From Data to Models: Systems biology methods and potential applications to toxicoinformatics

I.P. Androulakis

Biomedical Engineering Department

Chemical & Biochemical Engineering Department

RUTGERS

School of Engineering

ebCTC environmental bioinformatics and Computational Toxicology Center

From Data to Models: What is it?

Disclaimer: this a rather philosophical discussion so I would not necessarily spend too much time thinking about this page. However, it may be useful to set the stage

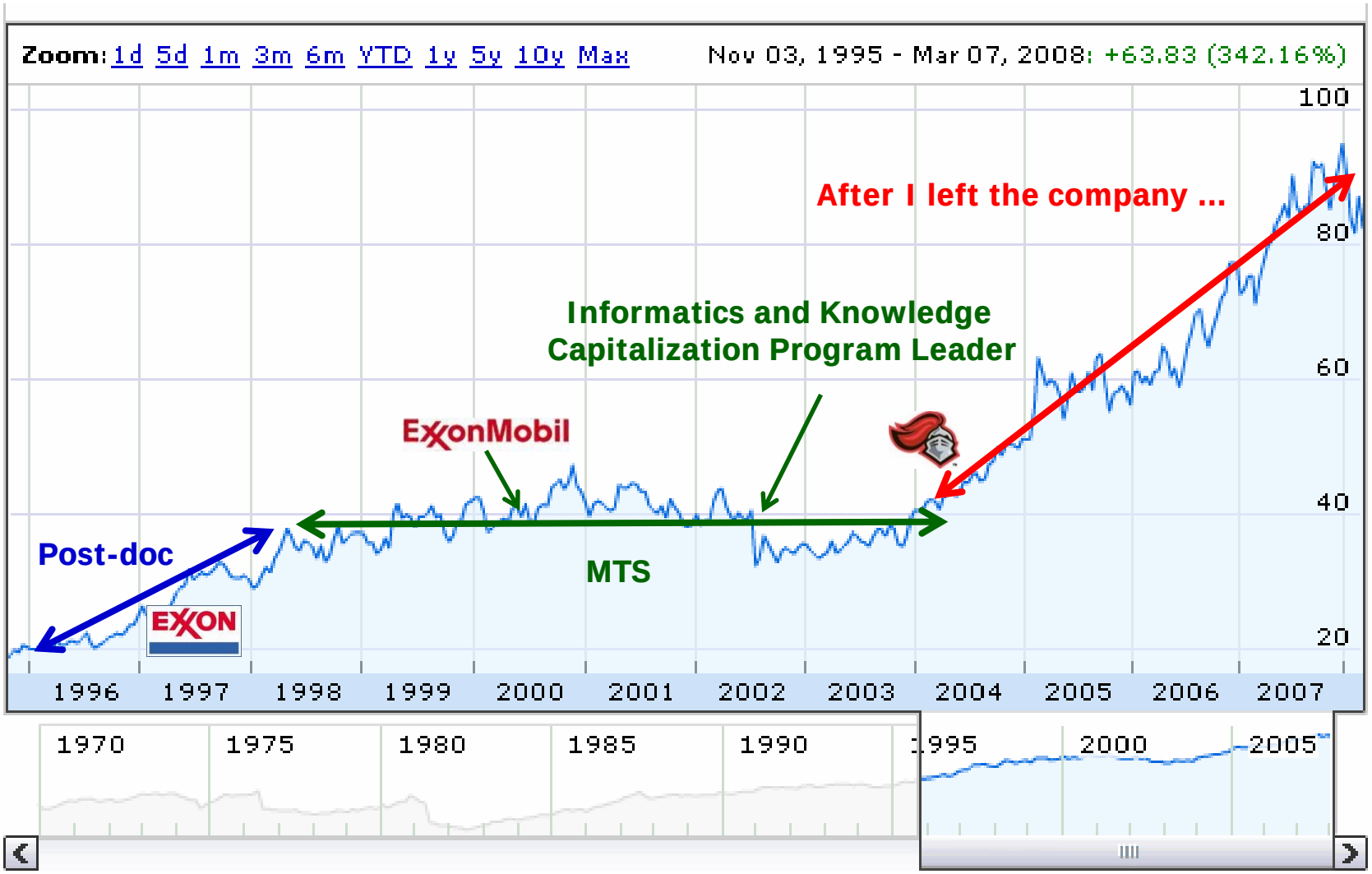
From Data to Petterns

- It is an undeniable fact that data is everywhere and we have to do something with it … not sure what sometimes
  - *Beer and nappies – A data mining urban legend*
- The idea of collecting, annotating, warehousing, and analyzing data for the purpose of unraveling possible patterns has been extensively discussed and will **not** be part of this talk

