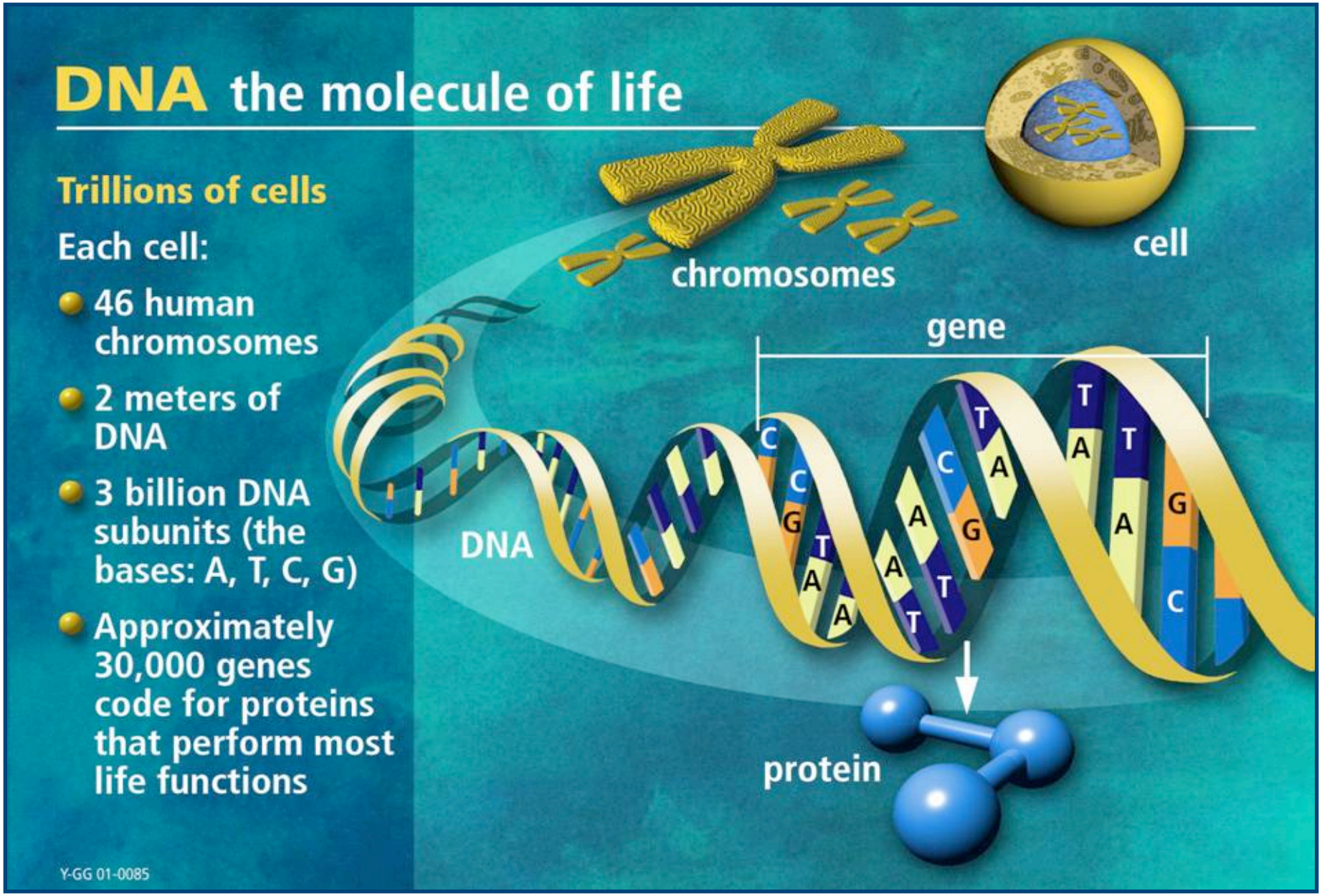
Functional Genomics Lecture

ECE Genomic Signal Processing (Dr. Rosen)

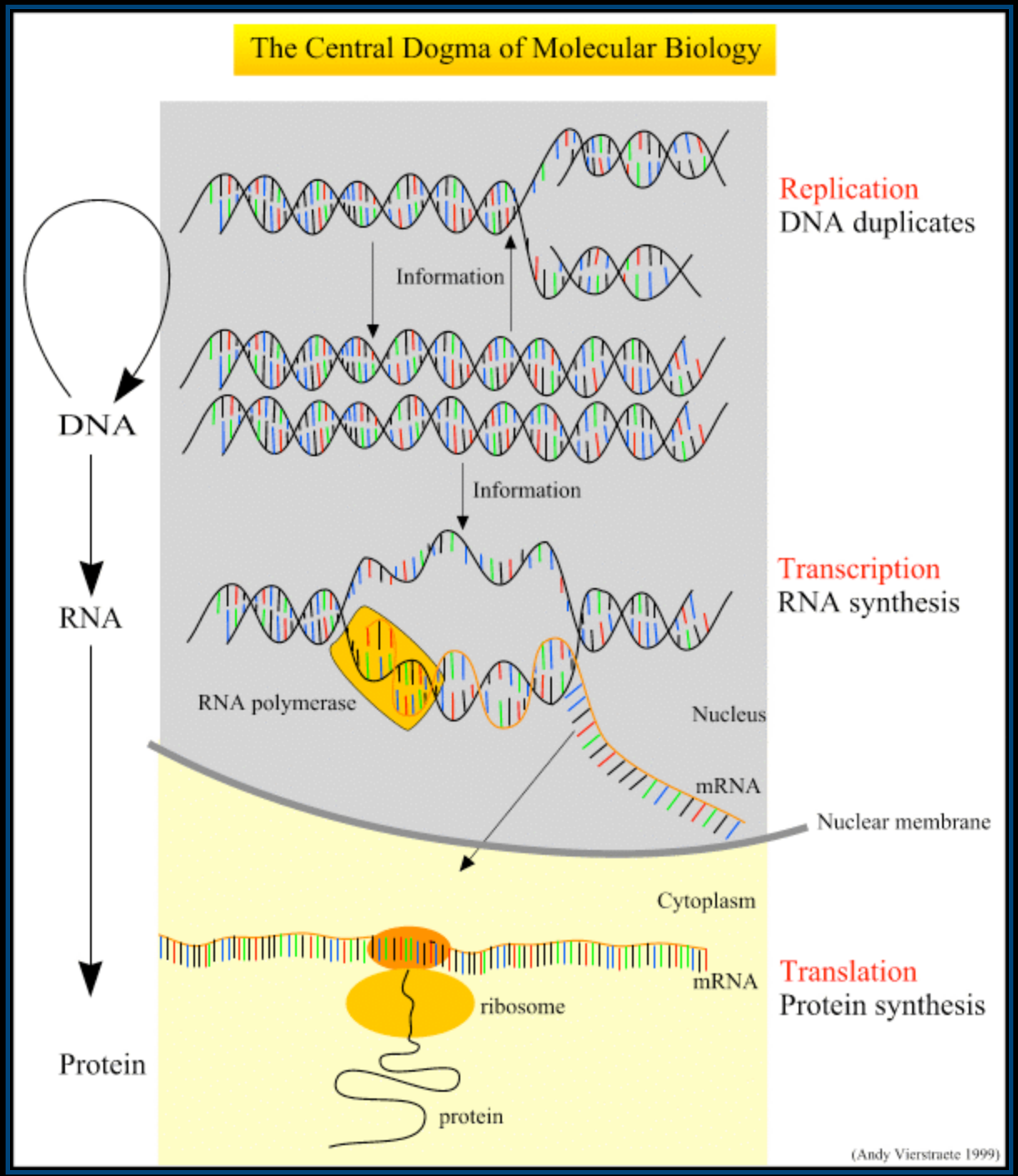
Tuesday, Apr. 21, 2009, 6-8 pm, 616 Bossone

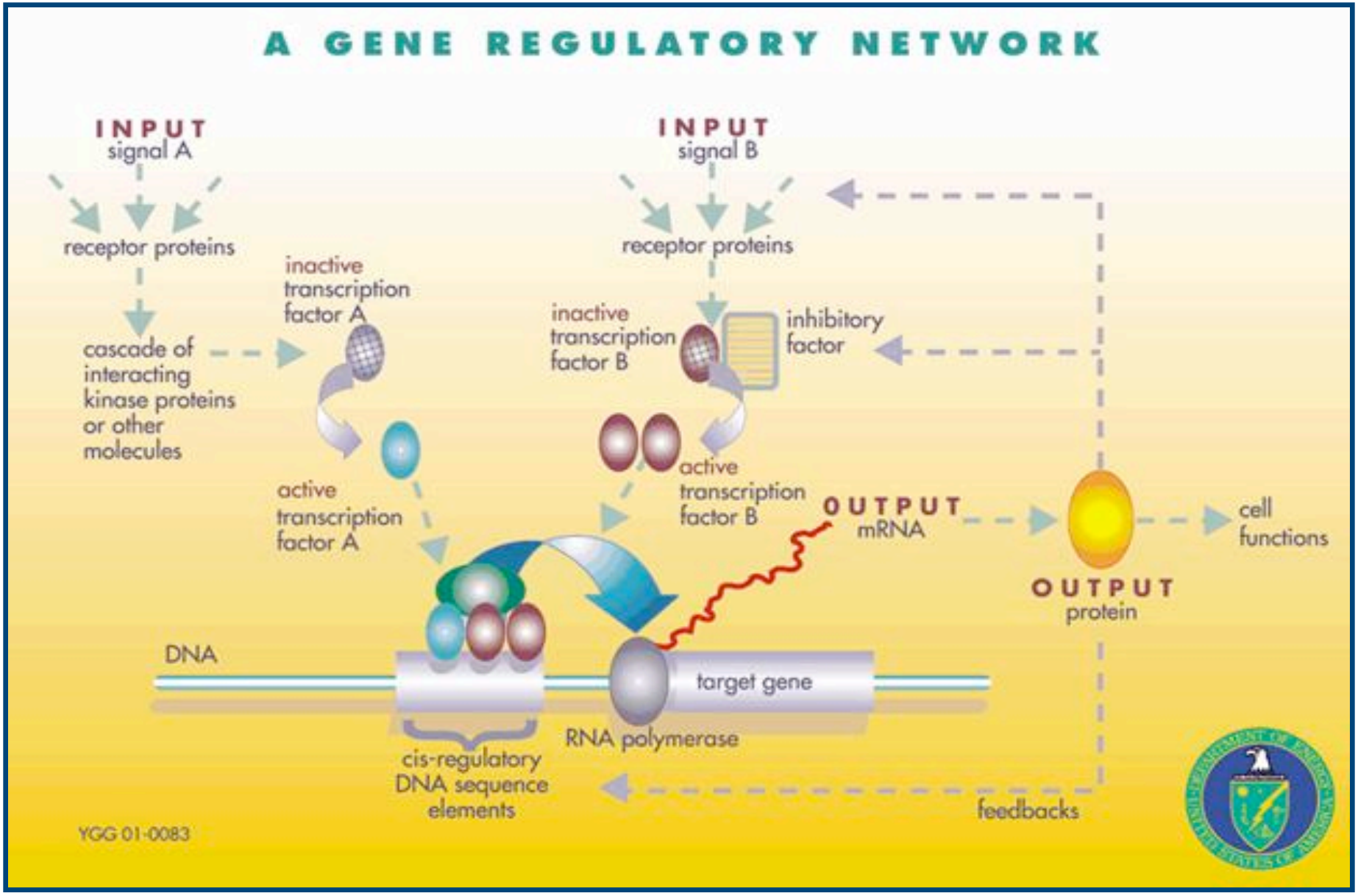
Bahrad A. Sokhansanj, PhD

School of Biomedical Engineering, Science & Health Systems



U.S. Department of Energy Human Genome Program, http://www.ornl.gov/hgmis.


The Central Dogma of Molecular Biology
DNA
RNA
Protein
Information
Information
Replication
DNA duplicates
Transcription
RNA synthesis
RNA polymerase
Nucleus
mRNA
Nuclear membrane
Cytoplasm
mRNA
ribosome
protein
Translation
Protein synthesis
(Andy Vierstraete 1999)



U.S. Department of Energy Genomics:GTL Program, http://doegenomestolife.org.

