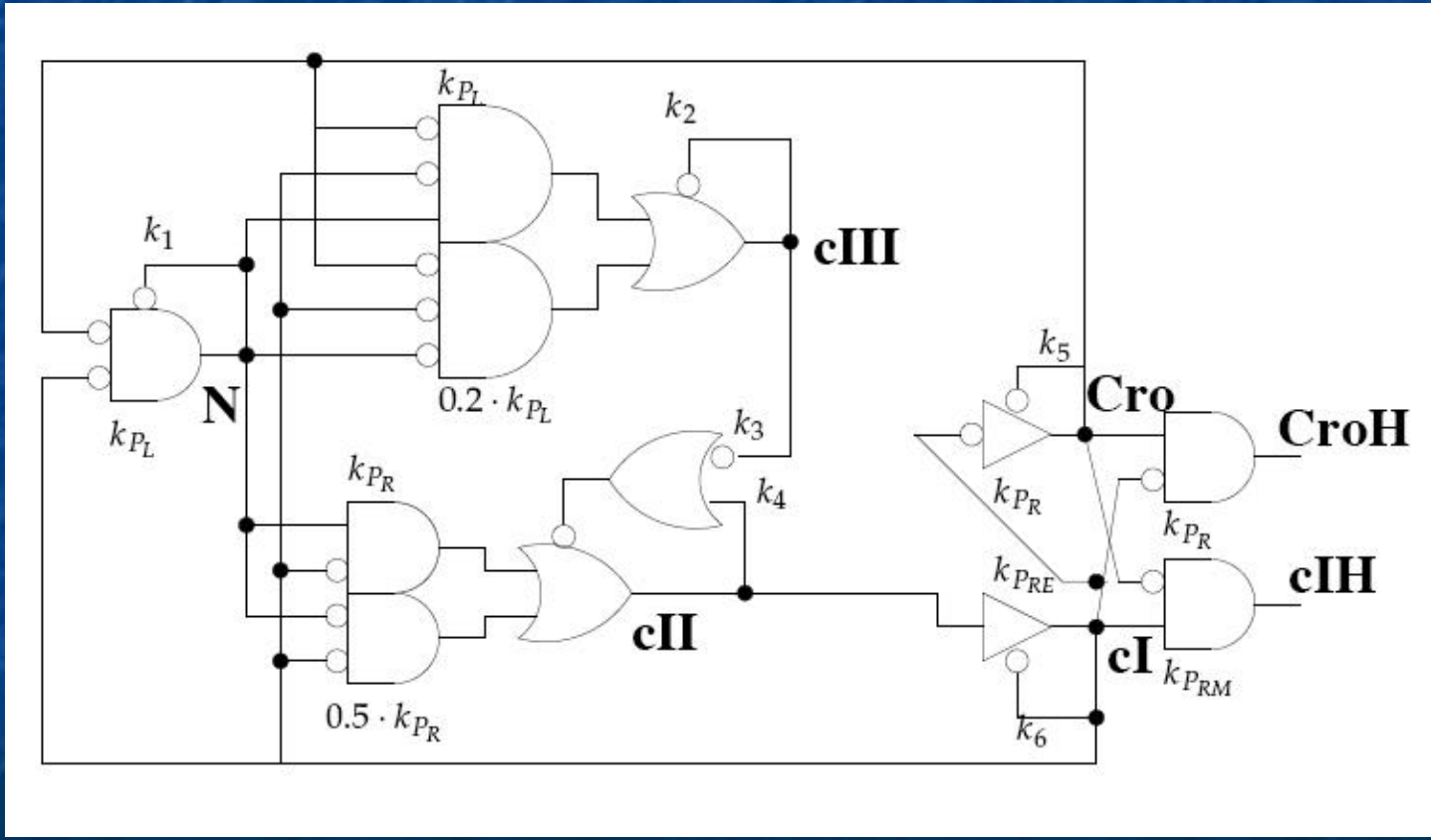


[MYERS 2003]

Models of gene evolution

- Finite-state behavior
  - FSMs
- Transients
  - ODEs
  - Hybrid systems
- Stochasticity
  - Markov models
- Asynchrony

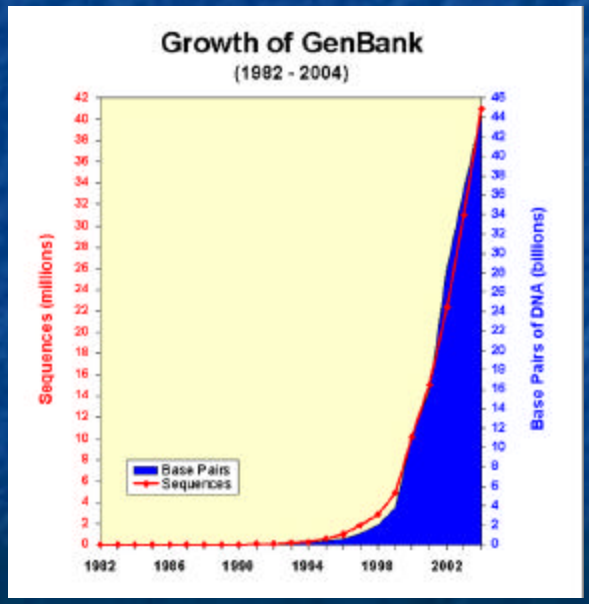
Gene regulatory networks
Example: Phage λ asynchronous model