From Data to Models

- A pattern is simply a coincidence or a potentially useful observation, if repeated at a very high rate, unless it can be interpreted using available laws or can be used to develop new laws that explain old behaviors and predict new
  - *Warning: This is an expression of my personal bias*
- The model is a quantification, not necessarily in closed form, of a law
- Actions and testable hypotheses in science and engineering are better designed with models rather than "knowledge"

What's in a Pattern?



www.finance.google.com

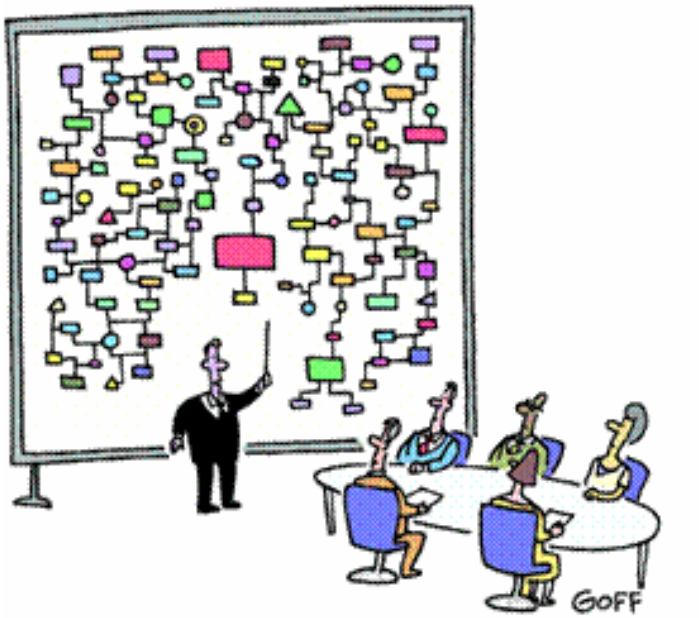
From Data to Models: Why Now?

Complexity and emergence are old concepts, so why this suspicious interest?

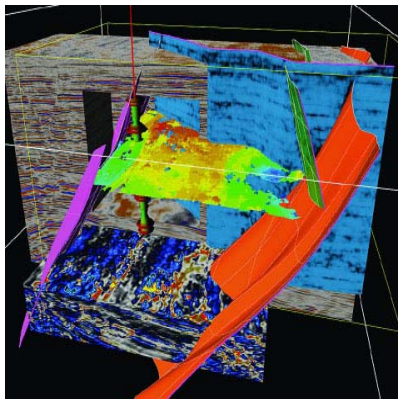
Technological advances allowed the handling of overwhelming amounts of data

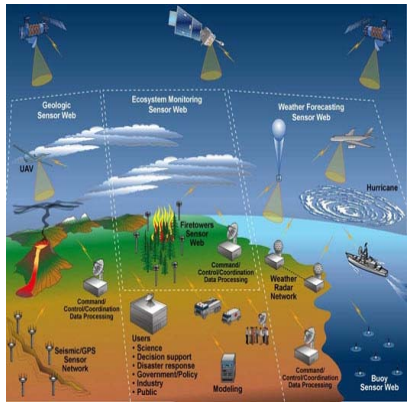
- Subsurface Imaging; GIS; Fraud Detection; HDHA; Oilfield Sensors

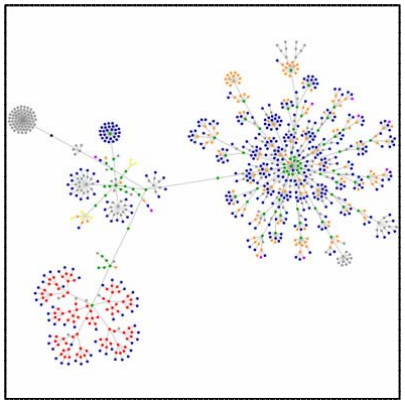
Complex systems require better management

