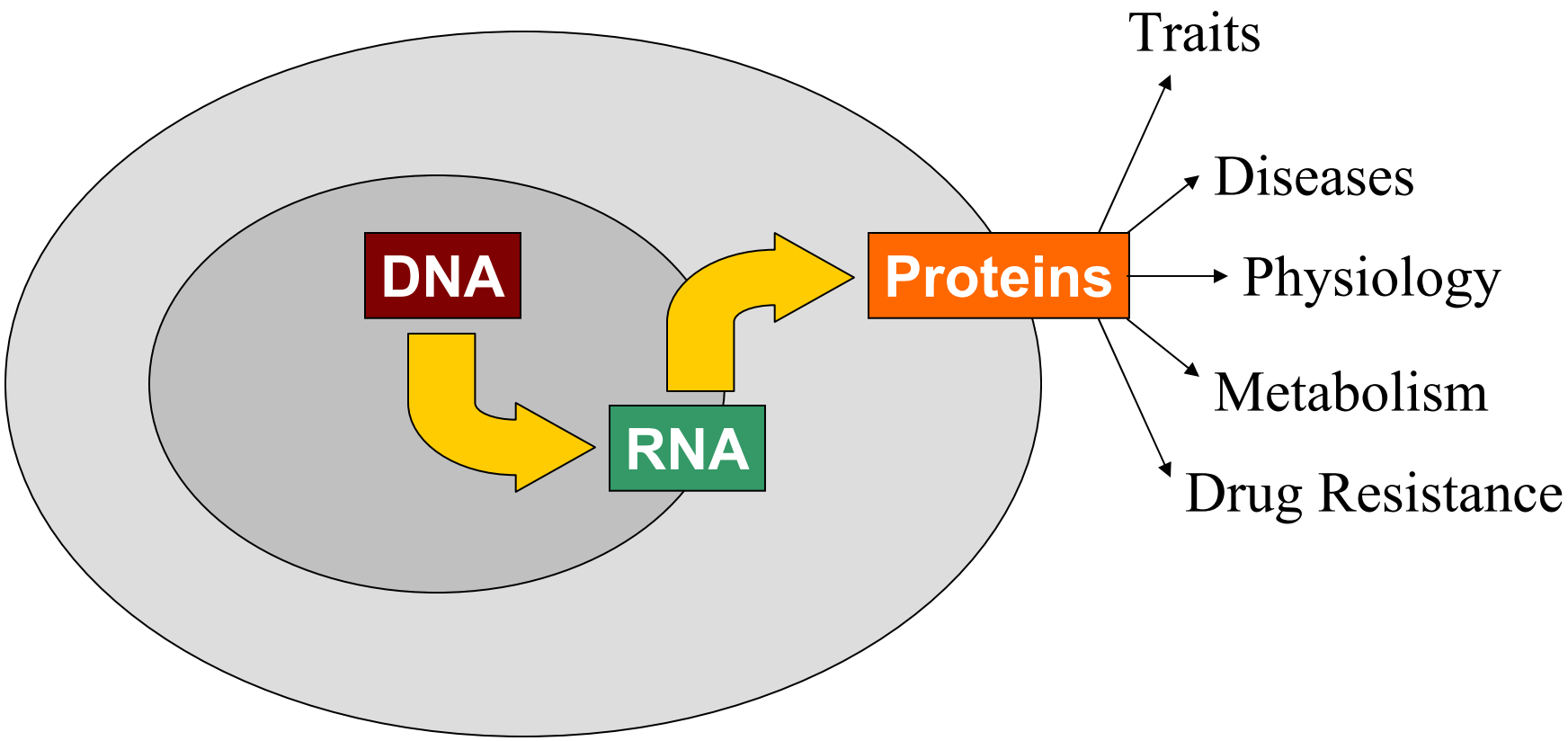
Building (Causal) Models from Experimental Data

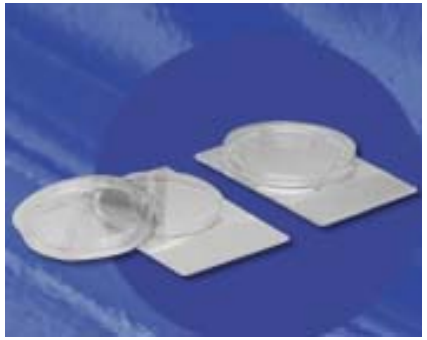
Marco Ramoni
Children's Hospital Informatics Program
Harvard Partners Center for Genetics and Genomics
Harvard Medical School

HST 950

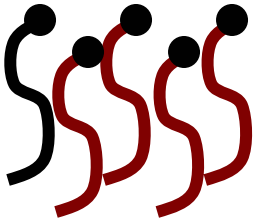
Central Dogma of Molecular Biology



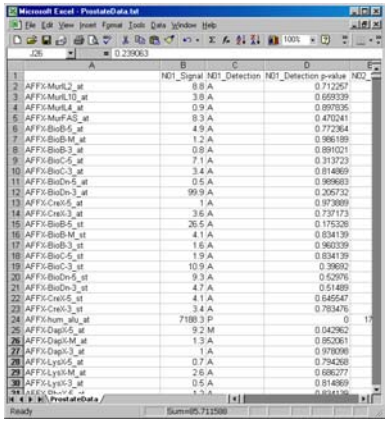
From Tissues to Microarrays



Tissues



Tagged by fluorescent dye



Data

Fluidics Station

Scanner

Image

Microarray Technology

Scope: Microarrays are reshaping molecular biology.

Task: Simultaneously measure the expression value of thousands of genes and, possibly, of entire genomes.

Definition: A microarray is a vector of probes measuring the expression values of an equal number of genes.

Measure: Microarrays measure gene expression values as abundance of mRNA.

Types: There are two main classes of microarrays:

- cDNA: use entire transcripts;
- Oligonucleotide: use representative gene segments.

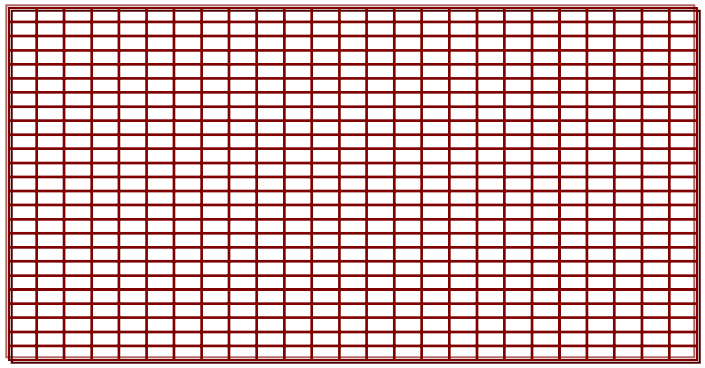
Statistical Challenges

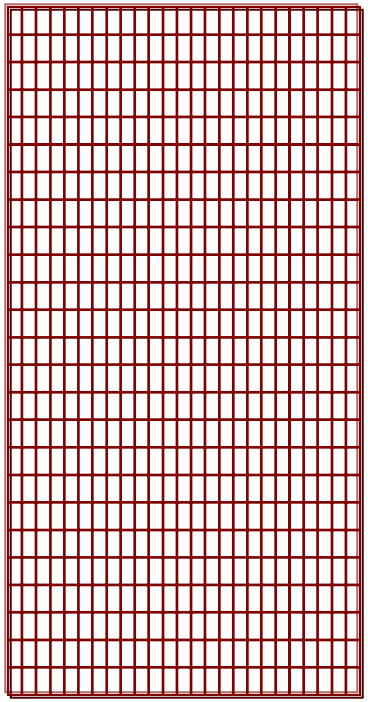
Small N large P: Many variables, few cases.