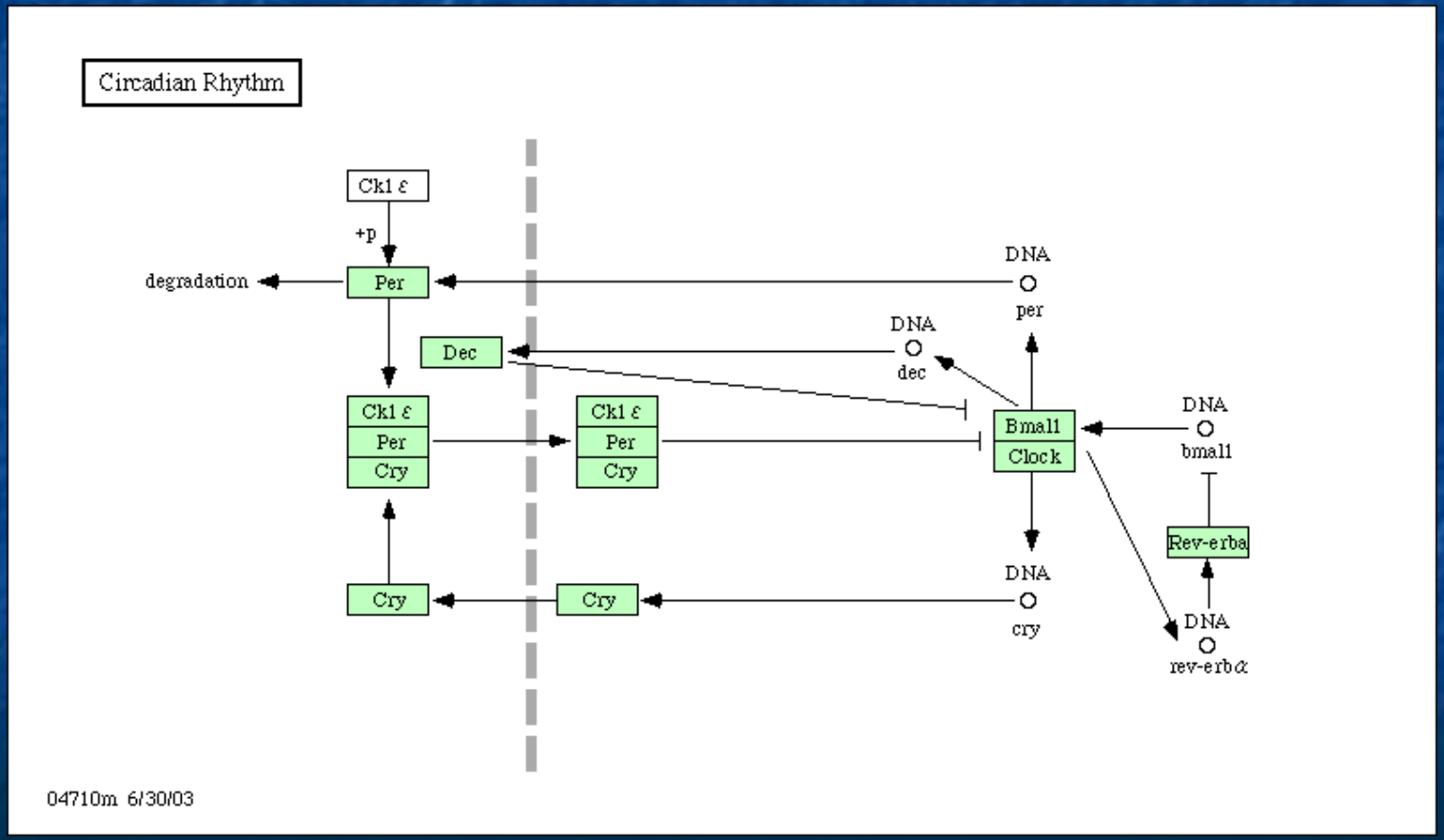
Gene regulatory networks

Databases

- Pathways
  - HS, animals and plants
- Genomic
  - BLAST searches
- Chemical
  - Compound structures
- BRITE
  - Drug taxonomy



Gene regulatory networks
Circadian cycle in homo sapiens



[KEGG]

Summary and conclusions

- Genomic information has been exploding
  - Its effective use is still limited
- Application areas include:
  - Clinical medicine, pharmaceutics, biology
- Technologies include:
  - Modeling, abstraction, circuit design, data analysis, dedicated languages, optimal control

CAS researchers master these technologies and can widen their horizons to these application areas