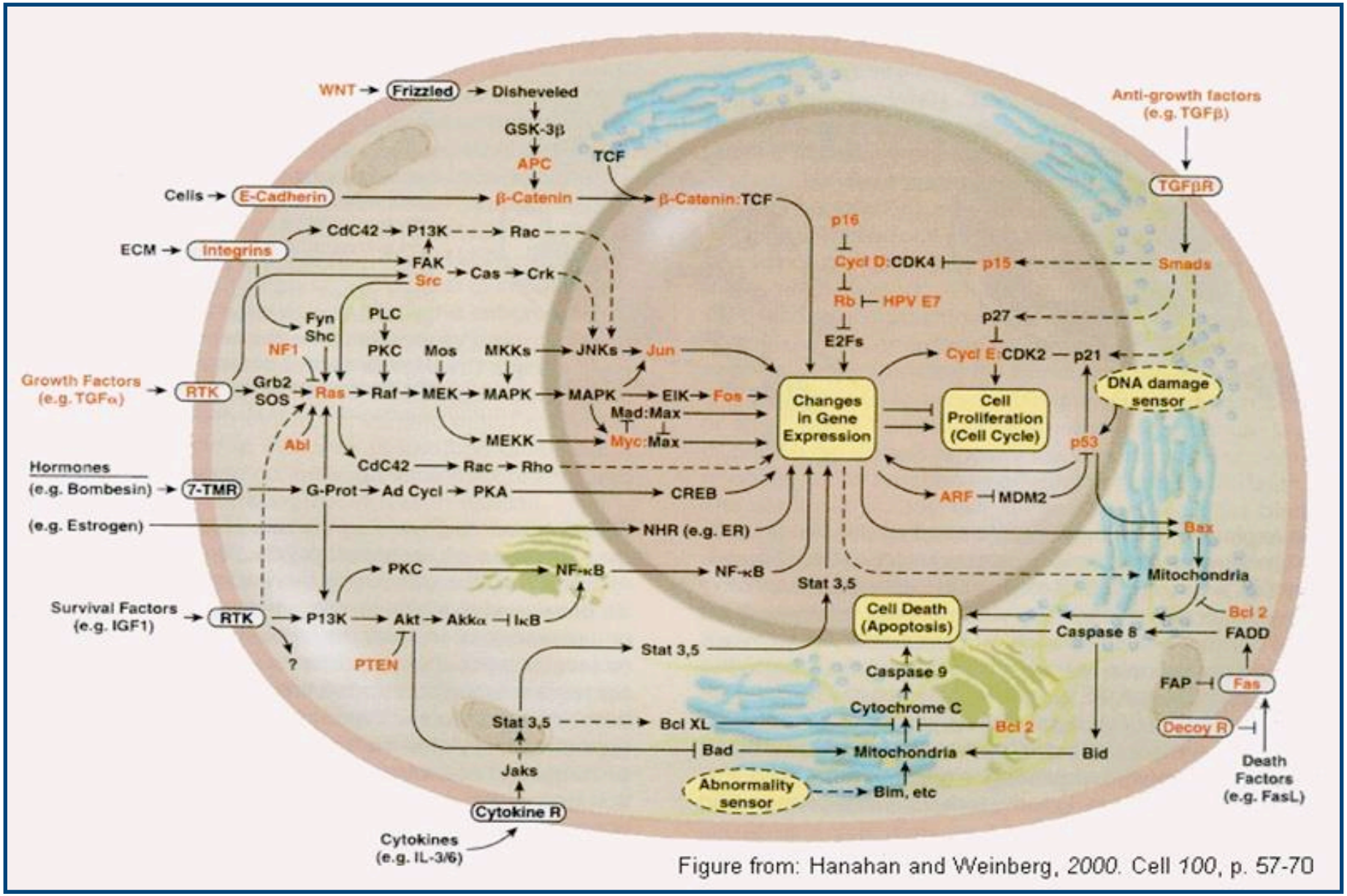
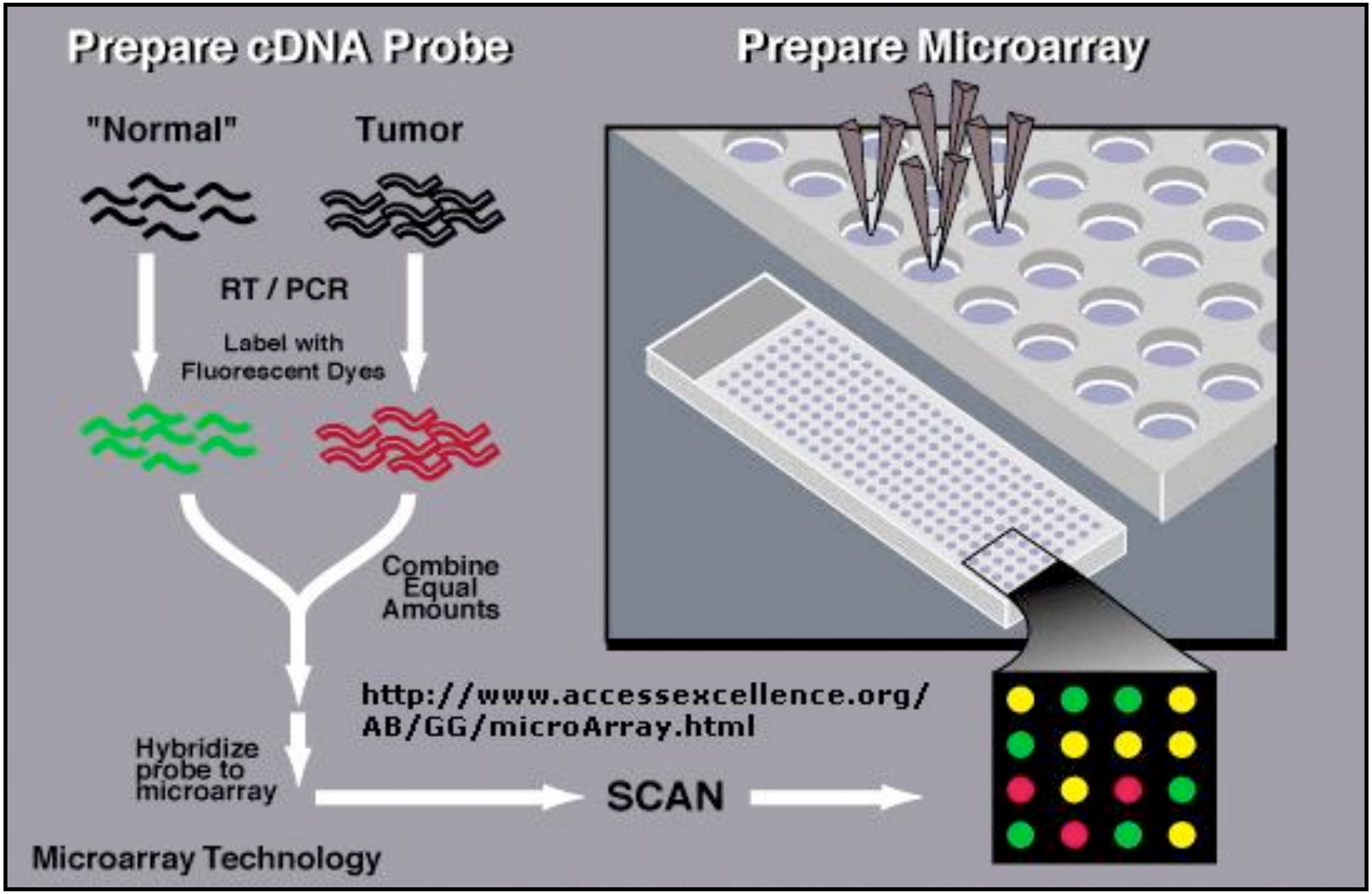
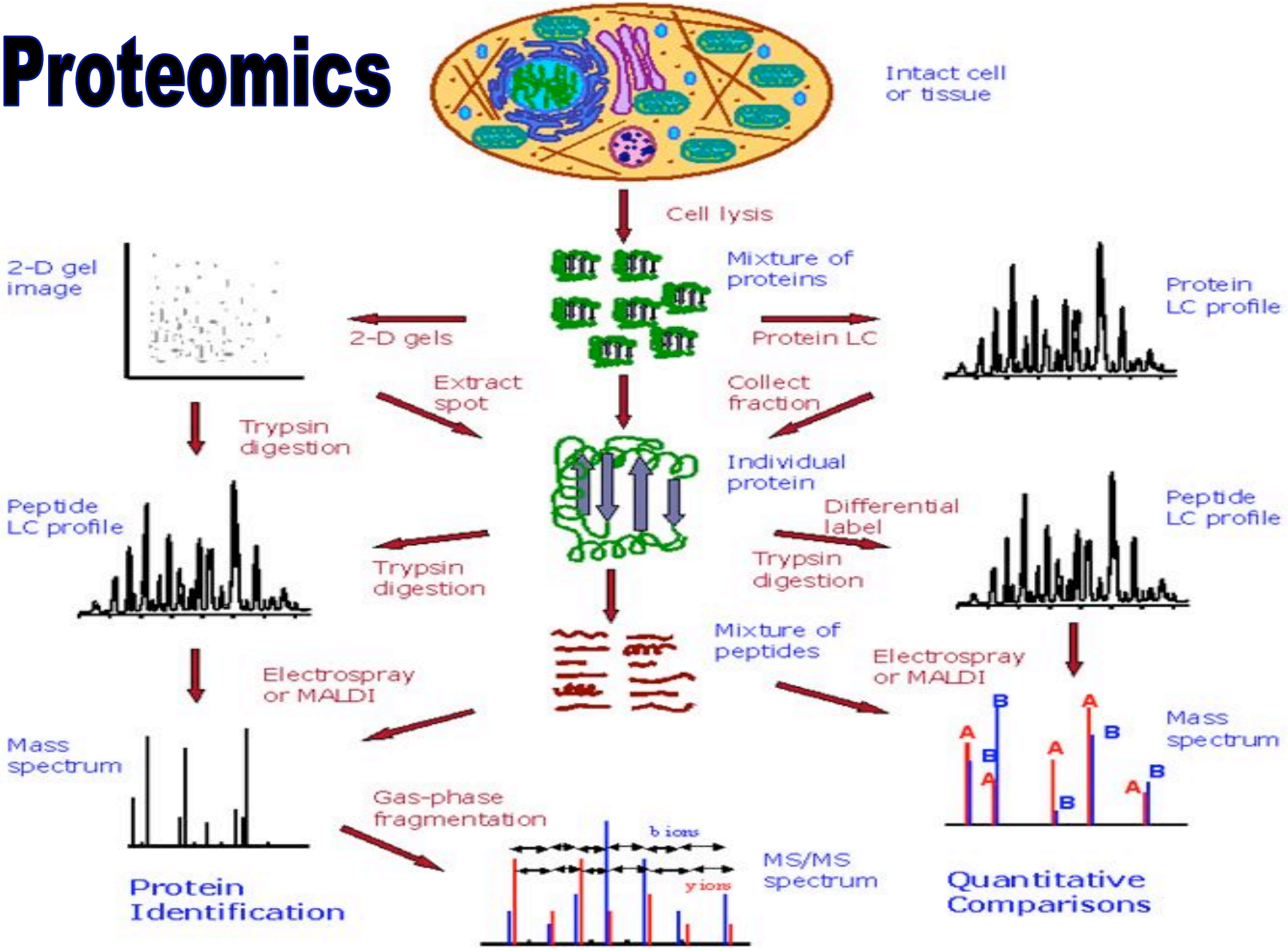


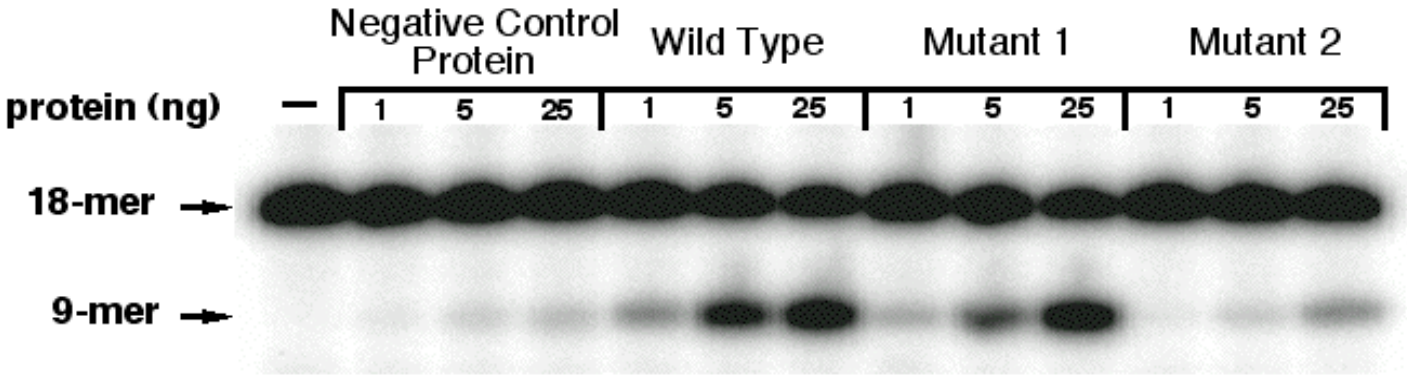
Figure from: Hanahan and Weinberg, *2000*. Cell *100*, p. 57-70


Prepare cDNA Probe
Prepare Microarray
"Normal"
Tumor
RT / PCR
Label with
Fluorescent Dyes
Combine
Equal
Amounts
http://www.accessexcellence.org/
AB/GG/microArray.html
Hybridize
probe to
microarray
SCAN
Microarray Technology

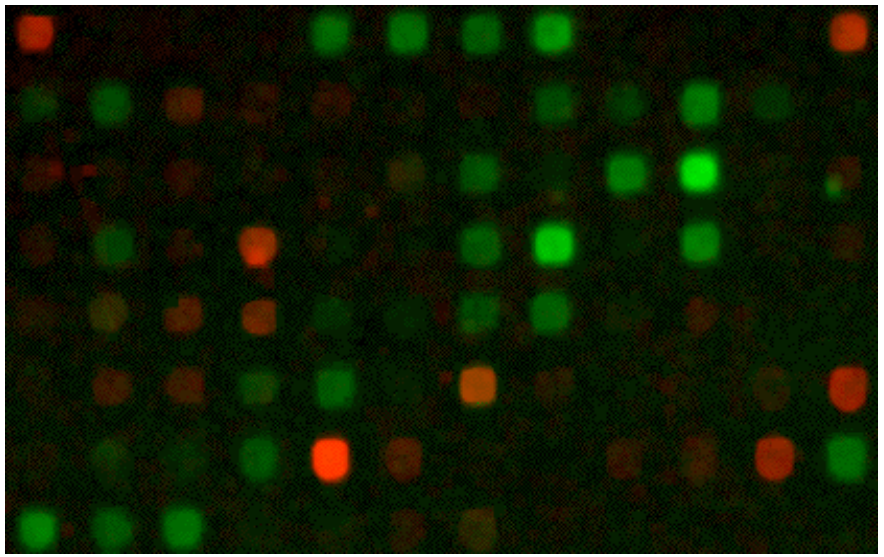
Proteomics


Intact cell or tissue
Cell lysis
Mixture of proteins
2-D gel image
2-D gels
Protein LC
Protein LC profile
Extract spot
Collect fraction
Trypsin digestion
Individual protein
Peptide LC profile
Differential label
Trypsin digestion
Trypsin digestion
Peptide LC profile
Mixture of peptides
Electrospray or MALDI
Electrospray or MALDI
Mass spectrum
Mass spectrum
Gas-phase fragmentation
b ions
y ions
MS/MS spectrum
Protein Identification
Quantitative Comparisons

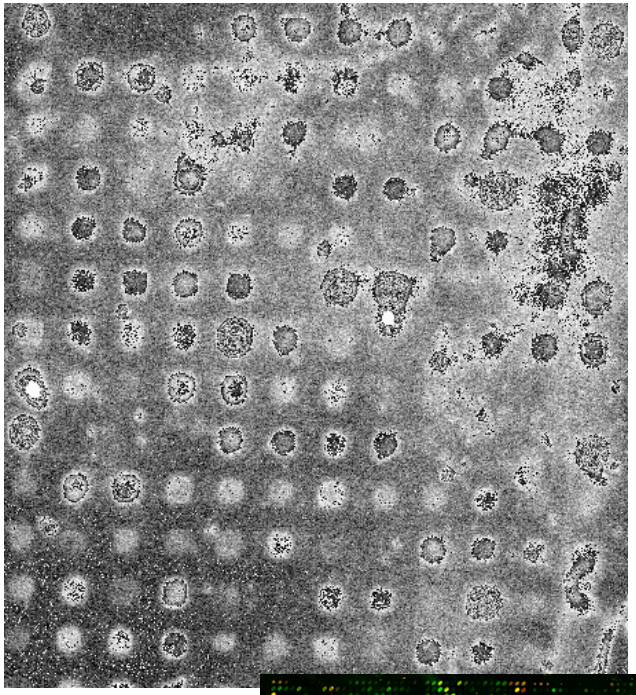
The Reality of Biological Data

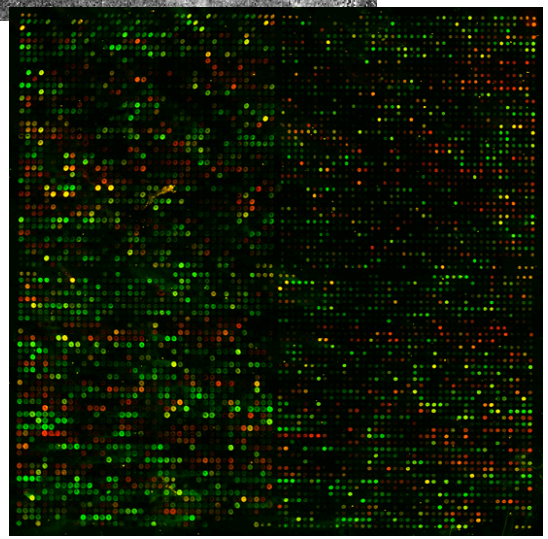


Can we use these data in a numerical model?