Noisy results: Measurements are vary variable.

Brittle conditions: Sensitive to small changes in factors.

Design: Platforms are designed without a clue about the analysis to be done.





Clustering for Causality

(Statistical) Rules for Causality:

- ✓ Correlation;
- ✓ Time-lag;
- ✓ No hidden-variables.

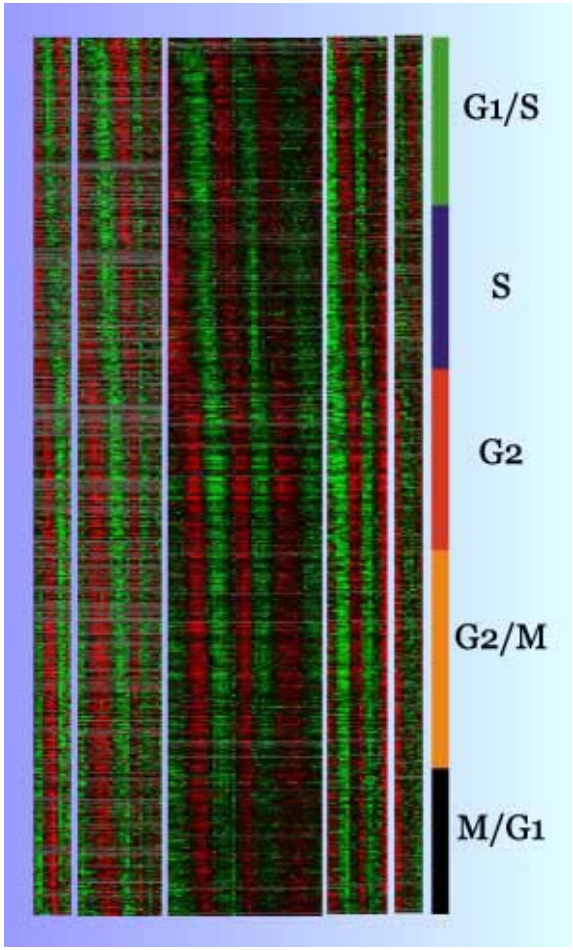
Challenge: data dimensionality.

Proof of concept: Cell cycle.

Method: clustering/eye-balling.

Argument: Identification of cell cycle phases.

Deficit: No method to identify gene control mechanisms.



Bayesian Networks

Qualitative: A dependency graph made by:

Node: a variable X, with a set of states $\{x_1,\ldots,x_n\}$.

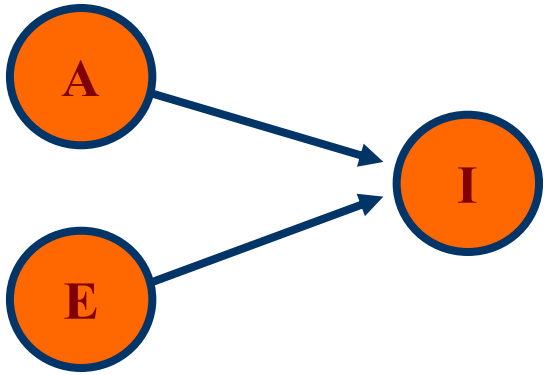
Arc: a dependency of a variable X on its parents Π.

Quantitative: The distributions of a variable X given each combination of states π_i of its parents Π.

Semantics: A graph encodes conditional independence.

A	p(A)
Y	0.3
O	0.7

E	p(E)
L	0.8
H	0.2

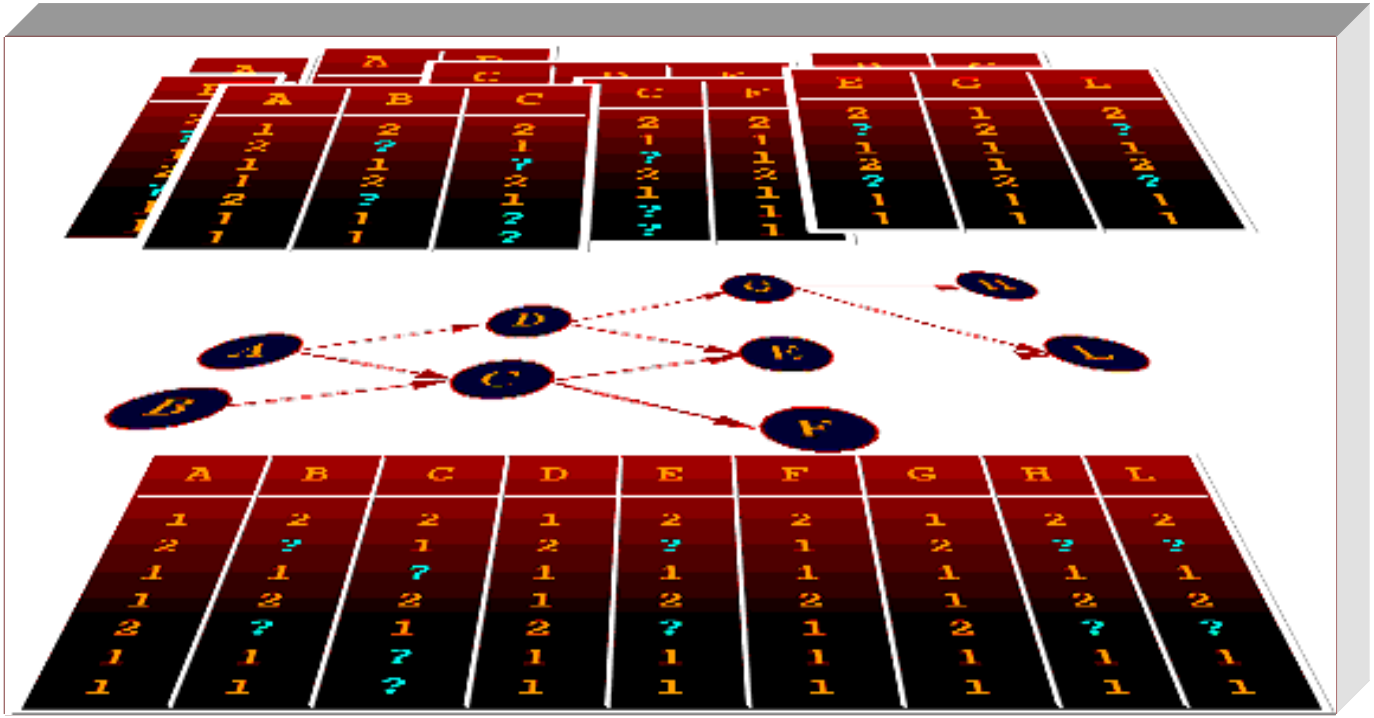


A	E	I	p(I|A,E)
Y	L	L	0.9
Y	L	H	0.1
Y	H	L	0.5
Y	H	H	0.5
O	L	L	0.7
O	L	H	0.3
O	H	L	0.2
O	H	H	0.8

A=Age; E=Education; I=Income

Factorization

- The graph factorize the likelihood: the "global" likelihood is the product of all local likelihood.

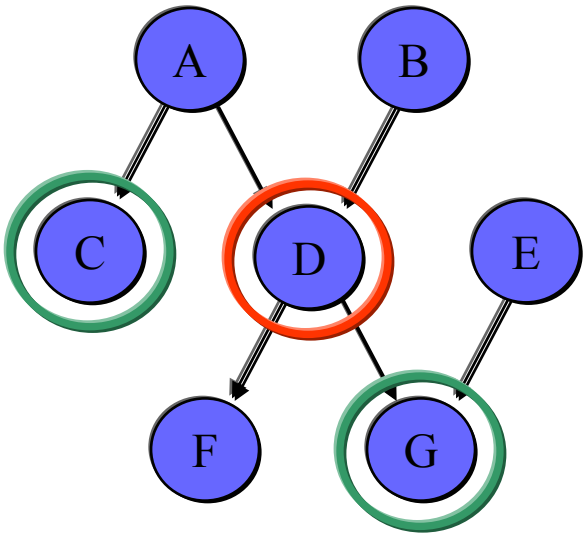


Reasoning

Components of a problem:

- Knowledge: graph and numbers.
- Evidence: e={c and g}.
- Solution: p(d|c,g)=?