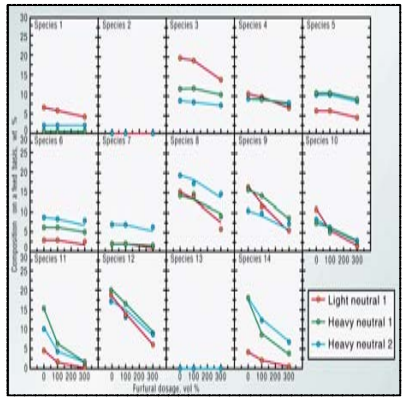


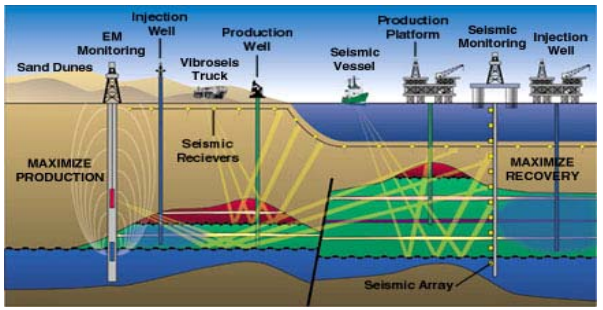
"And that's why we need a computer."









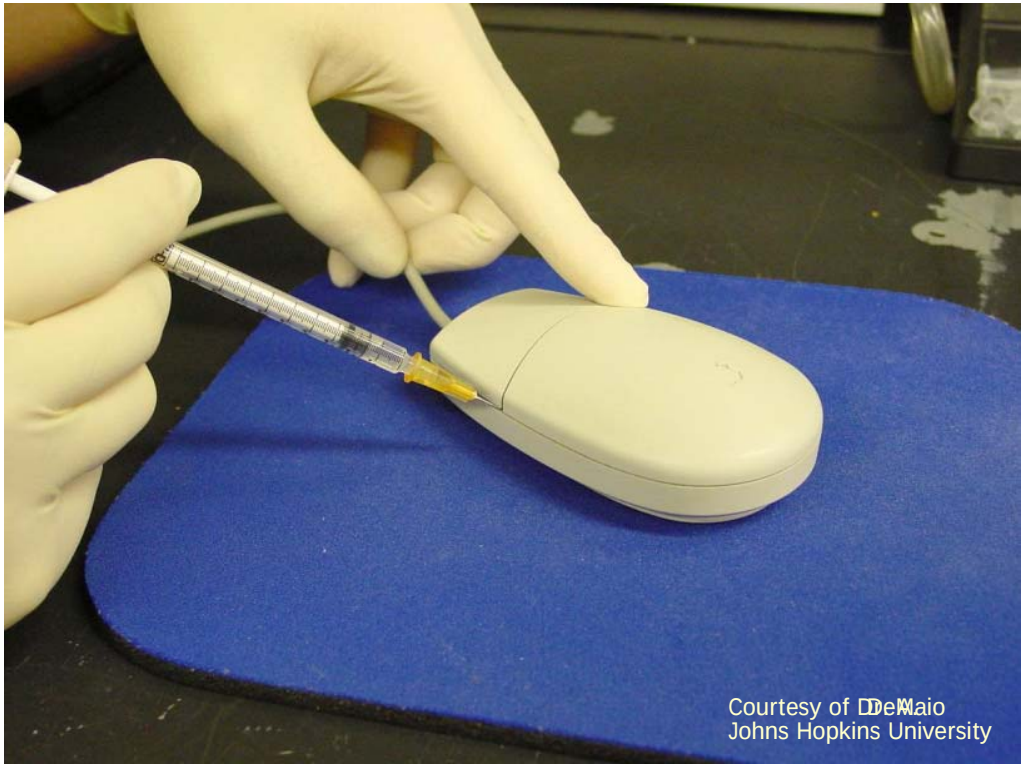


in silico Biology ?

Two major innovations opened up major opportunities

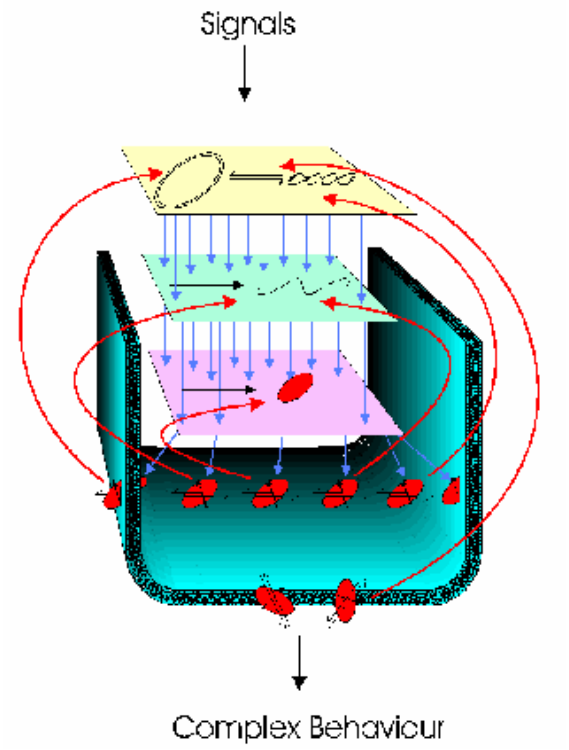
- Decoding of the (human) genome ⇔ State space definition
- High-throughout experimentation ⇔ Measurement of coordinated changes