Microarrays Produce Ugly Data

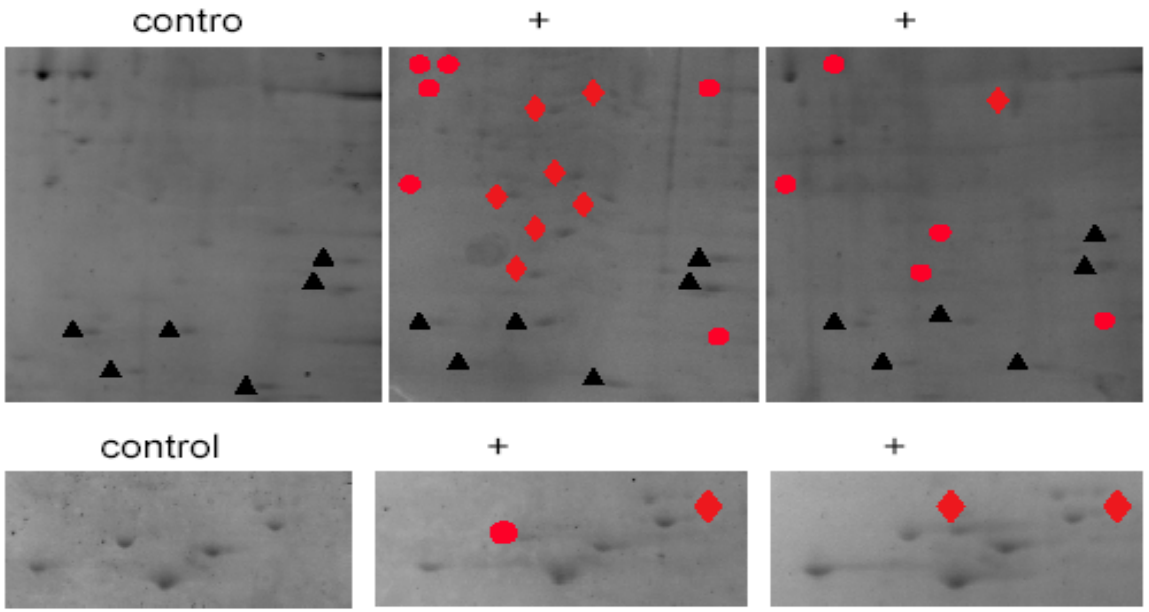




- Microarray fabrication is inconsistent: experiments have poor repeatibility
  - size, shape, and alignment of spots varies from array to array
  - defects in the slide and different washing protocols result in intra- & inter-slide variations in background
- Results are highly sensitive to image analysis
  - spot recognition, intensity quantification and normalization, background subtraction
- DNA chips are more consistent, but still not perfect - and they cost much more

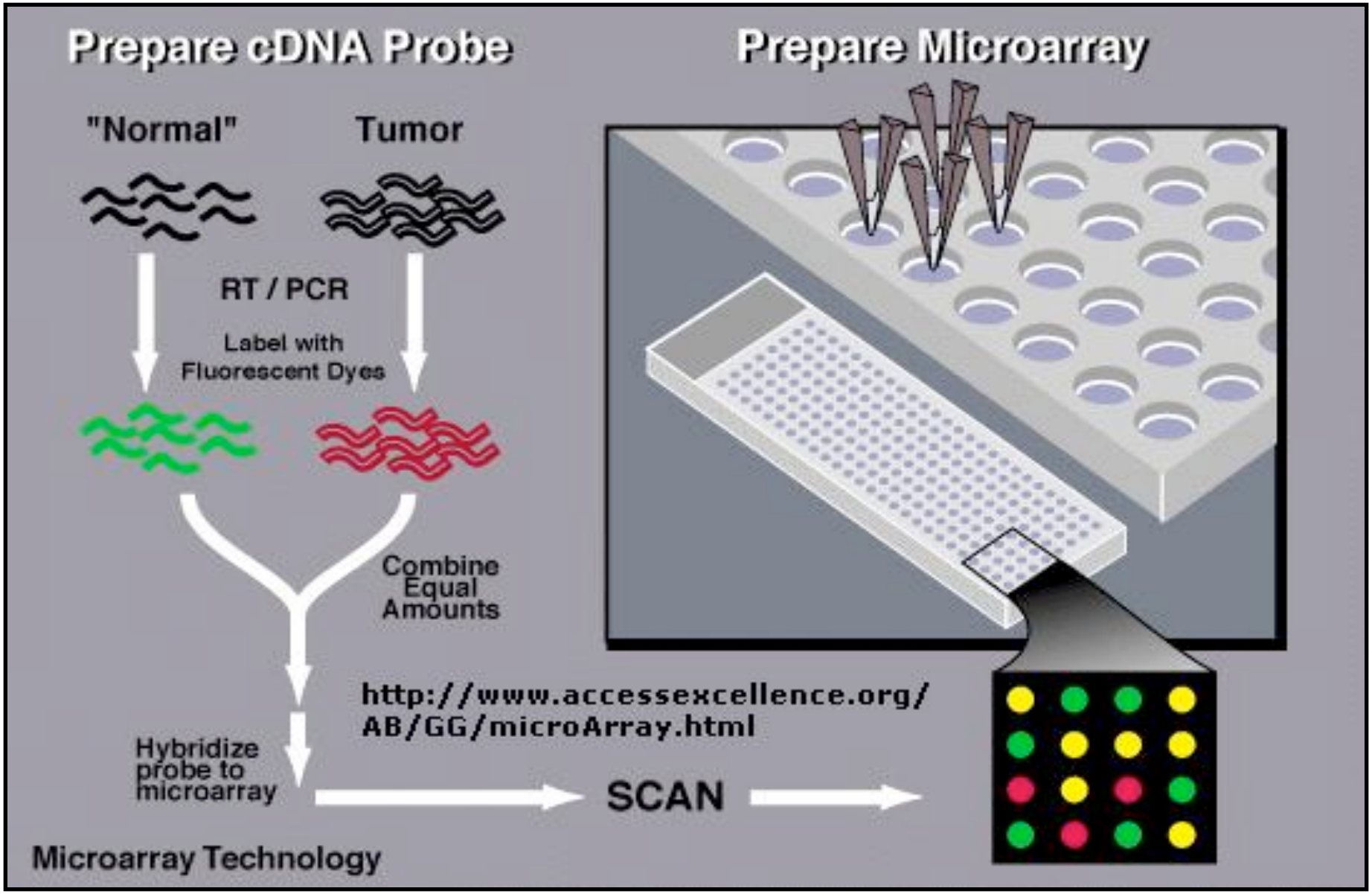
More Ugliness – Gels for Proteomics

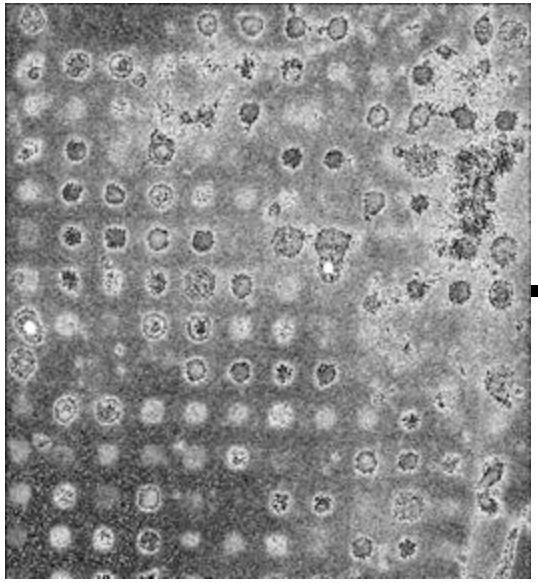
Monocyte cell line exposed to Y.pestis and Y.pseudotuberculosis

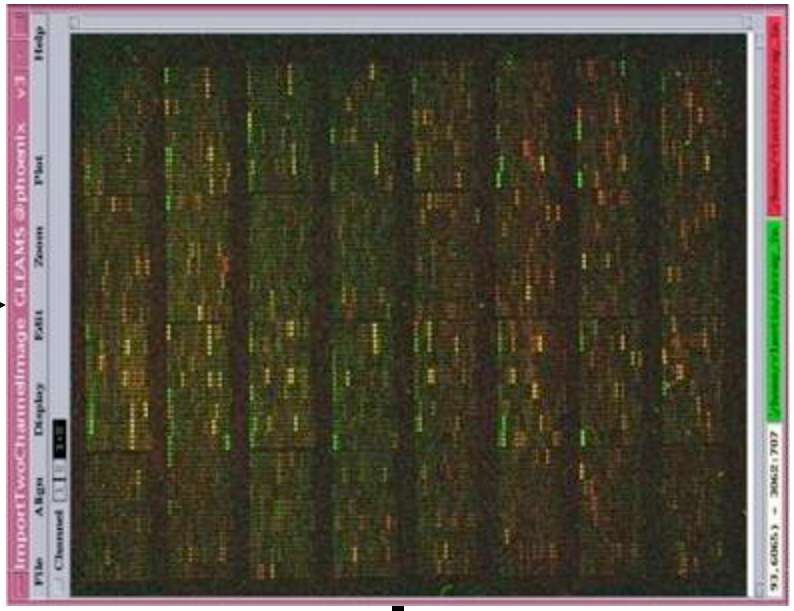


2DE of 4 hr exposures, MOI 5:1, control (no pathogen). Protein identification in progress.
▲ (unchanged), ◆ (upregulated), ● (downregulated). Two separate areas of gels

DIGE and MS are improvements, but still, at best, semi-quantitative.


Prepare cDNA Probe
Prepare Microarray
"Normal"
Tumor
RT / PCR
Label with
Fluorescent Dyes
Combine
Equal
Amounts
http://www.accessexcellence.org/
AB/GG/microArray.html
Hybridize
probe to
microarray
SCAN
Microarray Technology

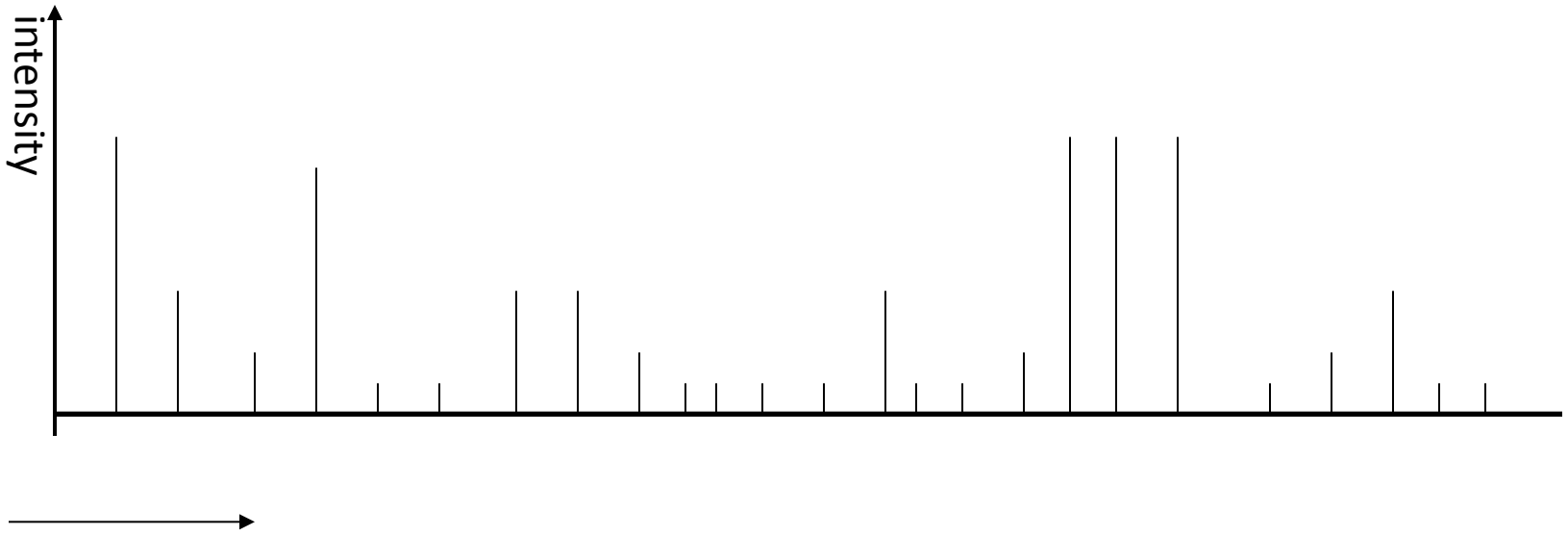




	HR 1	*HR 4*	*HR 10*
YPO3718/pgi	-1.609	-1.114	1.089
YPO2686/pgm	-1.157	-1.260	-1.172
YPO2995/arr	-1.861	1.094	-1.083
YPO0078/pfkA	-2.191	-1.212	-1.134
YPO0920/fbaA	-1.536	1.063	1.124
YPO1254/bglA	1.304	1.147	1.094
YPO0166/none	1.666	1.075	1.215
YPO2977/glk	-1.174	-1.308	-1.111
YPO1135/galM	2.292	-1.020	1.079

Information Derivable from Chip Data

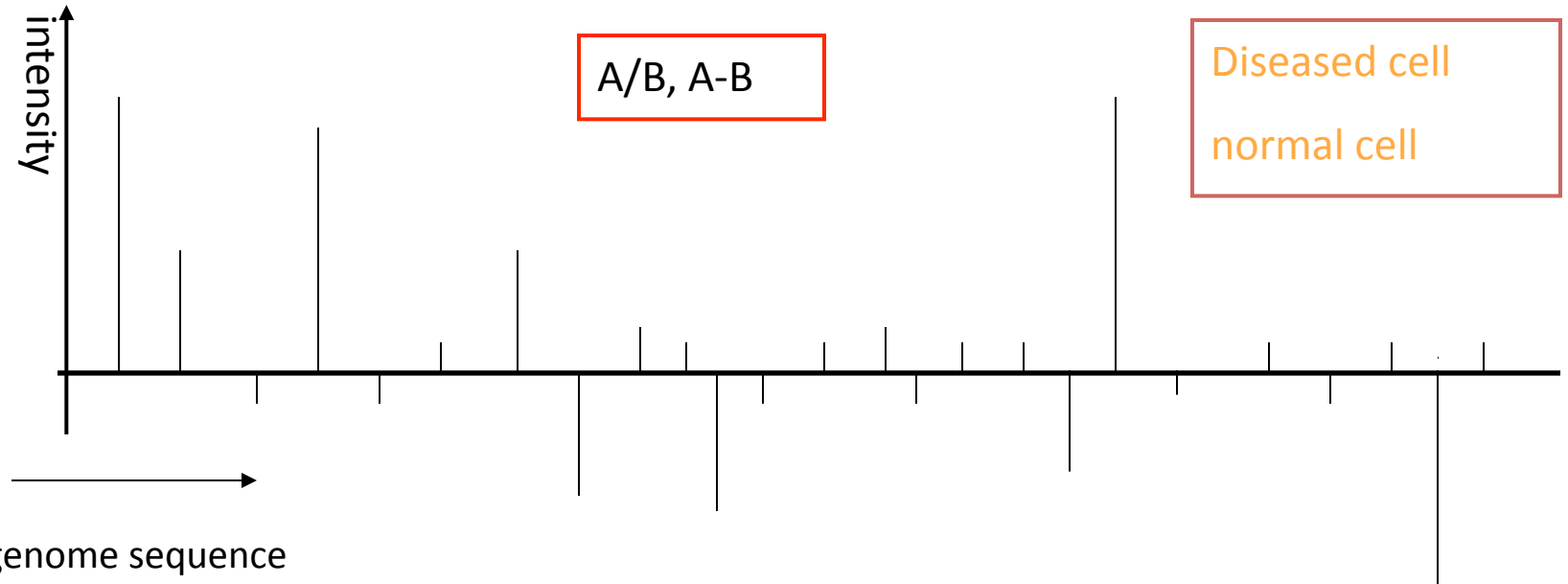
- By observing chip data, one can infer which genes are highly expressed or not expressed, or in general the relative expression levels of all genes



Inference: genes x, y, z are highly expressed under conditions W while genes a, b, c are not expressed

Information Derivable from Chip Data

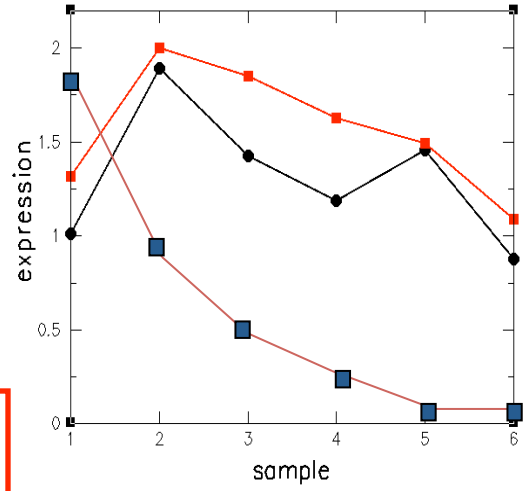
- By comparing gene expression levels under two conditions, one can infer which genes' expression levels are affected



Inference: gene X is significantly more highly expressed in diseased cell than in normal cell; hence gene X could potentially serve s a marker of the disease – differentially expressed genes

Information Derivable from Chip Data

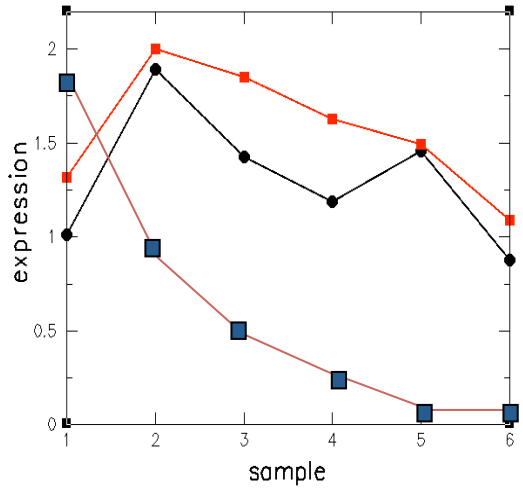
- By observing gene expression levels collected at different time points after a particular stimulus, one can infer how a gene's expression level changes with time



Inference: as the disease progresses, gene X's expression decreases; hence ….