Note: Lower case is an instance.



A	p(A)
0	0.3
1	0.7

B	p(B)
0	0.6
1	0.4

E	p(E)
0	0.1
1	0.9

A	C	p(C|A)
0	0	0.25
0	1	0.75
1	0	0.50
1	1	0.50

D	F	p(F|D)
0	0	0.80
0	1	0.20
1	0	0.30
1	1	0.70

A	B	D	p(D|A,B)
0	0	0	0.40
0	0	1	0.60
0	1	0	0.45
0	1	1	0.55
1	0	0	0.60
1	0	1	0.40
1	1	0	0.30
1	1	1	0.70

D	E	G	p(G|D,E)
0	0	0	0.90
0	0	1	0.10
0	1	0	0.70
0	1	1	0.30
1	0	0	0.25
1	0	1	0.75
1	1	0	0.15
1	1	1	0.85

Learning Probabilities

- Learning of probability distributions means to update a prior belief on the basis of the evidence.
- Probabilities can be seen as relative frequencies:

$$p(x_j \mid \pi_i) = \frac{n(x_j \mid \pi_i)}{\sum_j n(x_j \mid \pi_i)}$$

- Bayesian estimate includes prior probability:

$$p(x_j \mid \pi_i) = \frac{a_{ij} + n(x_j \mid \pi_i)}{\sum_j a_{ij} + n(x_j \mid \pi_i)}$$

α_{ij} / α_i represents our prior as relative frequencies.

Learning the Structure

Processes: Data are generated by processes.

Probability: The set of all models is a stochastic variable $\mathcal{M}$ with a probability distribution $p(\mathcal{M})$.

Selection: Find the most probable model given the data.

$$p(M \mid \Delta) = \frac{p(\Delta, M)}{p(\Delta)} = \frac{p(\Delta \mid M)\, p(M)}{p(\Delta)}$$

Computation: If we use the same data and we assume all models to be equally likely a priori, then:

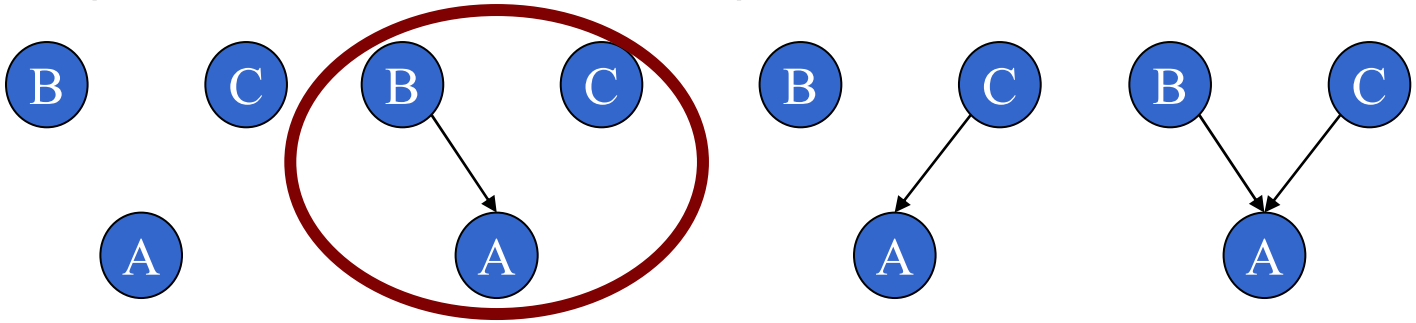
$$p(M|\Delta) \propto p(\Delta|M)$$

which is just the marginal likelihood.

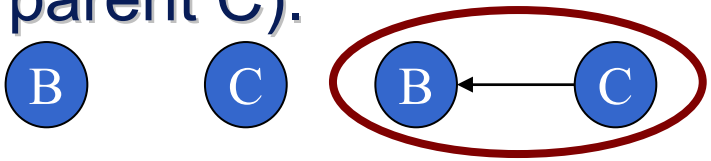
Strategy: Maximize the marginal likelihood

Local Model Selection

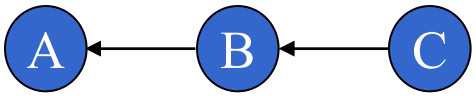
A (possible parents B; C):



B (possible parent C).

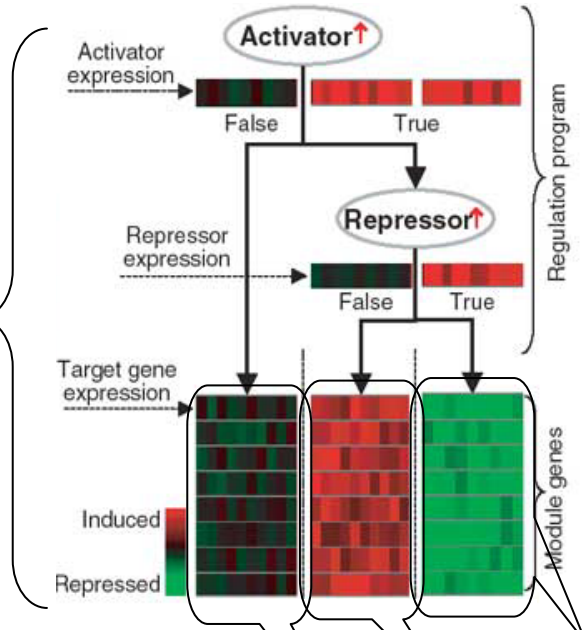


The model:



Module Networks

Model



~P(Activator)

~P(Repressor|Activator)