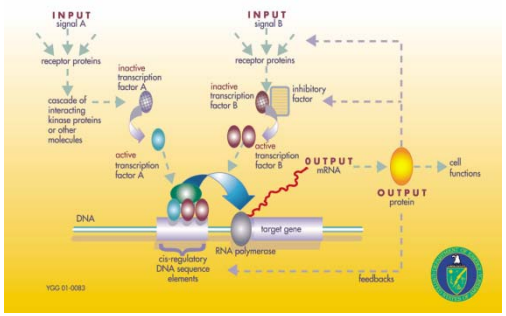
The system can be “systematically” probed and reverse-engineered to develop hypotheses for the next perturbation

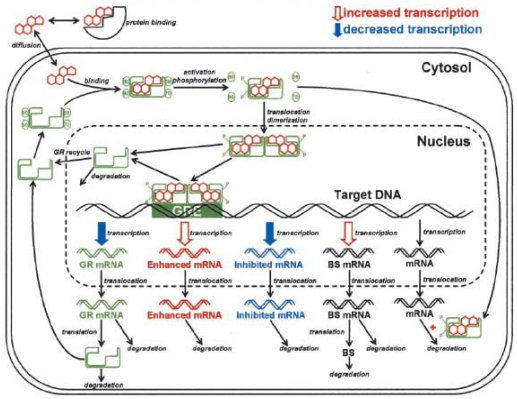


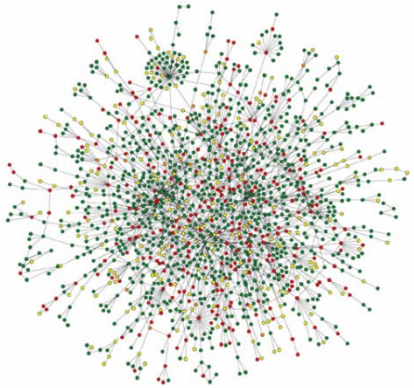
A Prototypical Example: *Systems Biology*

Biological systems propagate external perturbations across a complex **network** of interacting elements









→ Basic module

→ Systems of modules

→ Networks of systems

From Data to Models: Some Important Problems

Which of the features capture the structure in the data?

Which of the samples increase the information content of the data?

Which of the modules are important?

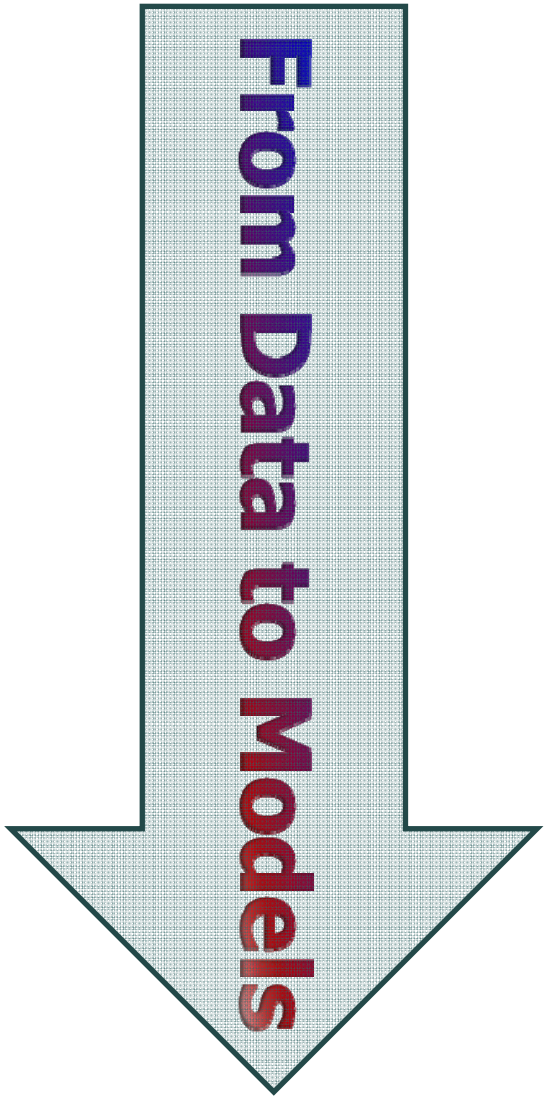
Which of the interactions among the modules are important?