Information Derivable from Chip Data

- By observing expression levels of two genes collected at different time after a particular stimulus, one can infer they have similar or different expression patterns

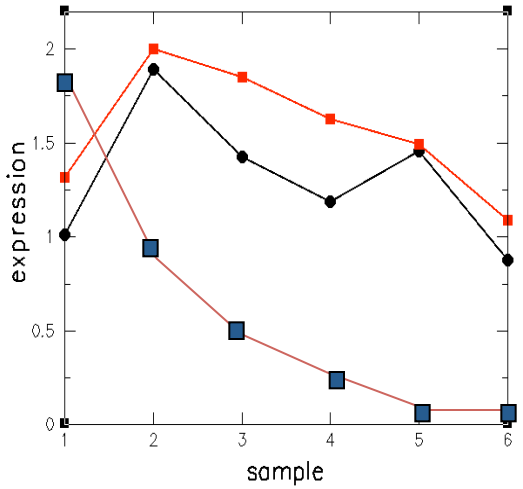


Inference: genes with similar expression patterns might be functionally related, e.g., working in the same pathway – co-expressed genes -> co-regulated

Information Derivable from Chip Data

- By comparing expression patterns of gene A collected when gene B is functioning and not functioning (e.g., knockout or mutation), one could possibly derive gene B's effect on gene A

Inference: genes A and B may interact directly or indirectly, or even B is the cause of A's altered expression patterns – interaction or causality relationship



Molecular Signatures – Genome, Proteome, Metabolome, Glycome, …

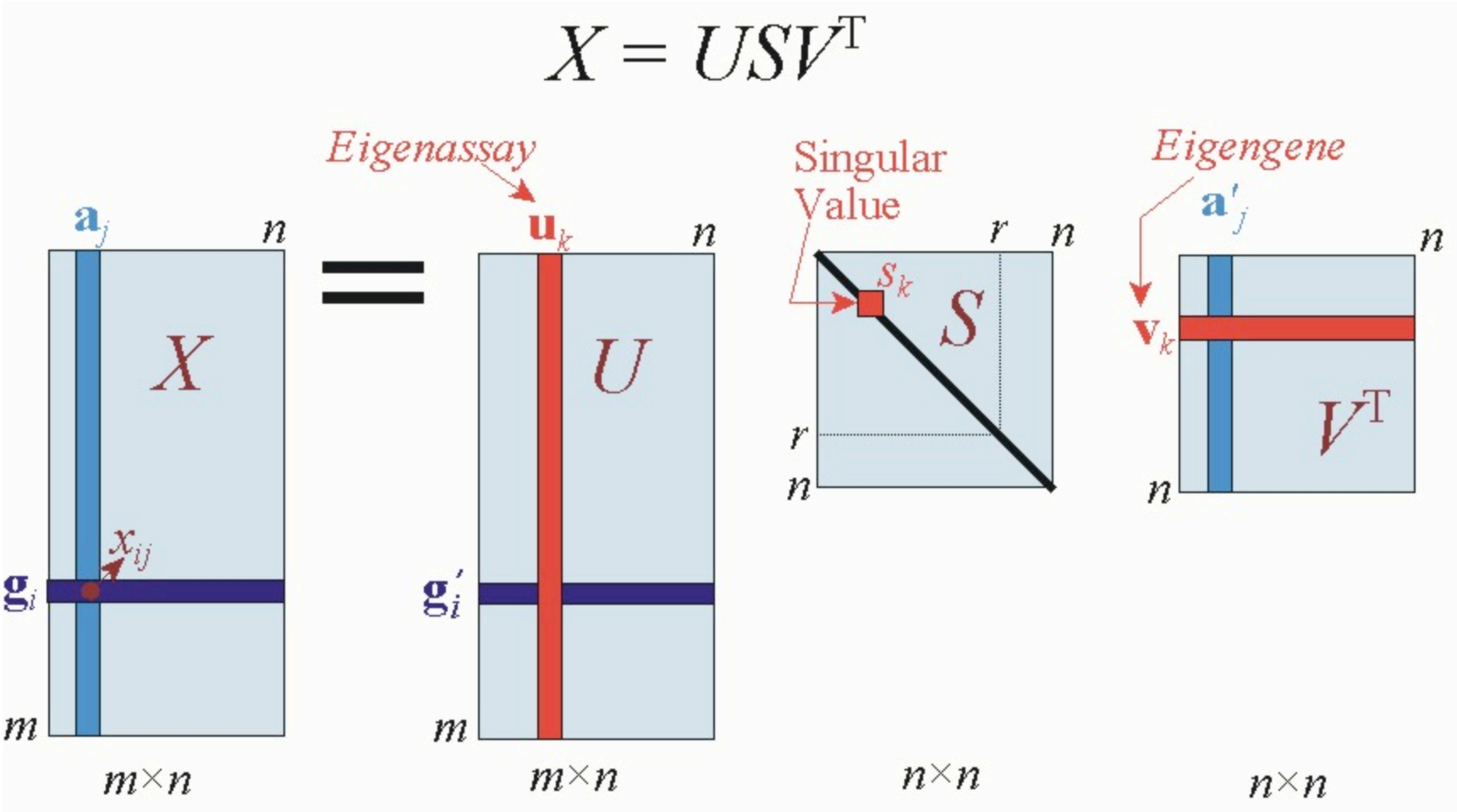
- New technologies allow large scale, parallel measurement of cell *state*:
  - transcription (mRNA expression; gene chips, RT-PCR)
  - translation (protein expression; gels, MS, protein chips)
  - protein modification (gels, MS)
  - protein-protein interaction (2-hybrid, protein chips, MS)
  - metabolites (MS, NMR)
  - carbohydrates and glycosylation (MS, ?)
  - (also large scale phenotypic changes)

- At first order, we can either/both
  - identify groups of cells/tissues/populations that share common patterns
  - detect patterns of gene/protein/metabolites/etc. that correlate with previously identified phenotypic groups

Many Ways to Identify Signatures

- Identifying major "components" of variation (potentially something that has to be removed from data, such as a fundamental difference between sampled groups)
  - singular value decomposition (SVD), principle component analysis (PCA), etc.

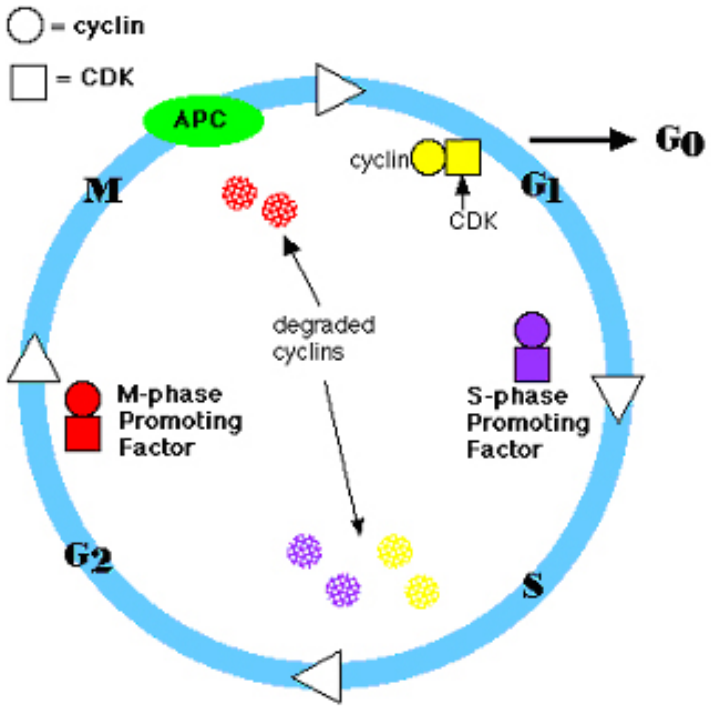
Singular Value Decomposition



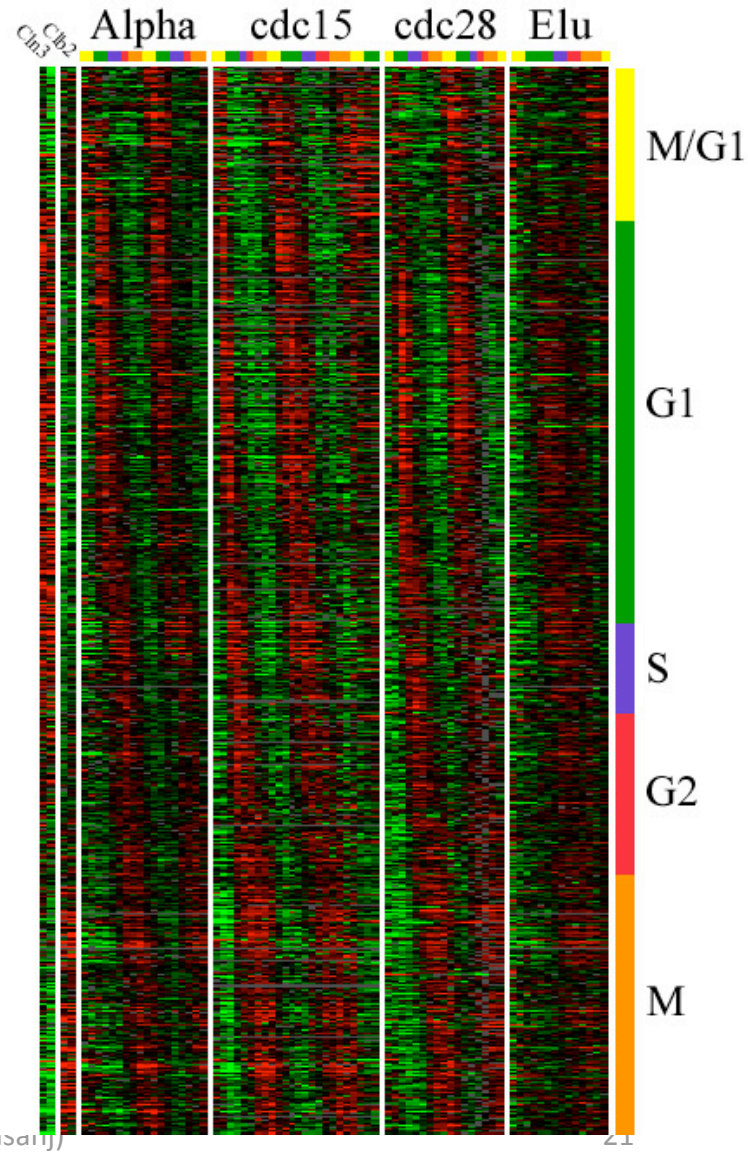
http://public.lanl.gov/mewall/kluwer2002.html

Complex Regulation Drives Yeast Cell Cycle

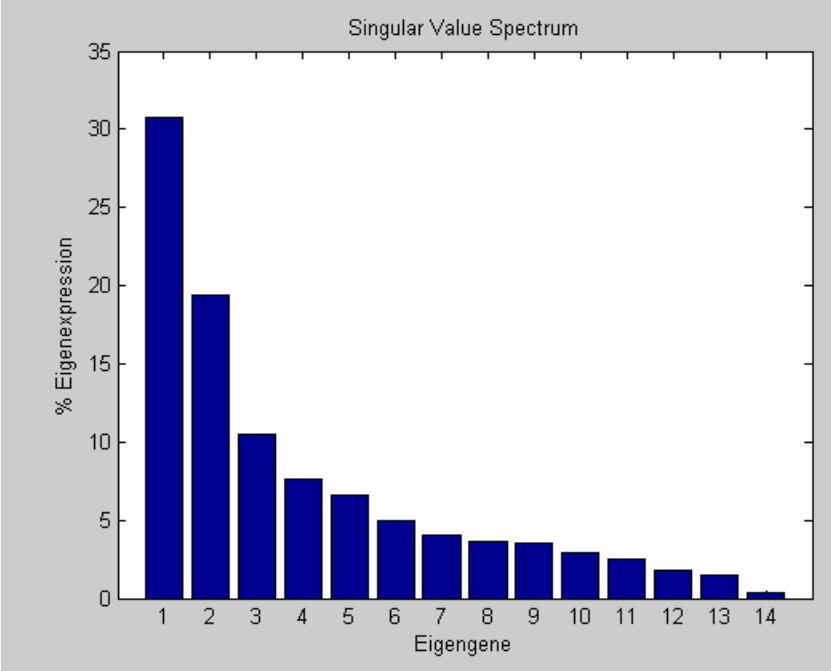
Yeast cyclins are proteins responsible for regulating cell cycle transitions. Cyclin gene (mRNA) expression data is taken from Spellman, et al., Molec. Biol. Cell, 9:3273, 1998.

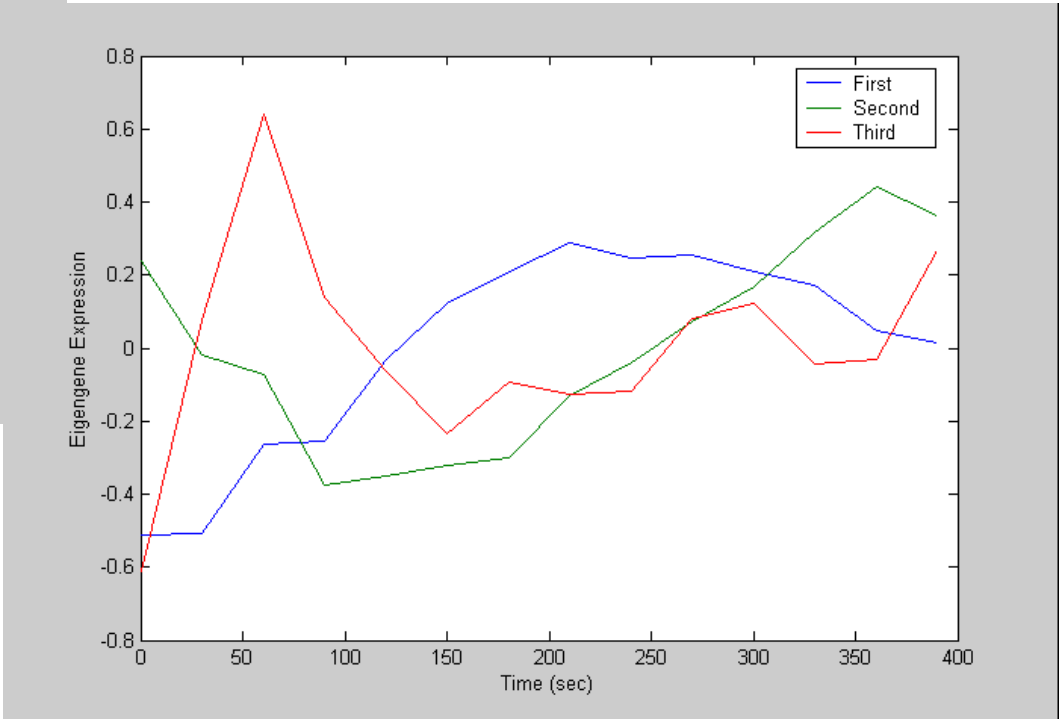


http://cyberia.cfdrc.com/datab/Applications/cell_tissue_bio/cellcycle/cellcycle.html



Singular Value Decomposition
(Yeast Cell Cycle Microarray Data)