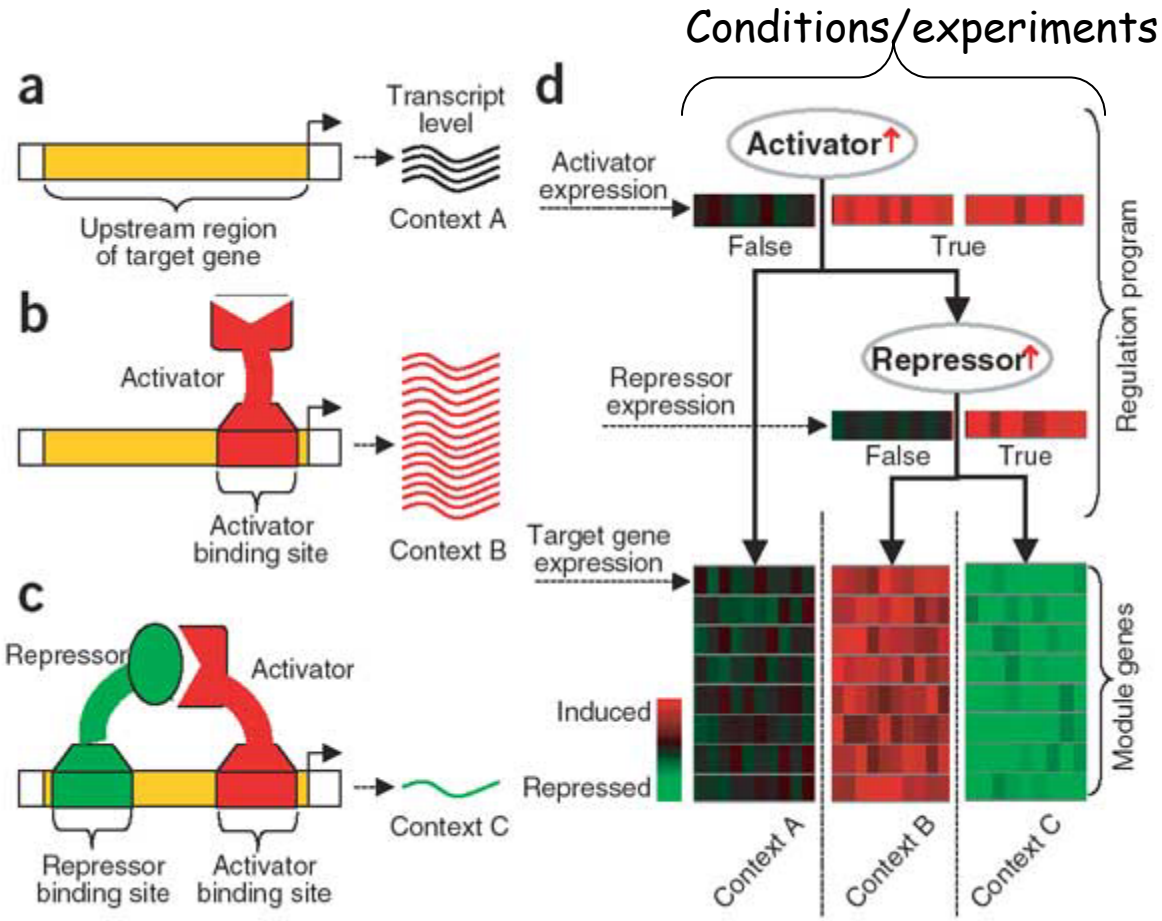
~P(Gene|A=up,R=up)

~P(Gene|A=up,R=down)

~P(Gene|A=down,R=down)

Segal, *Nat Genet*, 2003

Module Networks



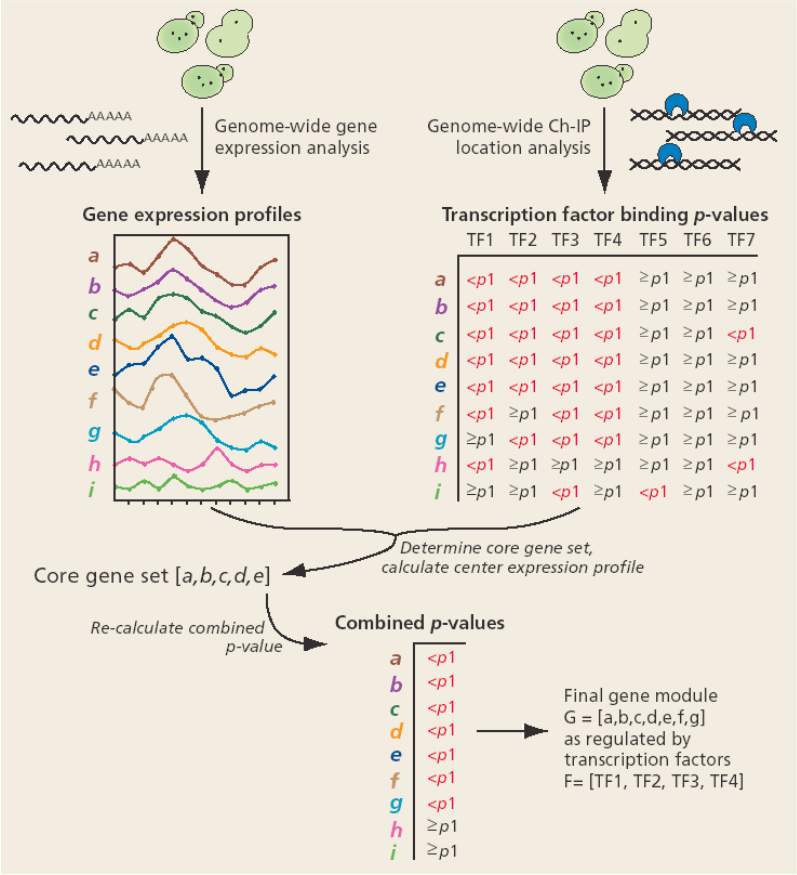
Chip ChIP Networks

Data: 500 expression datasets.

New Data: Chromatin Immuno-precipitation (ChIP) DNA arrays measure interaction of binding sites and transcription factors in vivo.

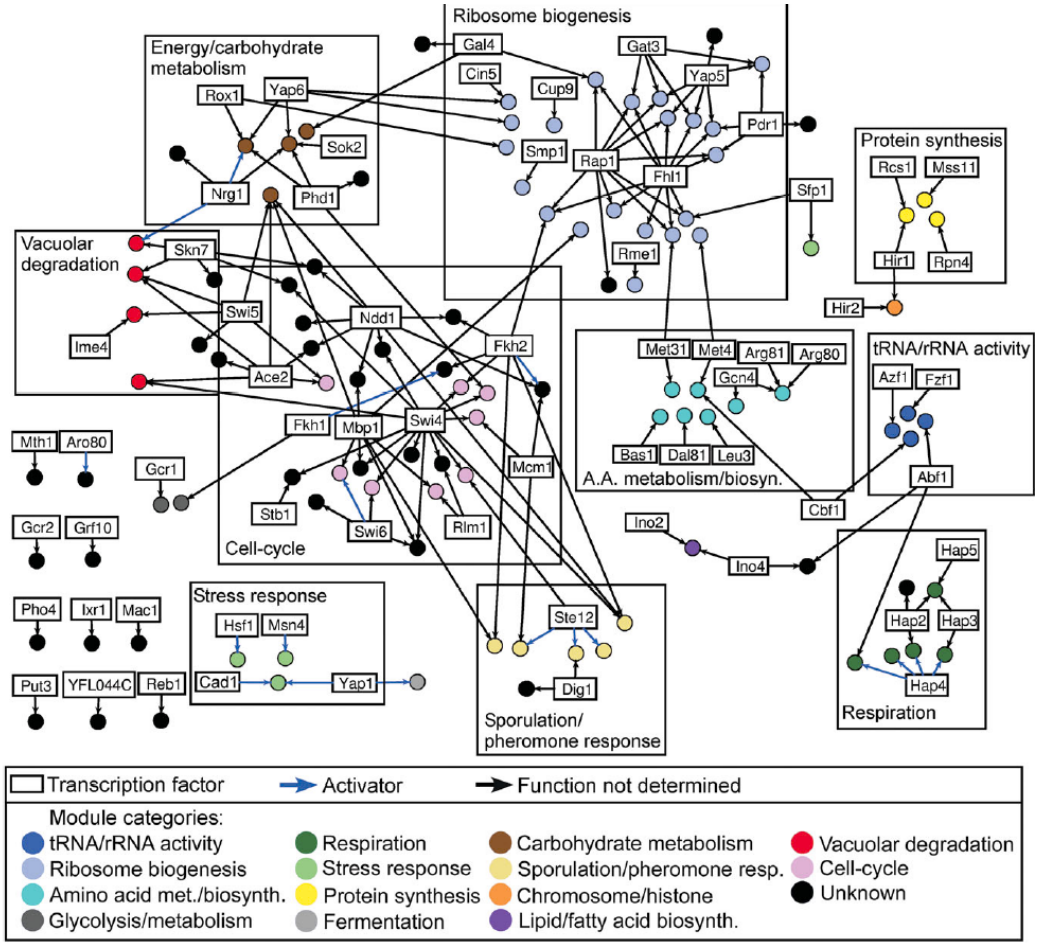
Results: 655 genes partitioned in 106 modules and 68 transcription factors working as hubs.