How are biological systems organized in the form of complex networks?

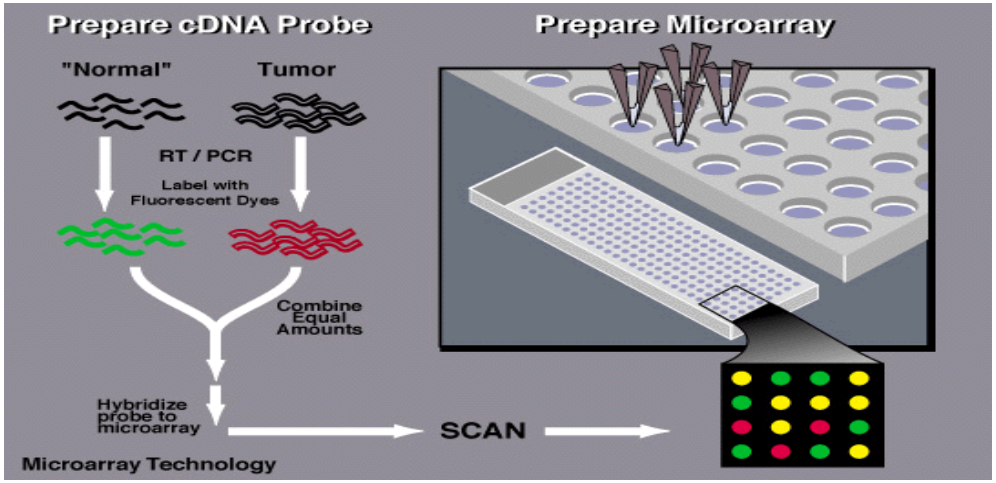
How can we develop models that explain the propagation of disturbances through the interaction of modules giving rise to observed emerging behaviors?