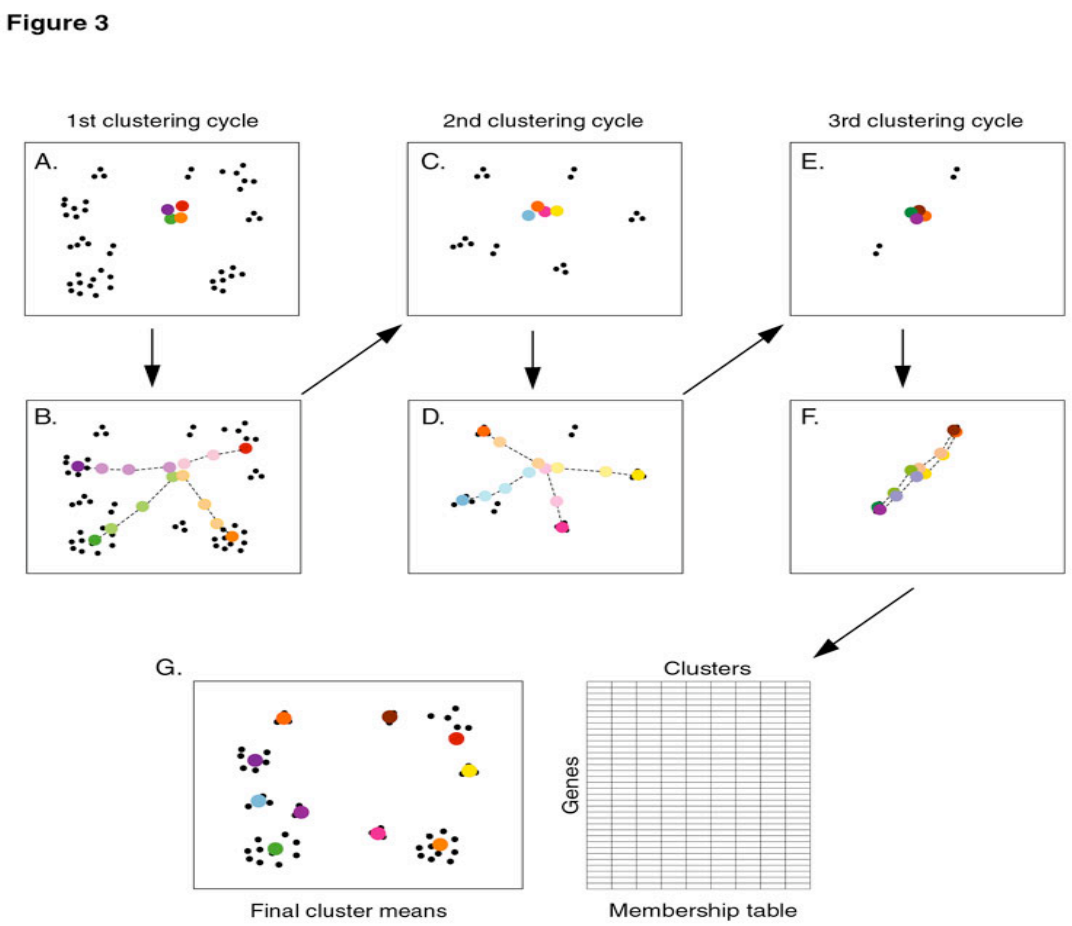


(based on 38 genes, *elu* expression dataset)

Many Ways to Identify Signatures

- Identifying major "components" of variation (potentially something that has to be removed from data, such as a fundamental difference between sampled groups)
  - singular value decomposition (SVD), principle component analysis (PCA), etc.
- Finding groups within data
  - clustering
  - self-organizing maps
  - support vector machines

Clustering (k-means)

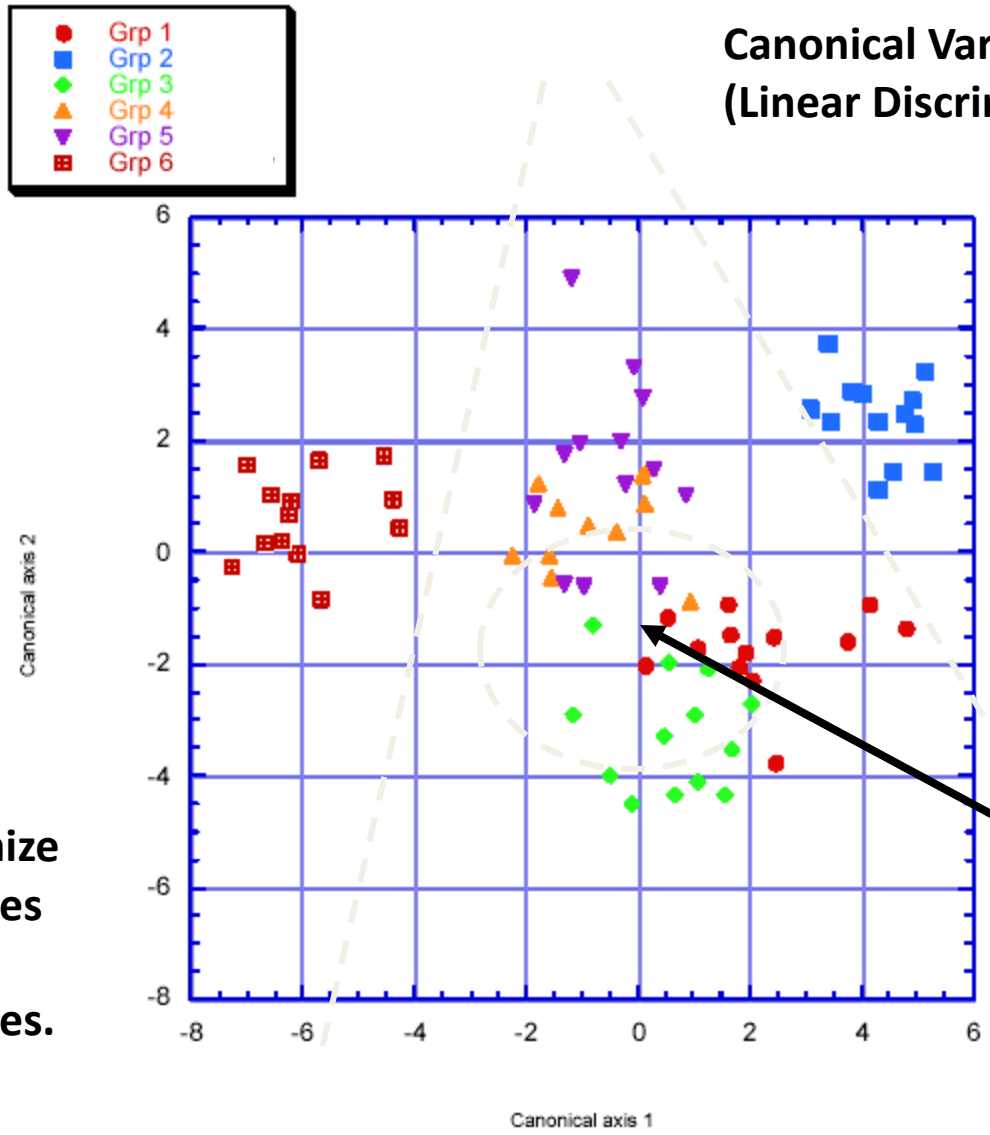


http://rana.lbl.gov/FuzzyK/images/figure3.html

Many Ways to Identify Signatures

- Identifying major "components" of variation (potentially something that has to be removed from data, such as a fundamental difference between sampled groups)
  - singular value decomposition (SVD), principle component analysis (PCA), etc.
- Finding groups within data
  - clustering
  - self-organizing maps
  - support vector machines
- Separating known groups
  - univariate methods (i.e. B-Tests, T-Tests on each gene, ANOVA on each gene)
  - horrible "capitalization on chance" problems
  - linear discriminant analysis / canonical variate analysis
    - these methods can be generalized for undetermined data, though the relative magnitudes of variables becomes significant in that case (but that filters out potentially noisy data) OR you get capitalization by chance by using stepwise methods

Linear Separation – Group Classification



Canonical Variate Analysis (Linear Discriminant Analysis)

Find optimal linear combinations of variables that maximize inter-group differences while minimizing intra-group differences.

Nonlinear Classification? (Kernel methods)

What Do We Get From Signatures

- Pattern for discrimination between groups (responders, non-responders, different genetic populations, etc.)
  - therapeutic design
  - diagnostics

- Lists of Genes
  - what genes appear to be the most significant in determining the difference between groups or cause the formation of distinct patterns within data?
    - you get long lists when you “capitalize on chance” using univariate methods
    - shorter lists from multivariate methods or when you use “honest” statistics modified for variable selection

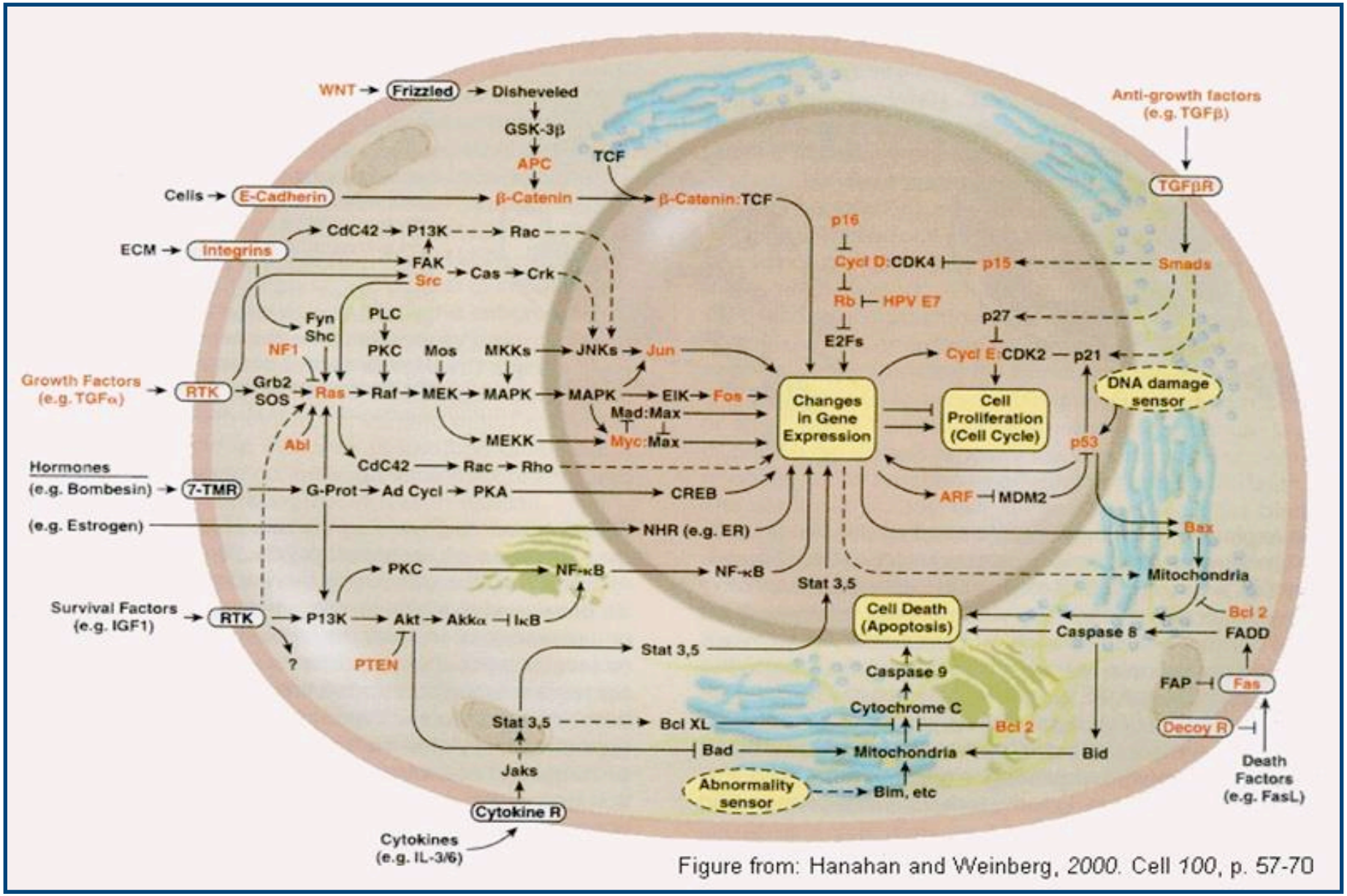
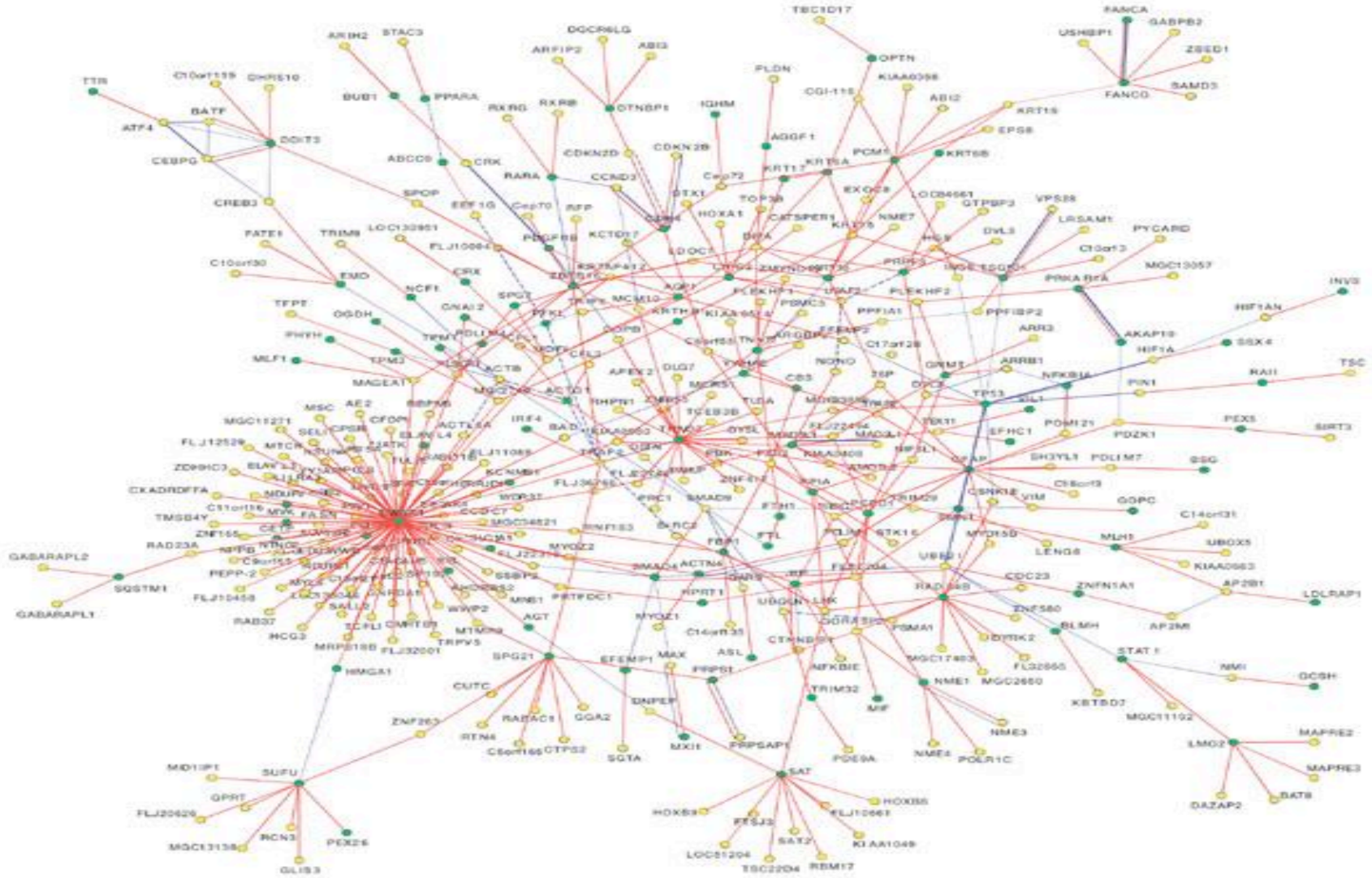
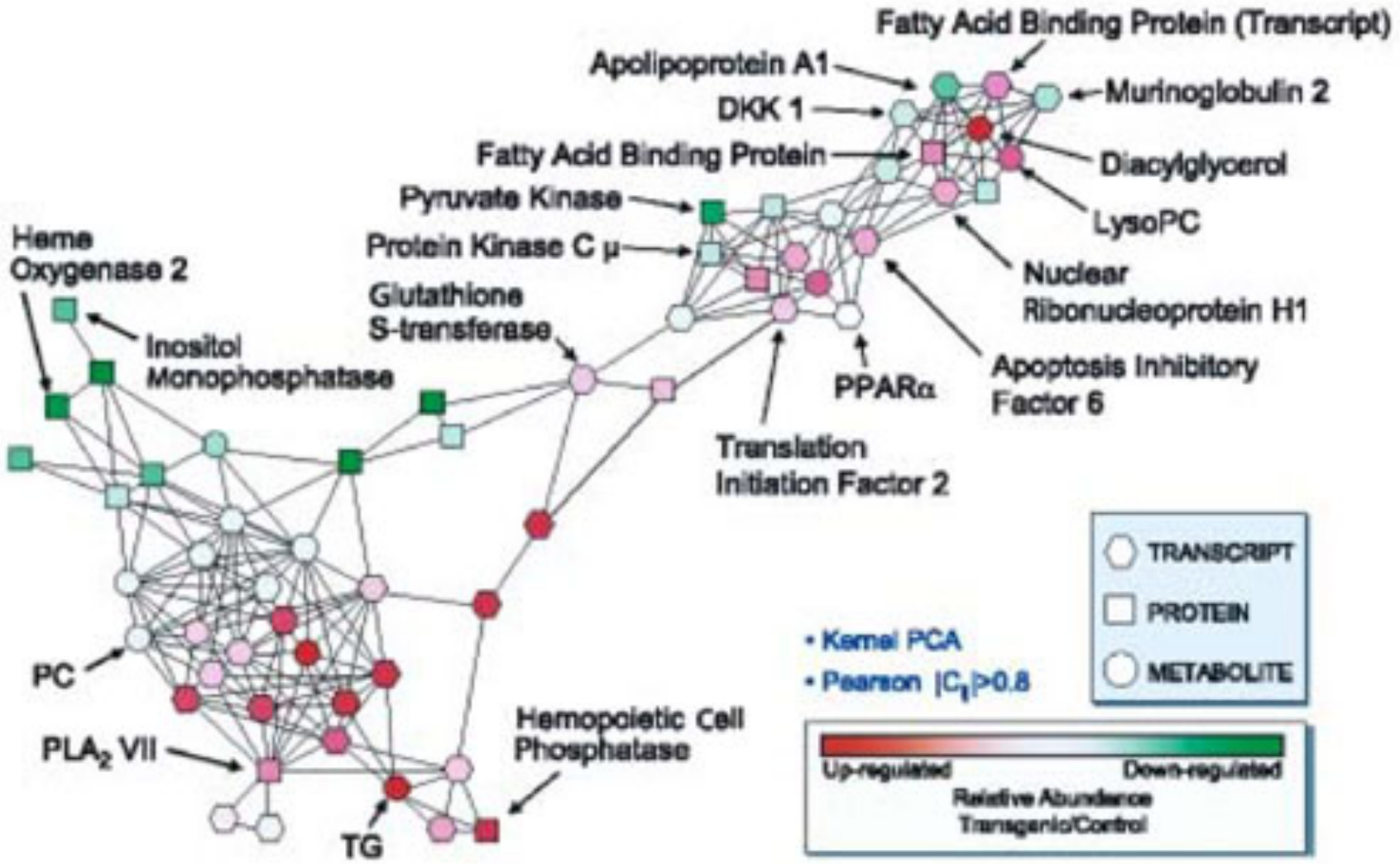