Validation: ChIP experiments to show activation of predicted transcription factors.



Bar-Joseph, *Nat Biotech*, 2004

Chip ChIP Networks



Scale Free Networks

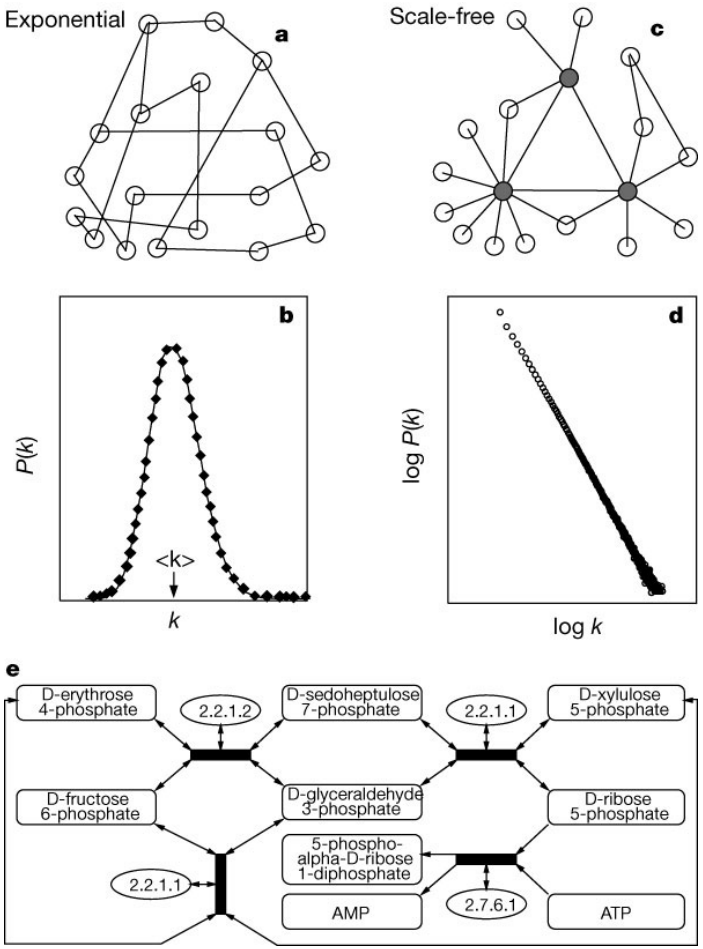
Q: Are these findings useful?

A: Yes, if we can learn something about the global structure of the network.

Scale free network: Natural interactions create robust substructures.

Method: Allow us to analyze global properties of a graph:

- ✓ Hubs/Authorities;
- ✓ Critical paths;
- ✓ Islands and holes.



Microarray Networks

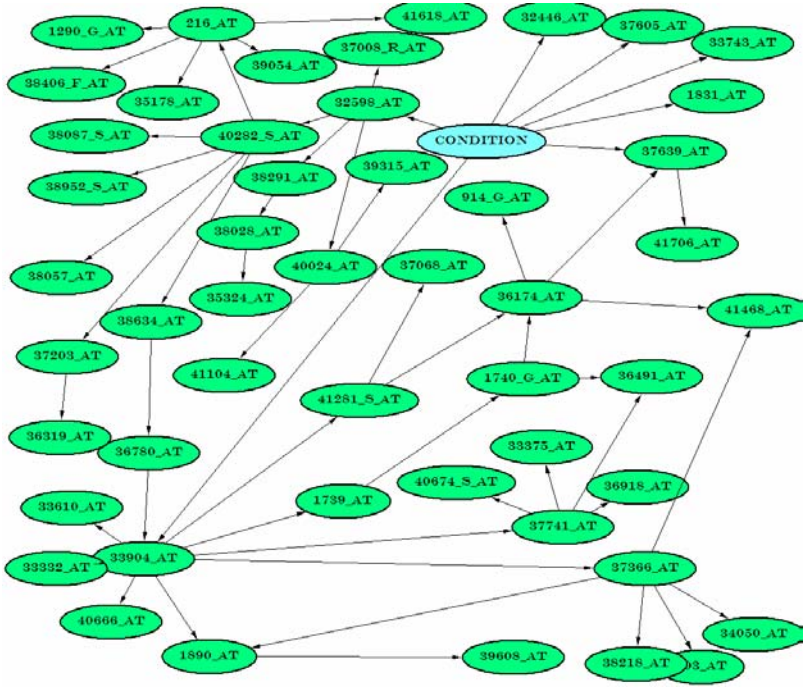
Data: 102 cases/control prostate cancer patients (Singh et al., 2002).

Task: Classification and dependency discovery.

Today: Genes are assumed independent to find best independent predictors.

Bayesian networks: discover the model of dependency and predictors.

Validation: Cross validation 92% of five fold.



Microarrays and Multiple Phenotypes

Data: 41 leukemia patients.

Measures: 72 candidate genes.

Phenotypes: 3 phenotypes.

Validation: Cross validation.

Oncogene Status: 97.56% (40)

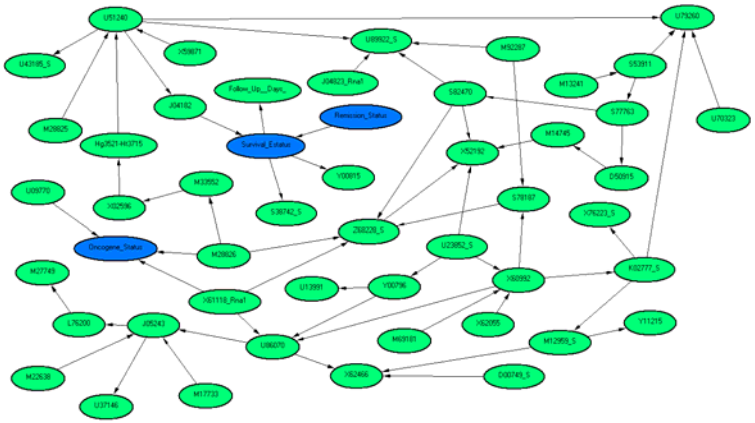
Average Distance: 0.03339

Survival Status:100% (40)

Average Distance: 0.00414

Confidence: Bayes factor -

$$\frac{P(M_1|D)}{P(M_2|D)}$$



Oncogene_Status

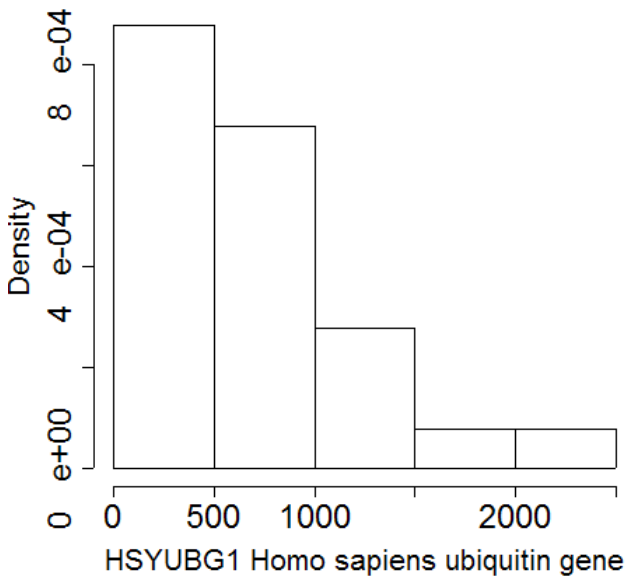
X61118_Rna1	M28826	U09770		1
U09770	X61118_Rna1	M28826	S53911	7
U09770	X61118_Rna1	M28826	M69181	56
X61118_Rna1	M28826	M17733		63
X61118_Rna1	M28826	S77763		315
U09770	X61118_Rna1	M28826	U43185_S	447
U09770	X61118_Rna1	M28826	J05243	447
X61118_Rna1	M28826			973
X61118_Rna1	M28826	S38742_S		1016
X61118_Rna1	M28826	J05243		1534
U09770	X61118_Rna1	M28826	M17733	1804
U09770	X61118_Rna1	M28826	Z68228_S	1807
U09770	X61118_Rna1	M28826	J04823_Rna1	3558
U09770	X61118_Rna1	M28826	U37146	3564
U09770	X61118_Rna1	M28826	X62055	3564
U09770	X61118_Rna1	M28826		3570
X61118_Rna1	M28826	Y11215		3933
X61118_Rna1	M28826			4254
U09770	X61118_Rna1	M28826	Survival_Estatus	7369
X61118_Rna1	M28826	Survival_Estatus		11093