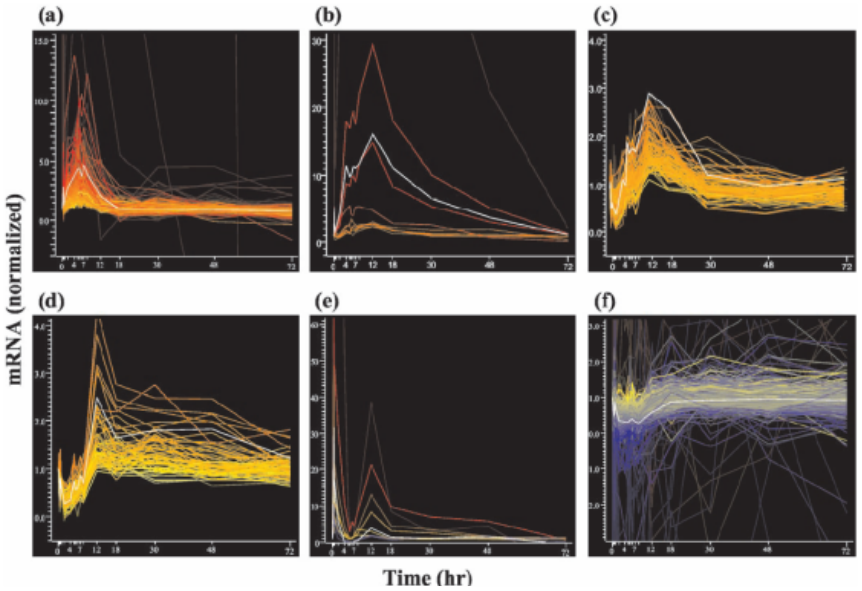
In this talk

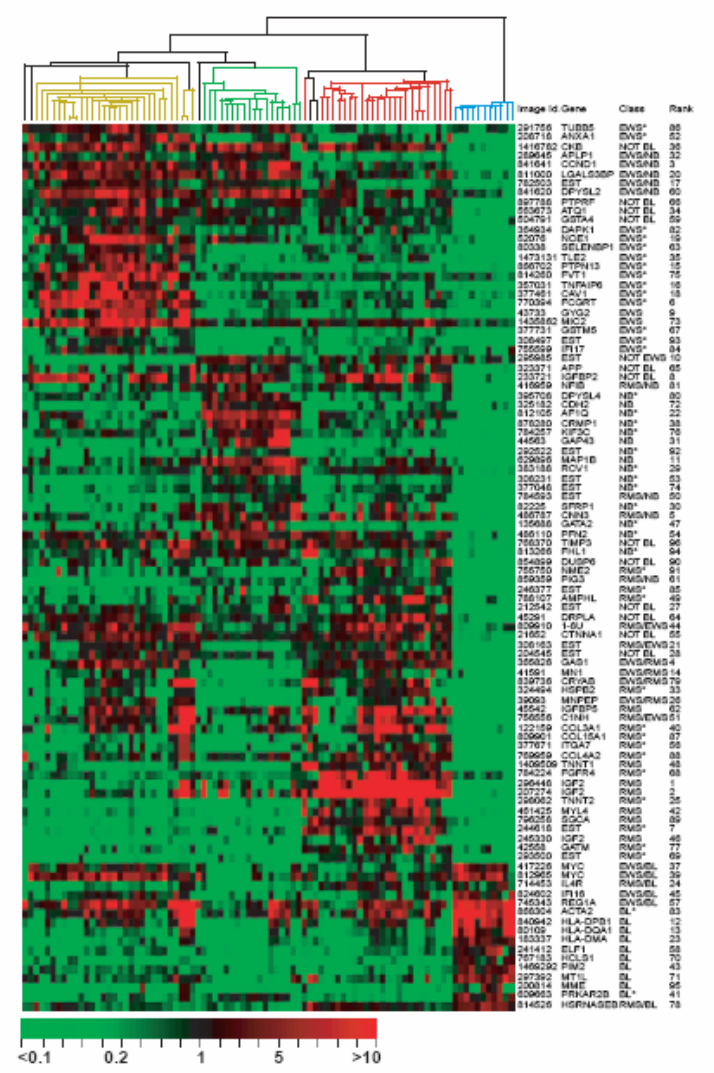
- How to use computational thinking
- This is not a comprehensive review
- This is not the end of the story