Figure from: Hanahan and Weinberg, *2000*. Cell *100*, p. 57-70



Rual et al., Nature 437: 1173, 2005.



Morel et al., Mayo Clin Proc. 2004;79:651-658

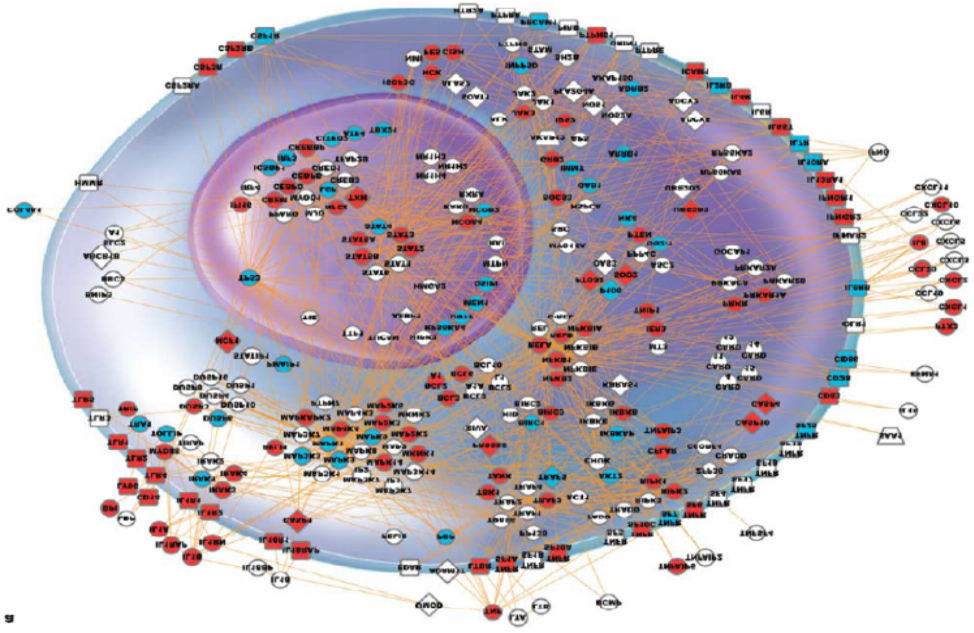
nature Vol 437|13 October 2005|doi:10.1038/nature03985

LETTERS

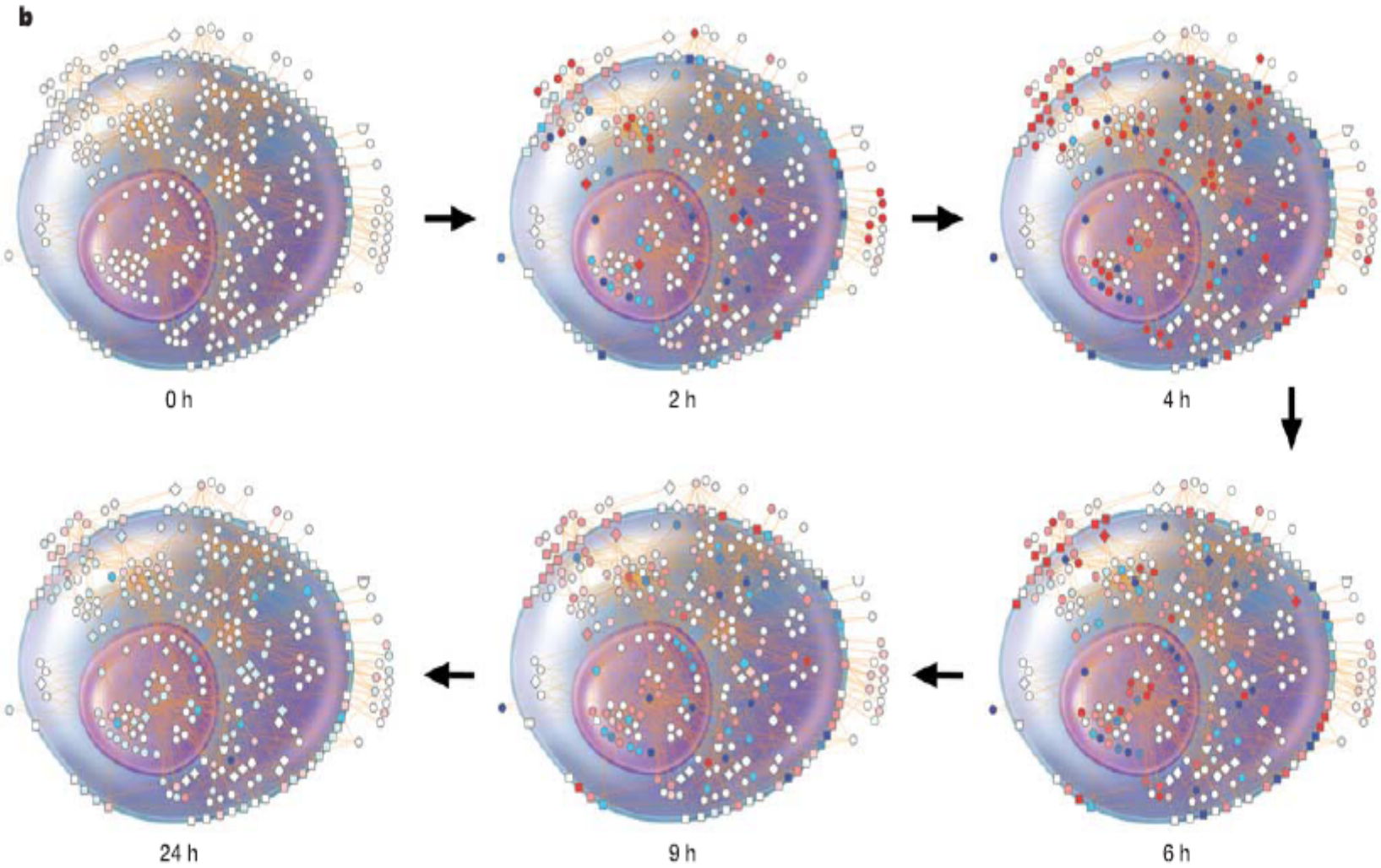
A network-based analysis of systemic inflammation in humans

Steve E. Calvano[1]*, Wenzhong Xiao[2]*, Daniel R. Richards[3], Ramon M. Felciano[3], Henry V. Baker[4,5], Raymond J. Cho[3], Richard O. Chen[3], Bernard H. Brownstein[6], J. Perren Cobb[6], S. Kevin Tschoeke[5], Carol Miller-Graziano[7], Lyle L. Moldawer[5], Michael N. Mindrinos[2], Ronald W. Davis[2], Ronald G. Tompkins[8], Stephen F. Lowry[1] & the Inflammation and Host Response to Injury Large Scale Collaborative Research Program†

Oligonucleotide and complementary DNA microarrays are being used to subclassify histologically similar tumours, monitor disease progress, and individualize treatment regimens[1–5]. However, extracting new biological insight from high-throughput genomic studies of human diseases is a challenge, limited by difficulties in recognizing and evaluating relevant biological processes from huge quantities of experimental data. Here we present a structured network knowledge-base approach to analyse genome-wide transcriptional responses in the context of known functional interrelationships among proteins, small molecules and phenotypes. This approach was used to analyse changes in blood leukocyte gene expression patterns in human subjects receiving an inflammatory stimulus (bacterial endotoxin). We explore the known genome-wide interaction network to identify significant functional modules perturbed in response to this stimulus. Our analysis reveals that the human blood leukocyte response to acute systemic inflammation includes the transient dysregulation of leukocyte bioenergetics and modulation of translational machinery. These findings provide insight into the regulation of global leukocyte activities as they relate to innate immune system



Calvano et al., Nature 2005, 473: 1032 -1037



Calvano et al., Nature 2005, 473: 1032 -1037

Problems:

1. Correlation of genomic and proteomic data does not provide functional information

Solution: Computational modeling as a tool

2. Complexity: Thousands of interactions that cannot be studied via single gene/protein approach

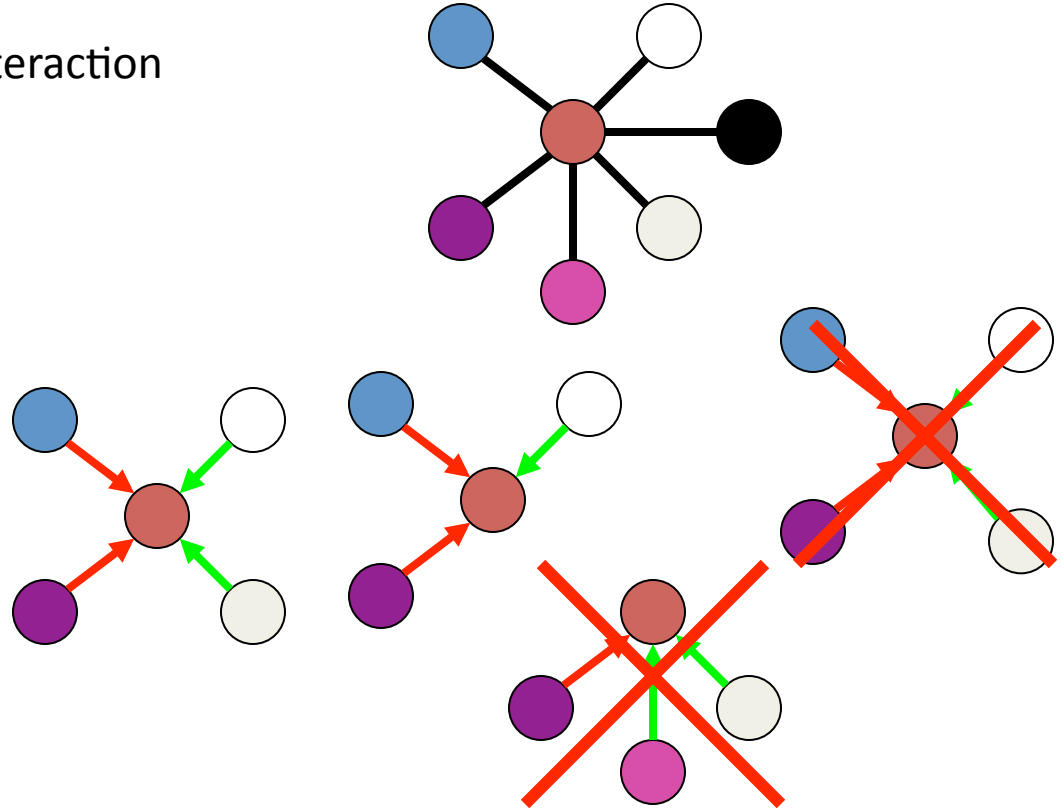
Solution: Systems Biology approach

Text-mining, protein-protein interaction databases, Gene Ontology...

identify

potential relationships

between variables



Dynamic Measurements, Functional Genomics, Functional Proteomics...