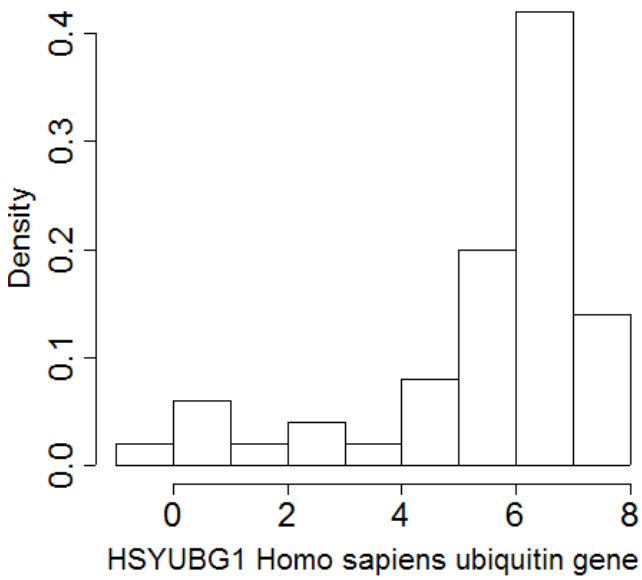
Distributional Assumptions

Microarrays produce data with skewed distributions.

Log-normal: take the logarithm, data are normal.

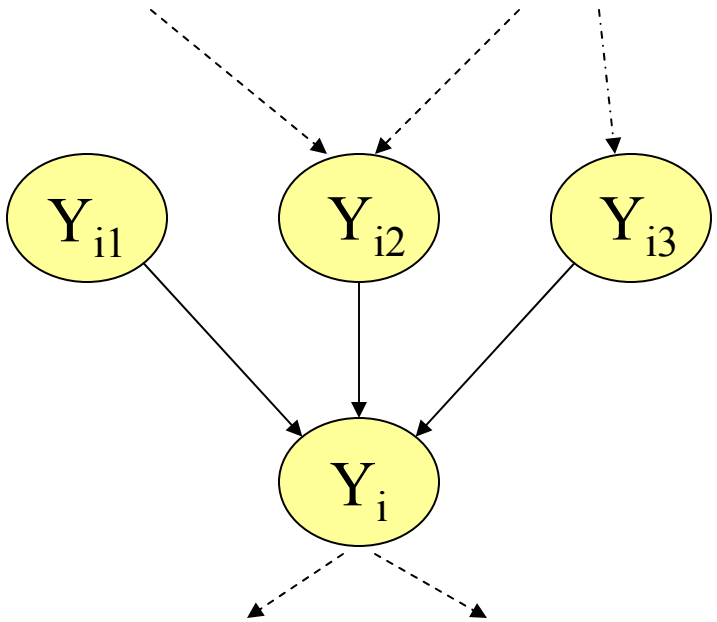
Gamma: they remain asymmetrical (exponential).





Generalized Gamma Networks

- Model gene expression data by Gamma distributions;
- Encode general non linear dependencies



$$\mu(pa(y),\theta) = \mu(\eta(pa(y),\theta))$$

$$\eta = \theta_o + \sum_j \theta_j f(y_{ij})$$

Can choose different link functions

$$\mu = \eta; \quad \eta = \theta_o + \sum_j \theta_j y_{ij}$$

$$\mu = 1/\eta; \quad \eta = \theta_o + \sum_j \theta_j / y_{ij}$$

$$\mu = \exp(\eta); \quad \eta = \theta_o + \sum_j \theta_j y_{ij};$$

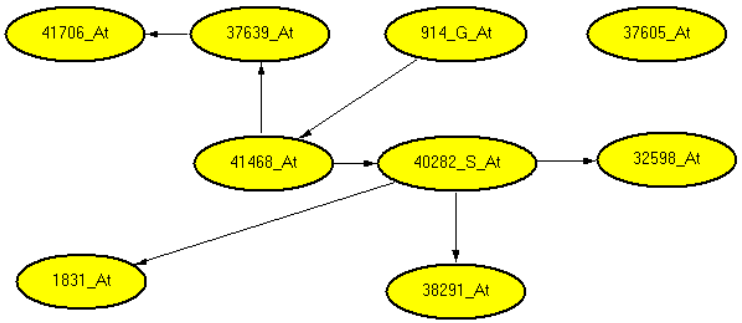
$$\eta = \theta_o + \sum_j \theta_j \log(y_{ij})$$

Differential Analysis

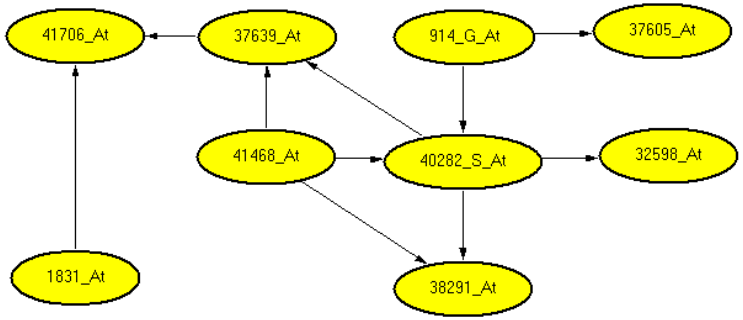
Data: Prostate cancer dataset.

Rationale: Cancer is a disease of control. Can we differentiate which control mechanism change between normal and cancer rather than genes?

Design: Learn two networks, one from normal and one from tumor specimens, and compare their dependency structure.