High-throughput Measurement of Gene Expression





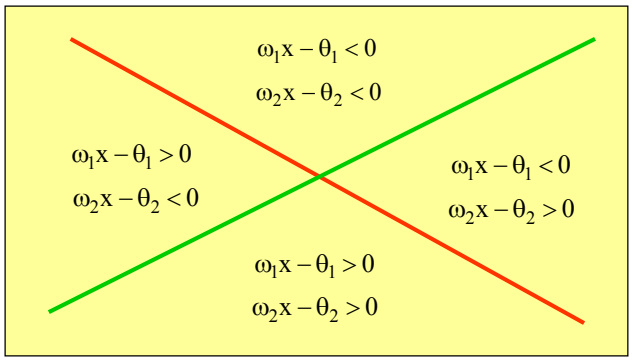


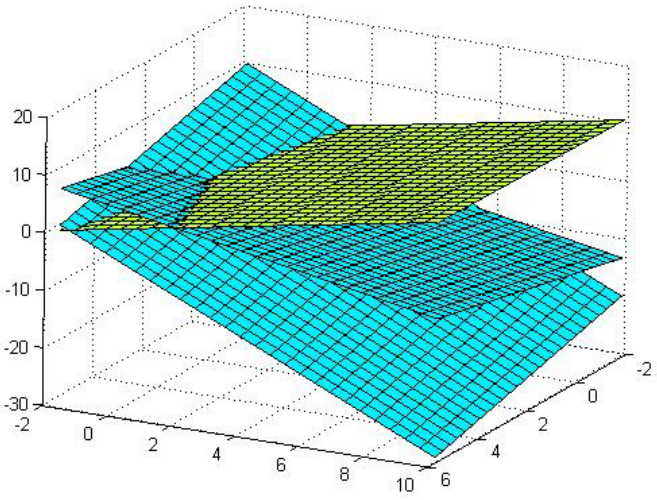
Feature Selection and Model Complexity

Oblique Multicategory Decision Tress

High dimensional spaces probably include redundant, i.e. uninformative, features

Formalize concept of model complexity





$$\min E$$

$$s.t.$$

$$\sum_{k,k'>k,\pi} h_{k,k',\pi} \le E$$

$$\sum_{k,k'>k,\pi} \hat{h}_{k,k',\pi} \le E$$

$$\sum_{k,\pi} y_{k,\pi} = N_\pi$$

$$q_{n,\pi} \le \frac{\sum_p c_1(p,\pi) z_{n,p} + c_0(\pi)}{N_p}$$

$$q_{n,\pi} \ge \sum_p c_1(p,\pi) z_{n,p} + c_0(\pi) - (N_p - 1)$$

$$\sum_n \sum_\pi q_{n,\pi} = N$$

$$v_{k,\pi} = \frac{\sigma_{k,\pi}}{N_k}$$

$$\sigma_{k,\pi} = \sum_n B(n,k) q_{n,\pi}$$

$$\sum_\pi v_{k,\pi} = 1$$

$$\sum_f D(n,f) w_{f,p} + U z_{n,p} \le U + \vartheta_p - z_{n,p}\varepsilon$$

$$\sum_f D(n,f) w_{f,p} + (u-\varepsilon) z_{n,p} \ge \varepsilon + \vartheta_p$$

$$y_{k,\pi} \le N_k v_{k,\pi}$$

$$y_{k,\pi} \ge v_{k,\pi}$$

$$w_{f,p} \ge \mu s_f$$

$$w_{f,p} \le M s_f$$

$$\sum_f s_f = N_f$$

$$h_{k,k',\pi} \le v_{k',\pi}$$

$$h_{k,k',\pi} \ge v_{k',\pi} - (1 - y_{k,\pi})$$

$$h_{k,k',\pi} \le y_{k,\pi}$$

$$\hat{h}_{k,k',\pi} \le v_{k,\pi}$$

$$\hat{h}_{k,k',\pi} \ge v_{k,\pi} - (1 - y_{k',\pi})$$

$$\hat{h}_{k,k',\pi} \le y_{k',\pi}$$

$$z_{n,p} = \begin{cases} 1 \text{ if } \sum_f D(n,f) w_{f,p} \le \vartheta_p \\ 0 \text{ if } \sum_f D(n,f) w_{f,p} \ge \vartheta_p \end{cases}$$

$$q_{n,\pi} = \begin{cases} 1 \text{ if point } n \text{ belongs to partition } \pi \\ 0 \text{ otherwise} \end{cases}$$

$\sigma_{k,\pi}$ = number of points of class k in partition π

$v_{k,\pi}$ = fraction of points of class k in partition π

$$y_{k,\pi} = \begin{cases} 1 \text{ if partition } \pi \text{ containts points of class } k \\ 0 \text{ otherwise} \end{cases}$$

$$s_f = \begin{cases} 1 \text{ if feature } f \text{ is used} \\ 0 \text{ otherwise} \end{cases}$$

$h_{k,k',\pi}$ = variable used to linearize the bilinear product $y_{k,\pi} v_{k',\pi}$

$\hat{h}_{k,k',\pi}$ = variable used to linearize the bilinear product $y_{k',\pi} v_{k,\pi}$

$$h_{k,k',\pi} \le E \Rightarrow \begin{cases} y_{k,\pi} = 0 \wedge v_{k',\pi} \le 1 \\ y_{k,\pi} = 1 \wedge v_{k',\pi} \le E \\ y_{k,\pi} = 0 \wedge v_{k',\pi} \le E \end{cases}$$

$$h_{k',k,\pi} \le E \Rightarrow \begin{cases} y_{k',\pi} = 0 \wedge v_{k,\pi} \le 1 \\ y_{k',\pi} = 1 \wedge v_{k,\pi} \le E \\ y_{k',\pi} = 0 \wedge v_{k,\pi} \le E \end{cases}$$

$w_{f,p}$ and ϑ_p = plane parameters, $Dw \le \vartheta$ (*variables*)

N = number of samples

N_π = desired number of occupied partitions (*parameter*)

N_p = number of planes (*parameter*)

N_f = number of features to be selected (*parameter*)

N_k = number of samples in class k (*known*)

u, U, μ, M = big-M parameters

c_0, c_1 = model parameters (*known*)