constrain

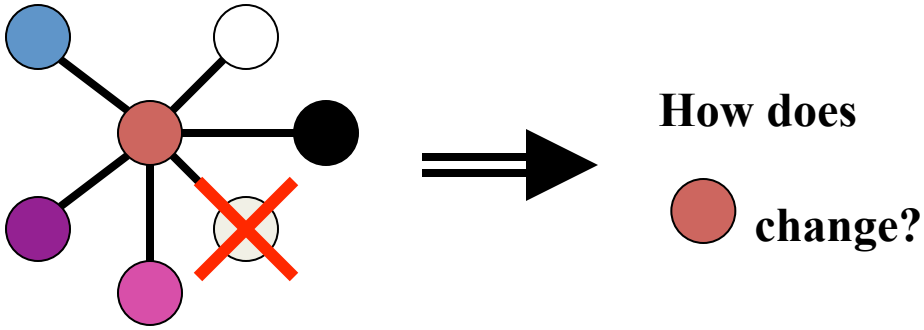
relationship models

Differences between models

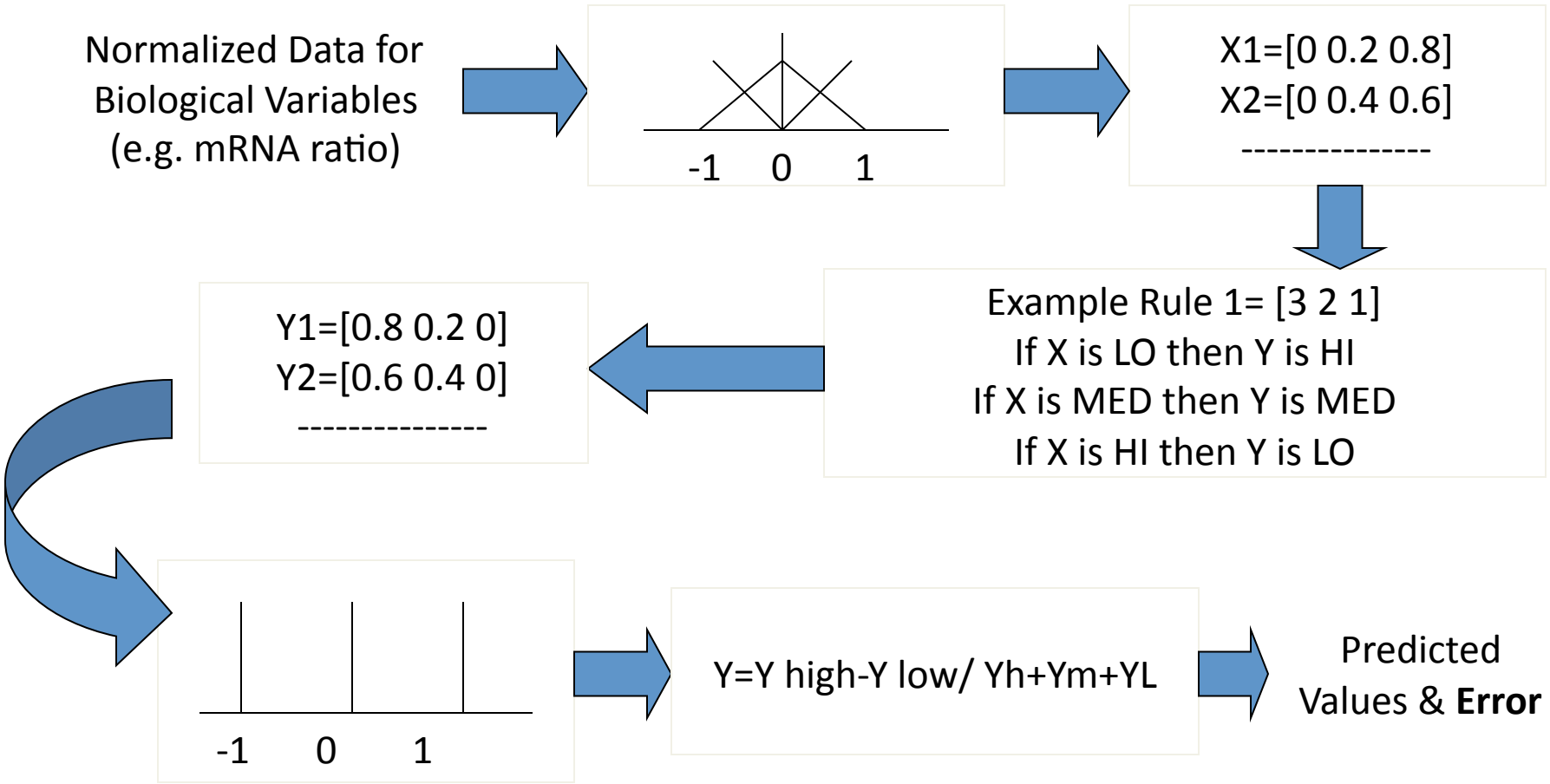
design

Hypothesis driven

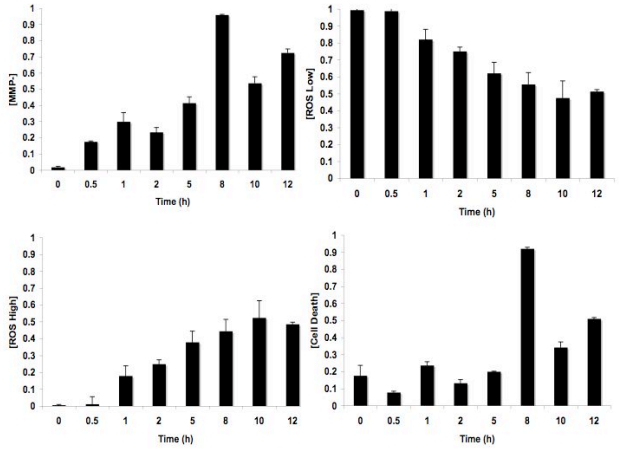
lab experiments & clinical trials

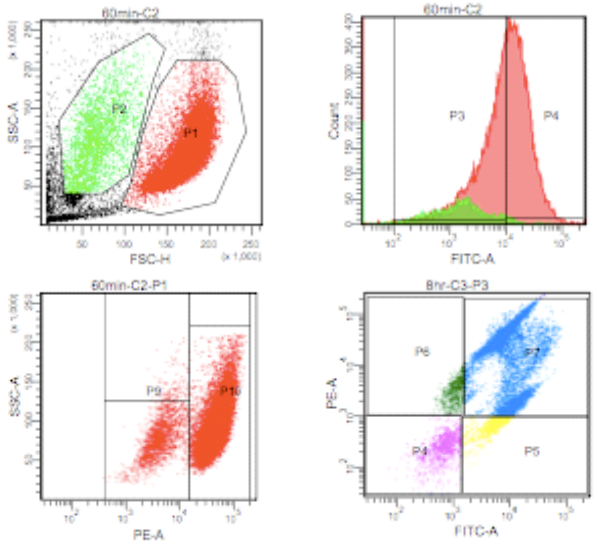


Fuzzy Simulation Procedure

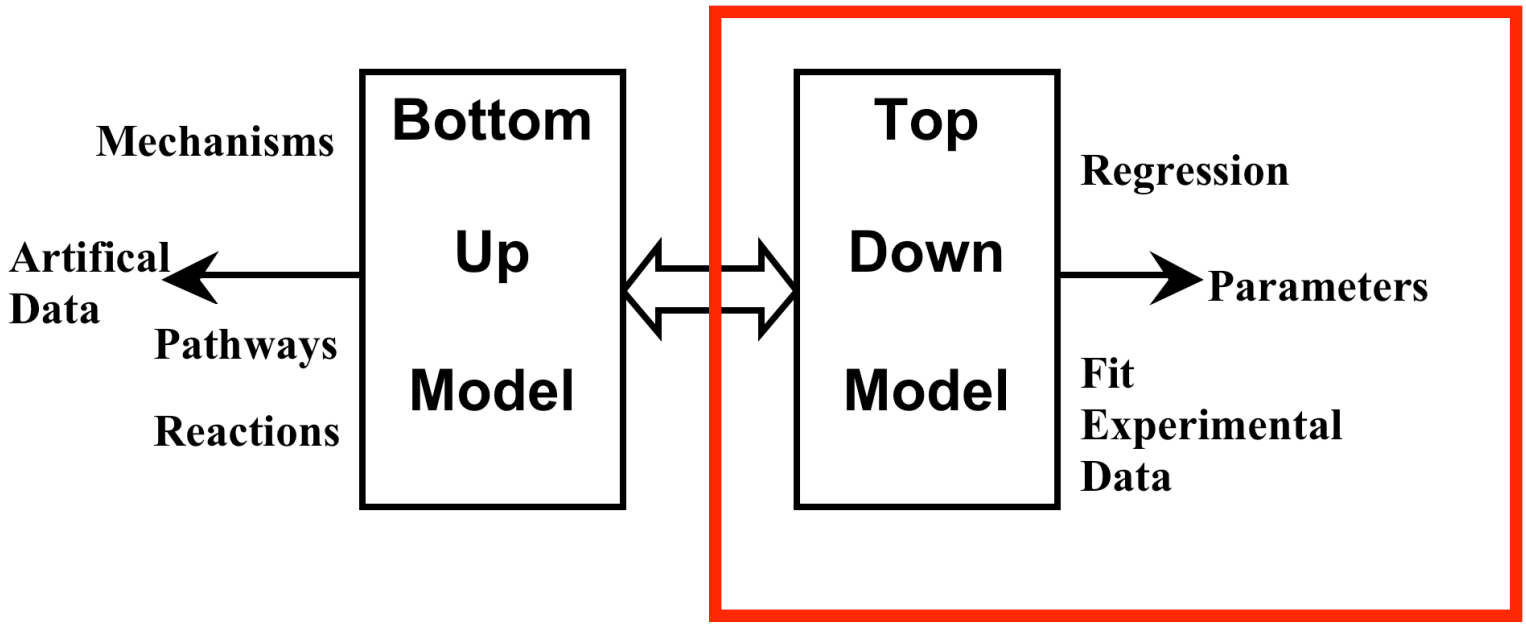


Quantitative Measurement

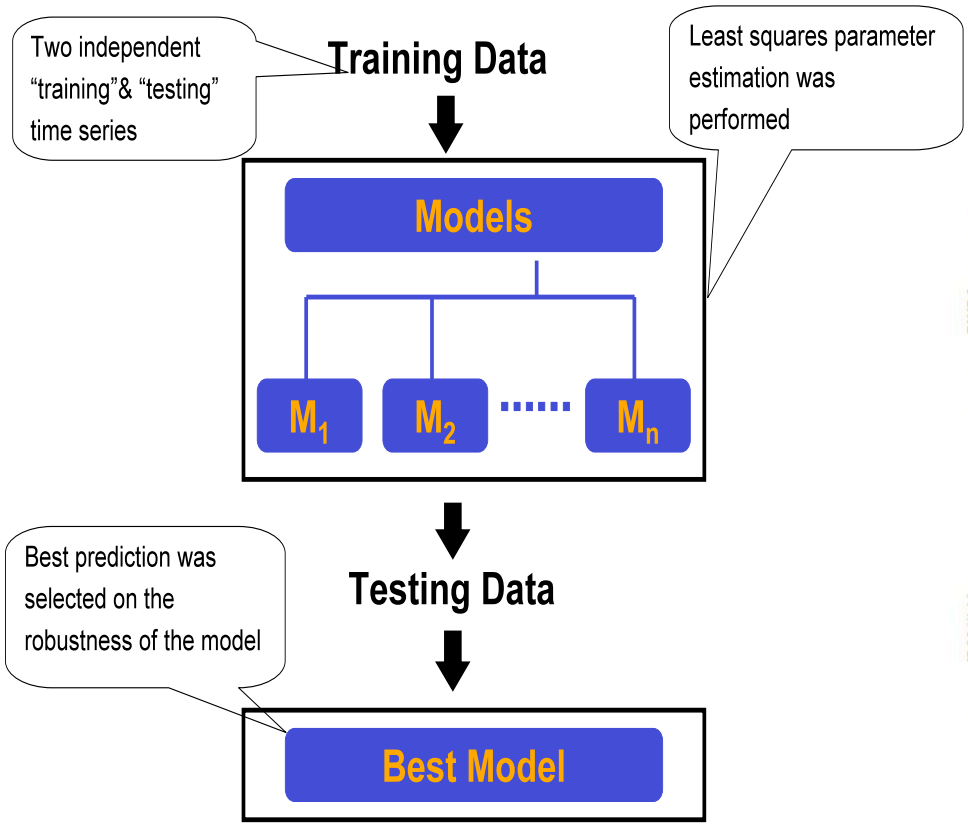


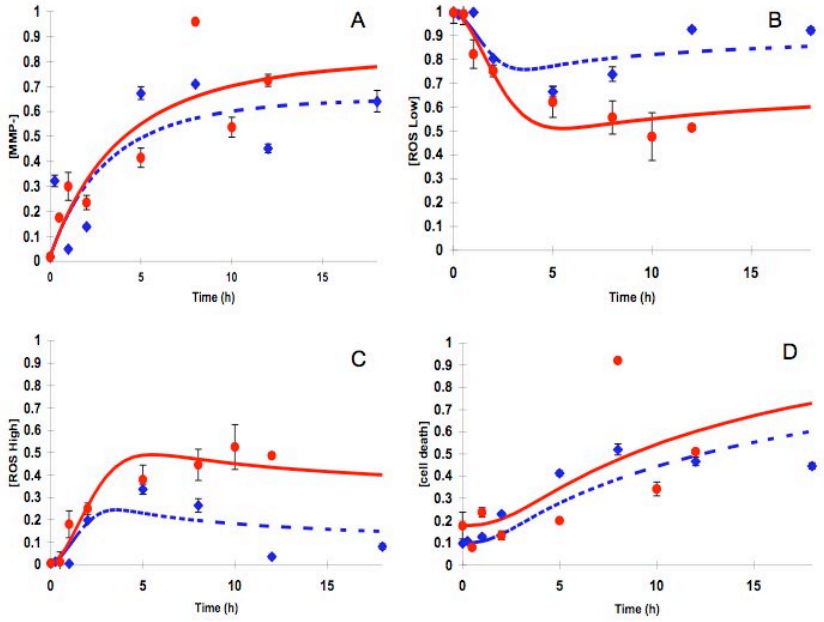


Modeling methodology



Developing Models based on Data

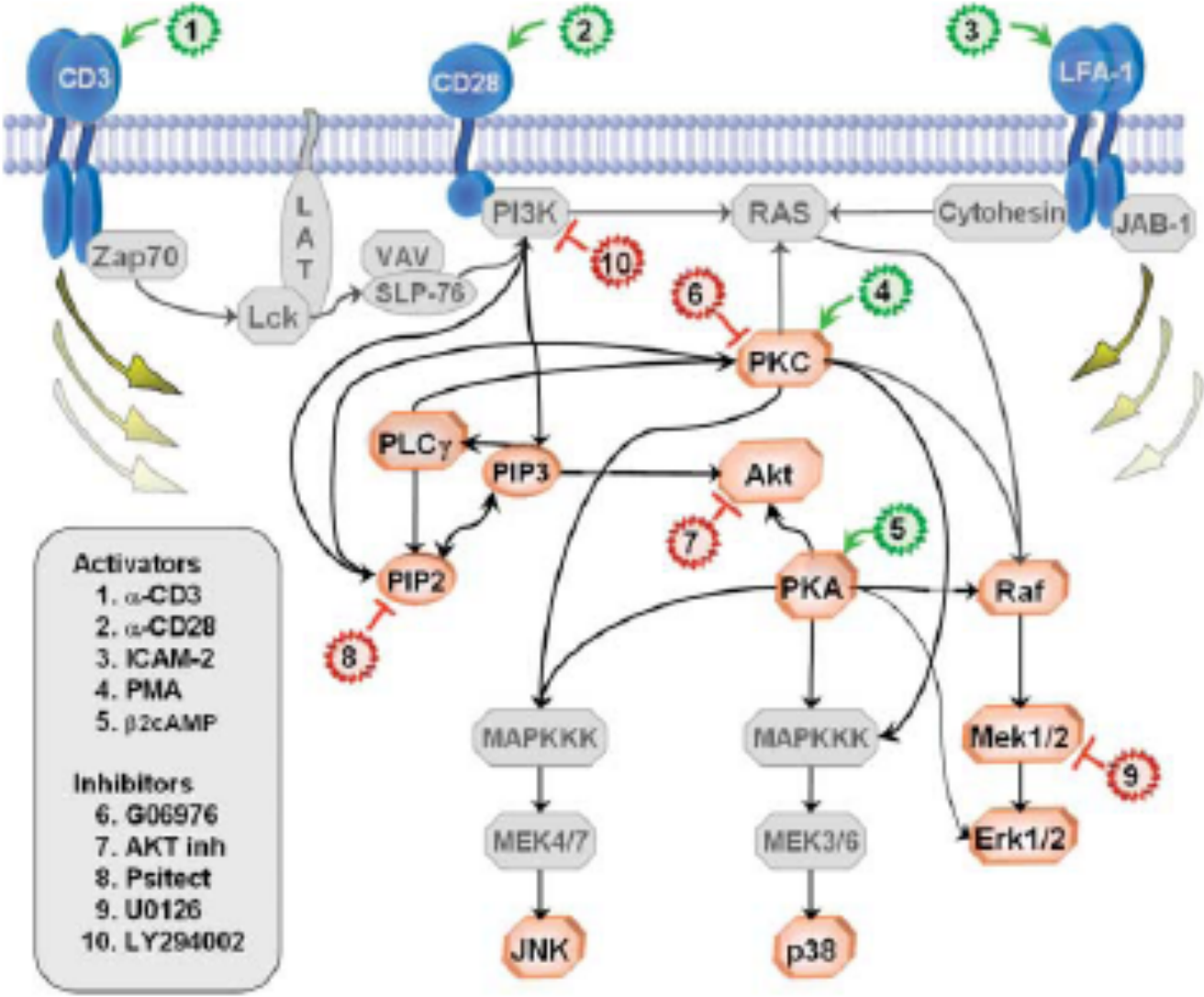




Bayesian Network Inference

- *Science*, Volume 308, pp 523-529, April 22, 2005
  "Causal protein-signaling networks derived from multiparameter single-cell data"
  K. Sachs, O. Perez, D. Pe'er, D. A. Lauffenburger, G. P. Nolan