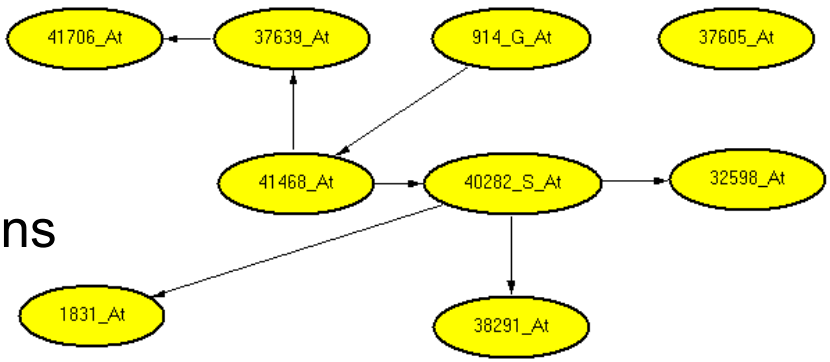
Normal specimens



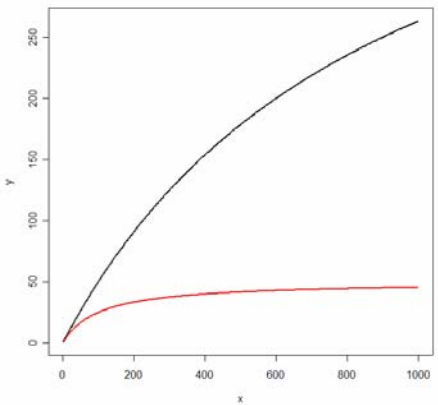
Tumor specimens

Functional Differences

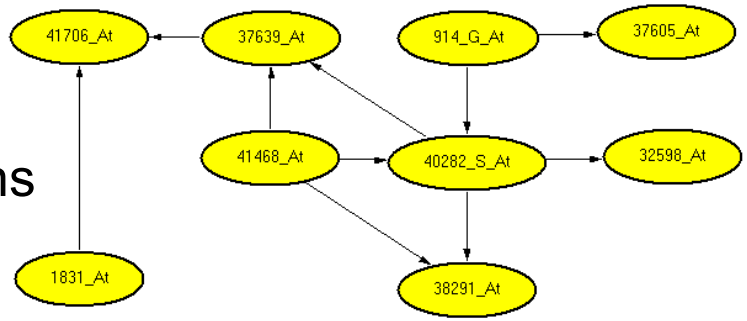
Normal specimens



$\mu=1/(.01+1.8/y.40282)$



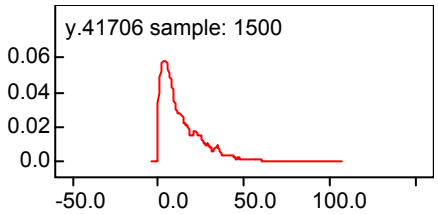
Tumor specimens



$\mu=1/(0.02+2/y.40282)$

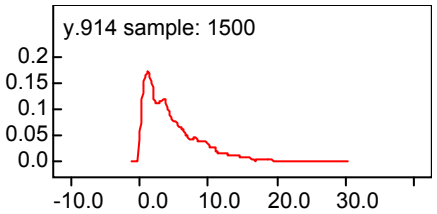
32598: gene with putative growth and transcription regulation functions

Normal Samples



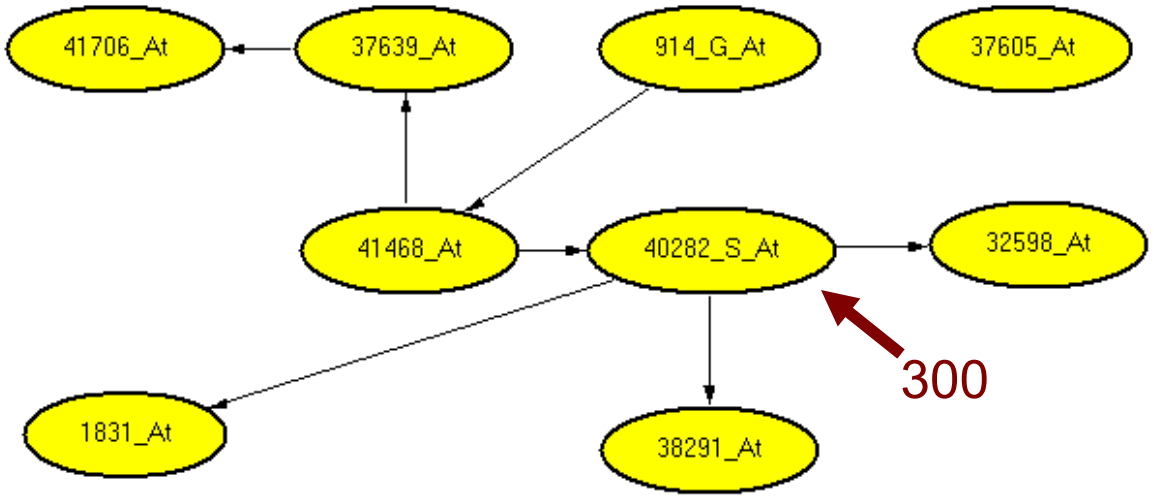
Mean=16

Tumor differentiation



Mean=1.3

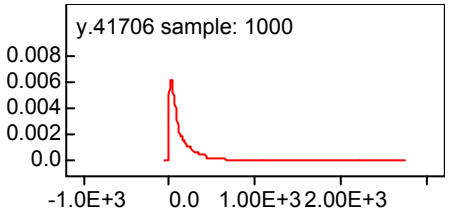
Oncogene



Growth/differentiation factors

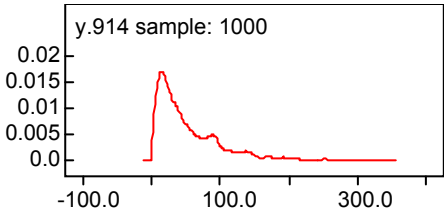
Observe 40282_S_AT=300 (average value in normal specimens).
Gene supposed to have a role in immune system.

Tumor Samples



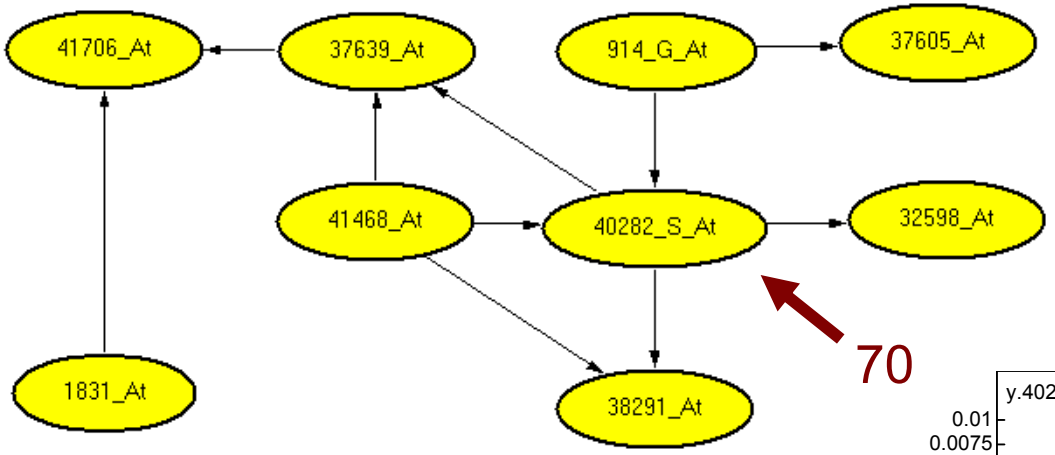
Mean=450

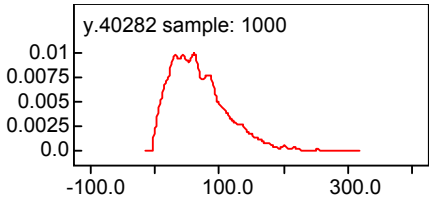
Tumor differentiation



Mean=66

Oncogene





Changes in 40282_S_AT determine changes in tumor markers.

SNPs Networks

Goal: Overt stroke in sickle cell anemia patients.

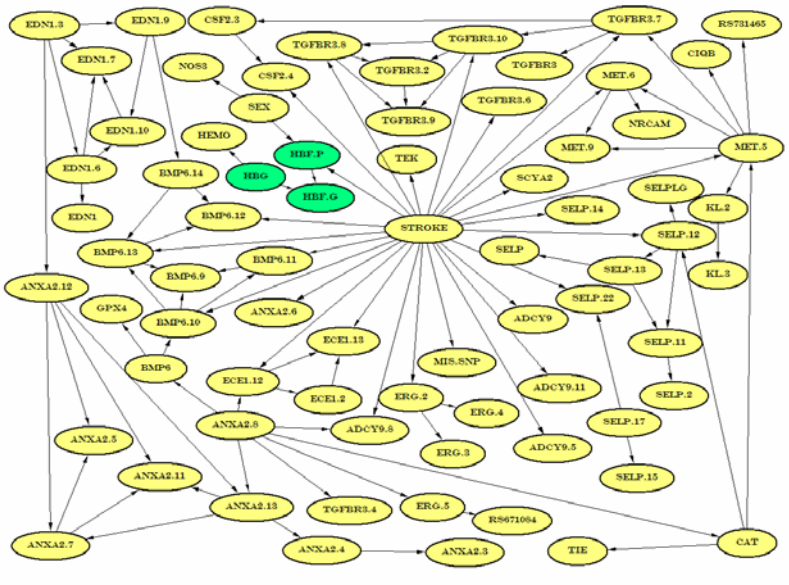
Subjects: 1392 case/control sickle cell anemia patients.

Genotypes: 80 candidate genes for approx 250 SNPs;

Risk factors: α-Thalassemia, clinical history, age, gender.

Validation: Stroke prediction of 114 subjects from a different population.

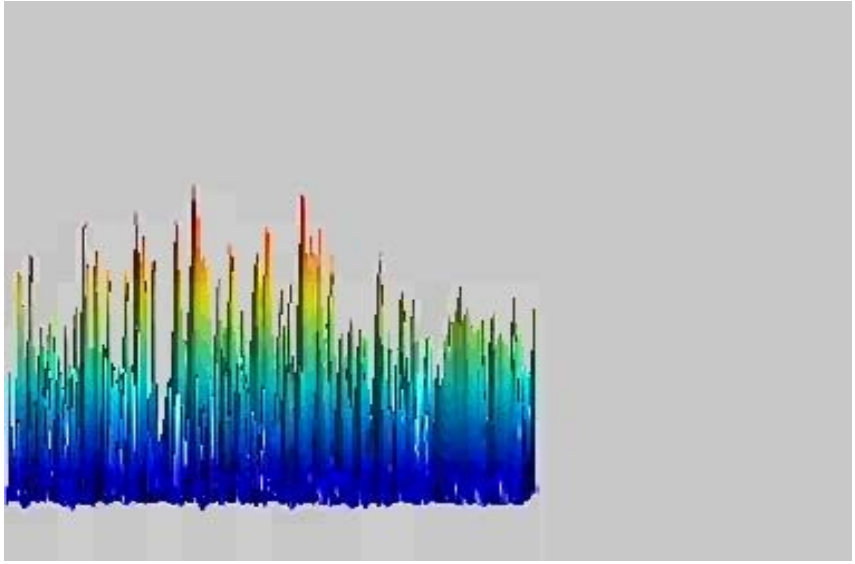
Results: 98.5% accurate (100% true positive rate).



Sebastiani et al, *in press*, 2004

Seldi-Time Of Flight Proteomics

Automation



Proteomic Data Streams

Proteomic Networks

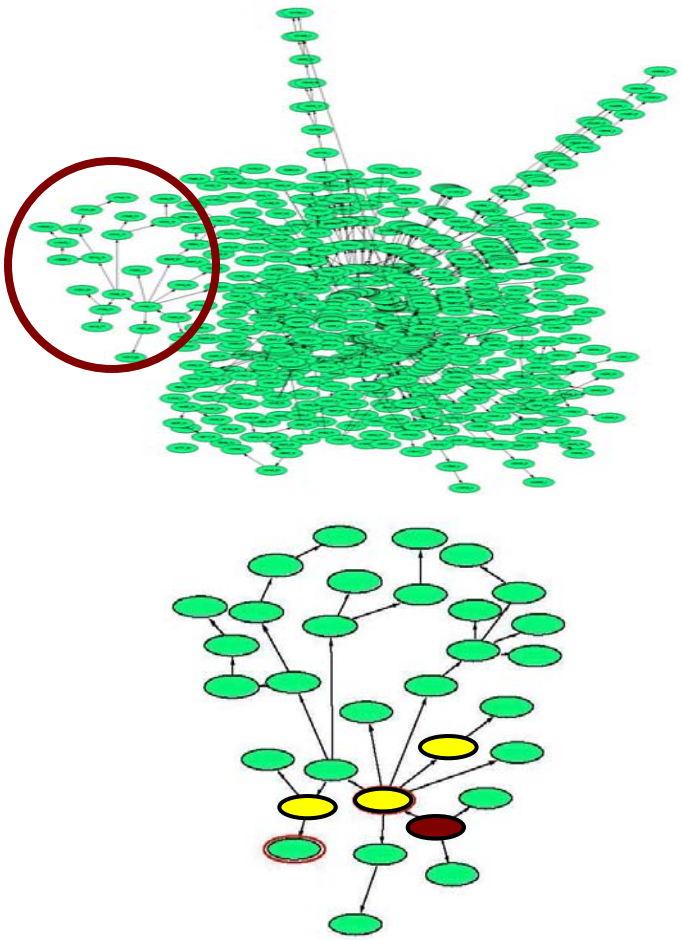
Domain: MDS (pre-leukemia).

Design: Over 100 case control patients to identify specific markers in peripheral blood.

Challenge: Identify proteins.

Model: A Bayesian network discovering dependencies and identify same/different proteins and controllers.

Results: G. Alterovitz, May 11th, Session 217-8 3:00pm, Seminar Room 217.



With G Alterovitz and T Libermann

Integrate SNPs and Proteins

Task: Find pathogenic SNPs with no phenotypes.

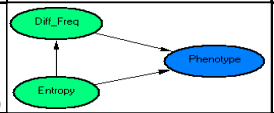
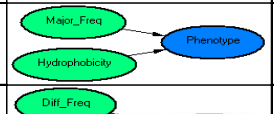
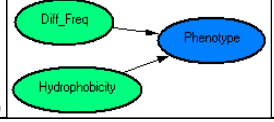
Rationale: Test SNPs that are more likely pathogenic.

Training set: Microbial data of aminoacid substitution cause of phylogenetic, biochemical or structural changes.

Test set: Human dataset of allele variances from OMIM.

Task: Find changes that induce pathogenic phenotype.

Results: less than 10% FPR.

Training set			Bayesian network	Bayes Factor
dataset	class 0	class 1		
LacI(1)	WT+Int (2940)	Sig (804)		5.45E+13
LacI(2)	WT (2710)	Sig (804)		6.79E+14
LacI(3)	WT (2710)	Int+Sig (1034)		1.38E+09
T4 lysozyme(1)	WT+Int (1388)	Sig (237)		6.66E+02
T4 lysozyme(2)	WT (1115)	Sig (237)		7.44E+10
T4 lysozyme(3)	WT (1115)	Int+Sig (510)		3.42E+04
LacI(1)+T4 lysozyme(1)	WT+Int (4328)	Sig (1041)		1.48E+21
LacI(2)+T4 lysozyme(2)	WT (3825)	Sig (1041)		4.17E+20
LacI(3)+T4 lysozyme(3)	WT (3825)	Int+Sig (1544)		2.17E+09

Cai et al, *Hum Mut*, 2004

Take Home Messages

Summary:

- Microarrays offer the opportunity to observe new phenomena, not only more genes.
- The opportunity is to identify global structures of control, that cannot be observed in isolation (Holistic vs Reductionistic).
- To grasp the opportunity, we need new, improved methods, and a new way to look at phenomena (Quantitative vs Qualitative).
- To prove our results, we need also a new standard of proof, adequate for the new attitude (Predictive vs Descriptive).

Challenges:

- Networks discover not only information but also domain specific emerging semantics (what does a link mean?).
- How do we translate these discoveries to humans?