Biomolecular Network

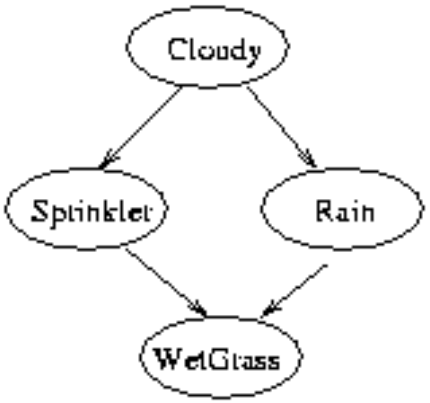


Application of Bayes Nets

- “Bayesian Network”, “Bayes Net”... a generalization of “Bayesian Inference”
- Directed, acyclic graphs representing relationships between variables
- NO FEEDBACK LOOPS
  - (these have to be handled using “Dynamic Bayesian Networks”)

Example of Bayesian Net Inference

P(C=F)	P(C=T)
0.5	0.5



C	P(S=F)	P(S=T)
F	0.5	0.5
T	0.9	0.1

C	P(R=F)	P(R=T)
F	0.8	0.2
T	0.2	0.8

S	R	P(W=F)	P(W=T)
F	F	1.0	0.0
T	F	0.1	0.9
F	T	0.1	0.9
T	T	0.01	0.99

http://www.cs.ubc.ca/~murphyk/Bayes/bnintro.html

Simplify Joint Probability Distributions

- By the chain rule of probability, the joint probability of all the nodes in the graph above is
- P(C, S, R, W) = P(C) * P(S|C) * P(R|C,S) * P(W|C,S,R)
- By using conditional independence relationships, we can rewrite this as
- P(C, S, R, W) = P(C) * P(S|C) * P(R|C) * P(W|S,R)

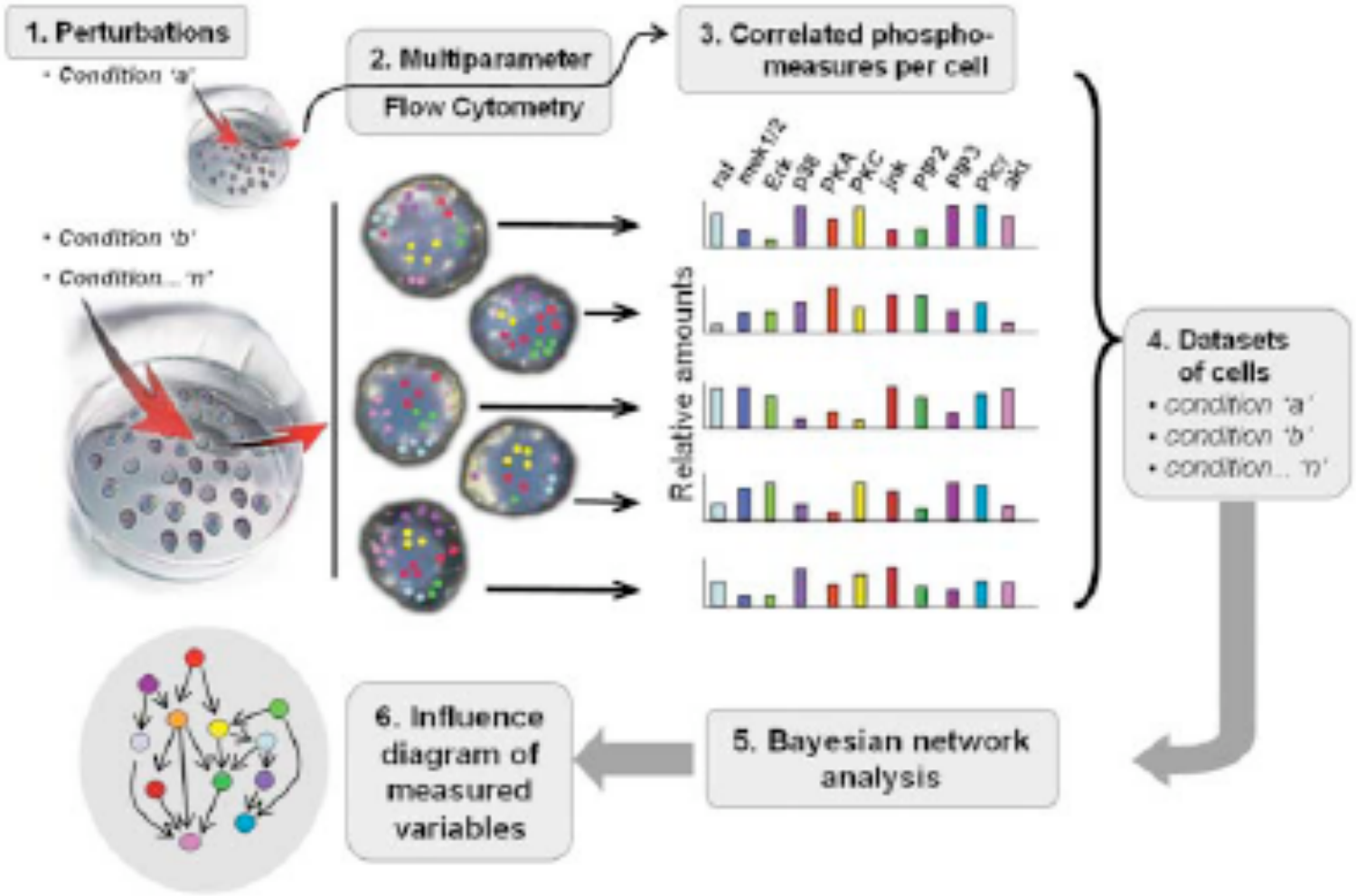
- where we were allowed to simplify the third term because R is independent of S given its parent C, and the last term because W is independent of C given its parents S and R.

- In general, if we had n binary nodes, the full joint would require O(2^n) space to represent, but the factored form would require O(n 2^k) space to represent, where k is the maximum fan-in of a node. And fewer parameters makes learning easier.

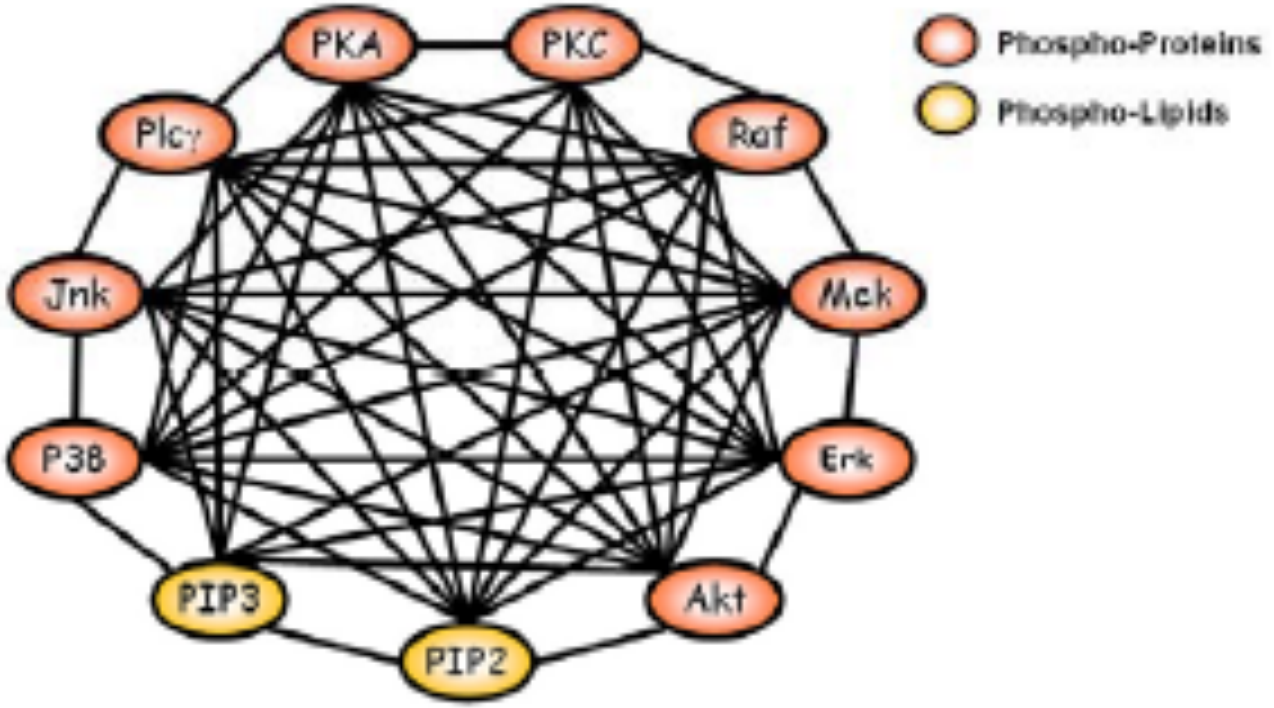
Identifying structure of Bayesian networks

- You can use MCMC (Markov Chain Monte Carlo) to “learn” parameter values based on data, or you can generate models and see how well they can predict correlations that you measure under different perturbations (the approach taken in the paper)

Network Inference from Experiments

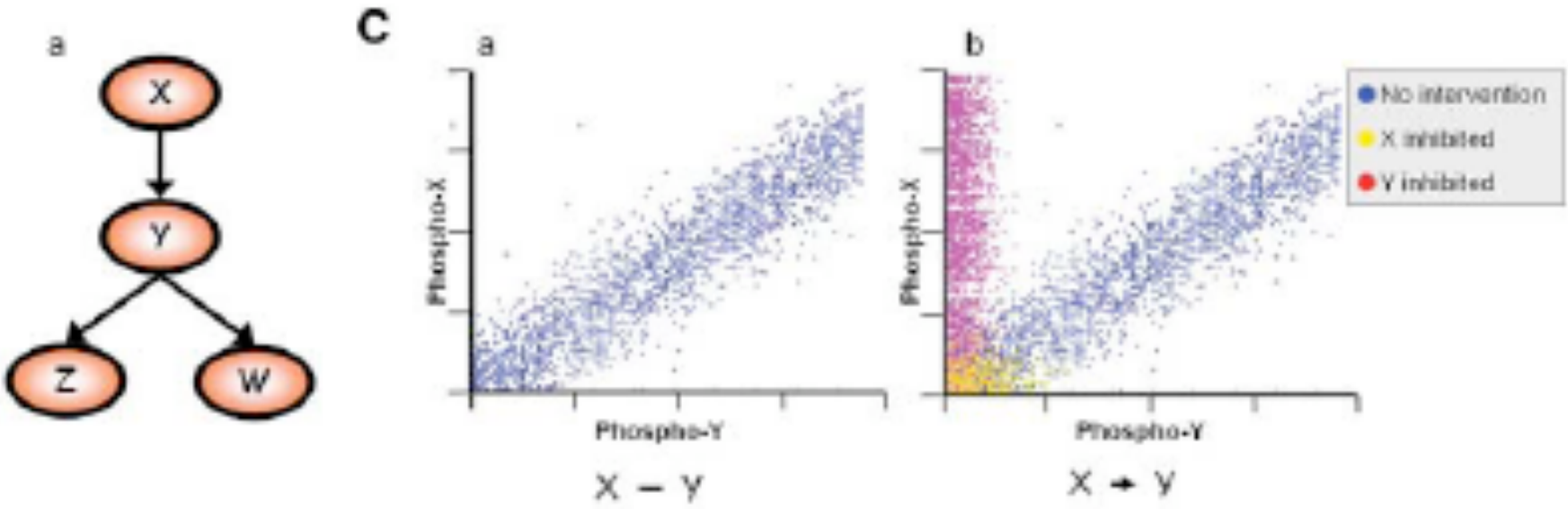


Looking at correlations = Big mess



52 out of 55 possible(!) correlations are found in data with reasonable p value

Identifying causality



X and Y are correlated... now, if we inhibit X, we see an effect on Y (yellow dots) But if we inhibit Y, there is no effect on X (purple dots). So, X influences Y, but not vice versa.

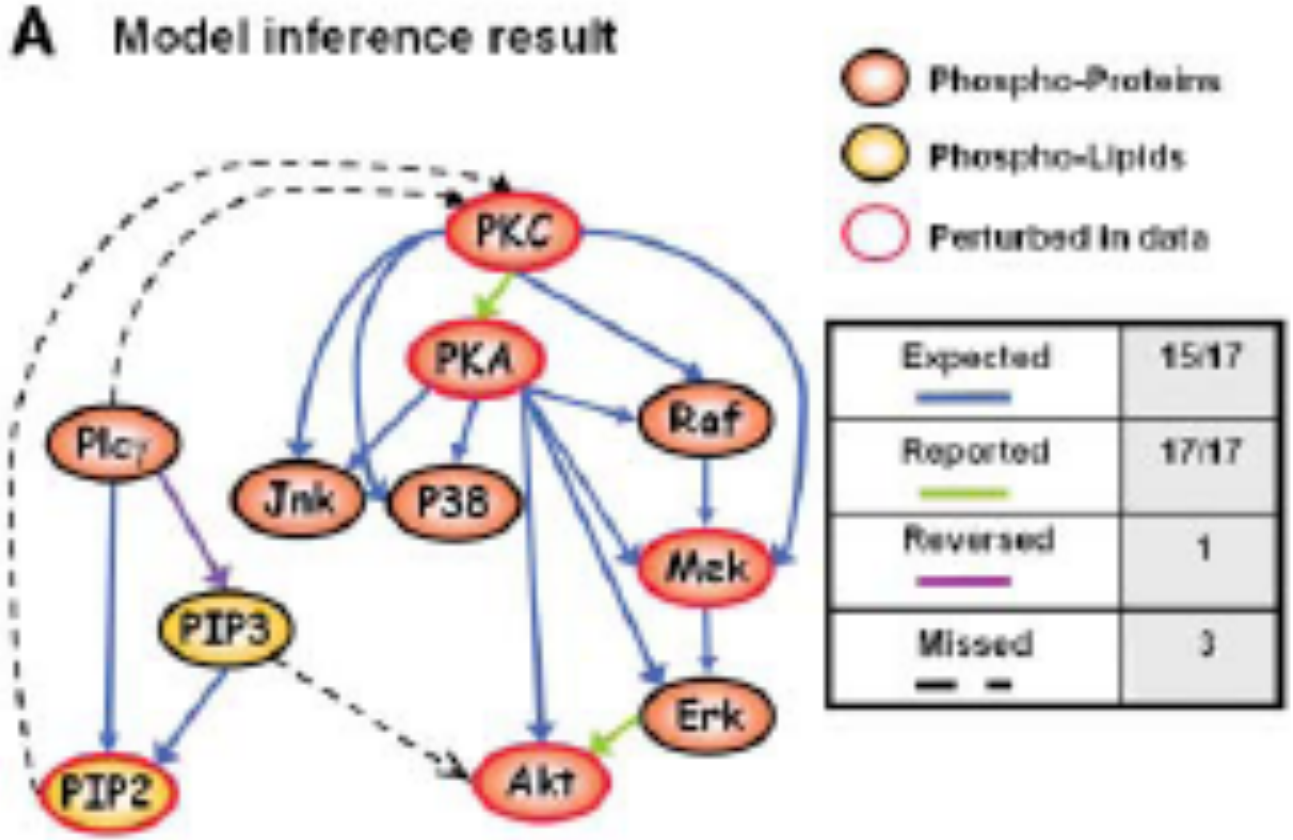
Identifying a complex network

- Simulated annealing was used, with operators adding or deleting arcs in the graph. In most cases, a change that led to a higher scoring (better fit) was accepted, but there was a small (and diminishing) probability of accepting a worse scoring solution (to avoid falling in a local minimum).

- The search starts with a random graph; it was repeated 500 times to explore the whole search space – all the highest scoring networks were combined to form a "consensus graph"

Consensus results

Final result for confidence > 85% graphs



Limitation of acyclicity

Some loops were excluded by confidence > 85% but could be biological significant (big limitation of Bayes nets – but what about Dynamic Bayesian Networks?)