$D(n,f), B(n,k)$ = data

p = *planes*

π = *partitions* = 2^p

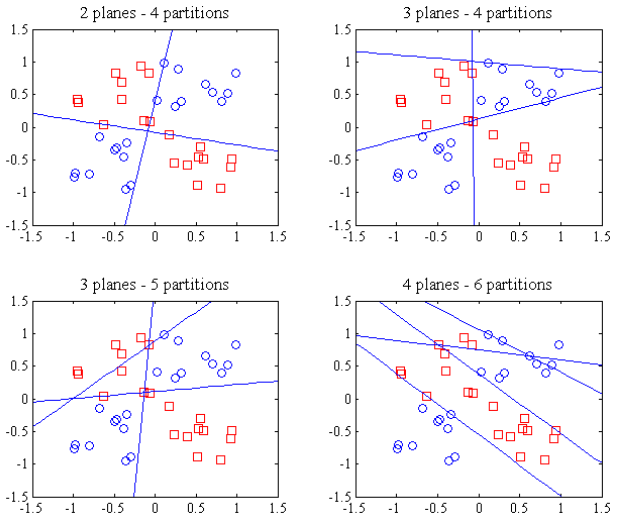
f = *features*

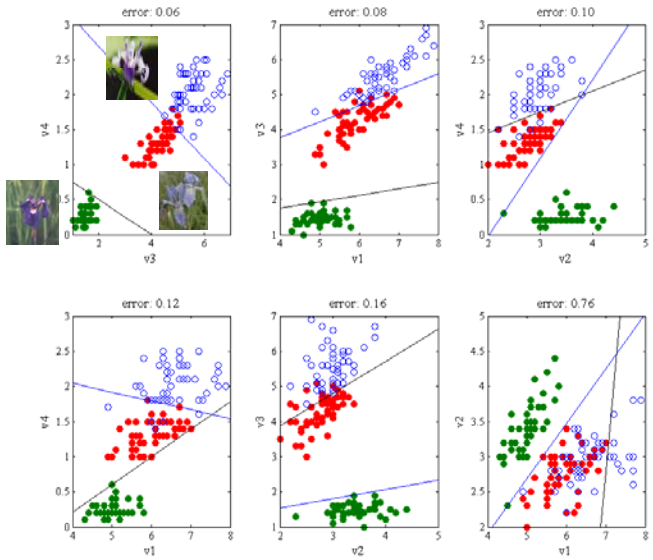
k = *classes*

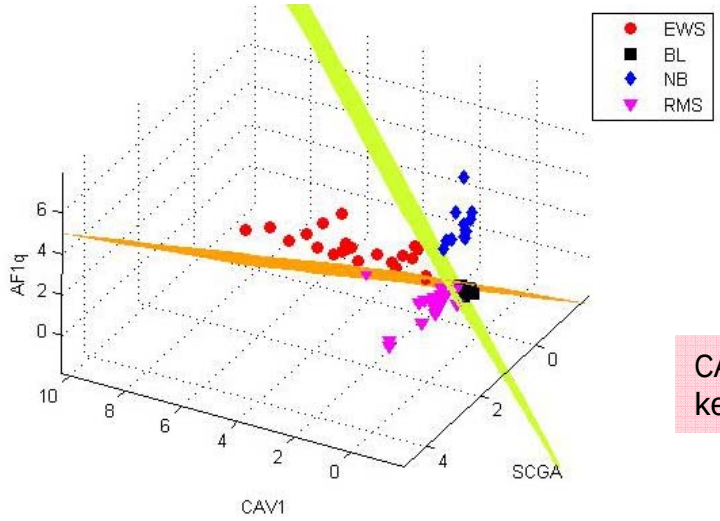
n = *samples*

Feature Selection and Model Complexity

Oblique Multicategory Decision Tress

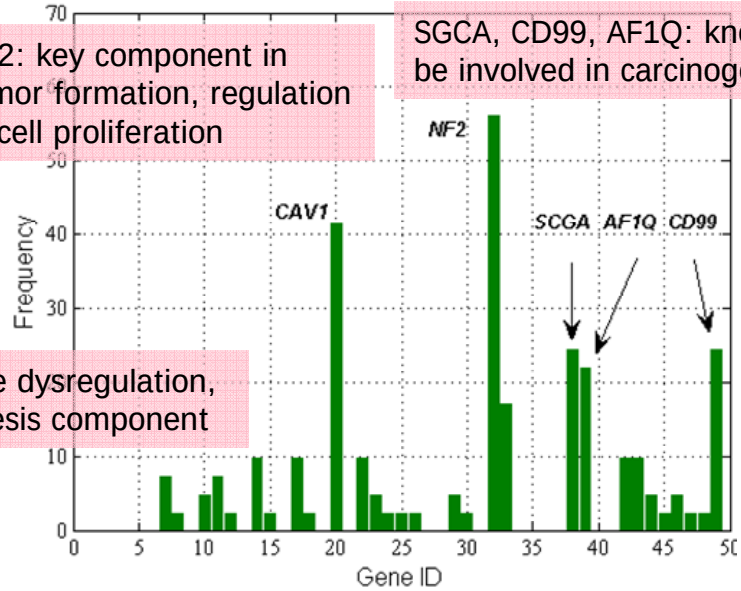






NF2: key component in tumor formation, regulation of cell proliferation

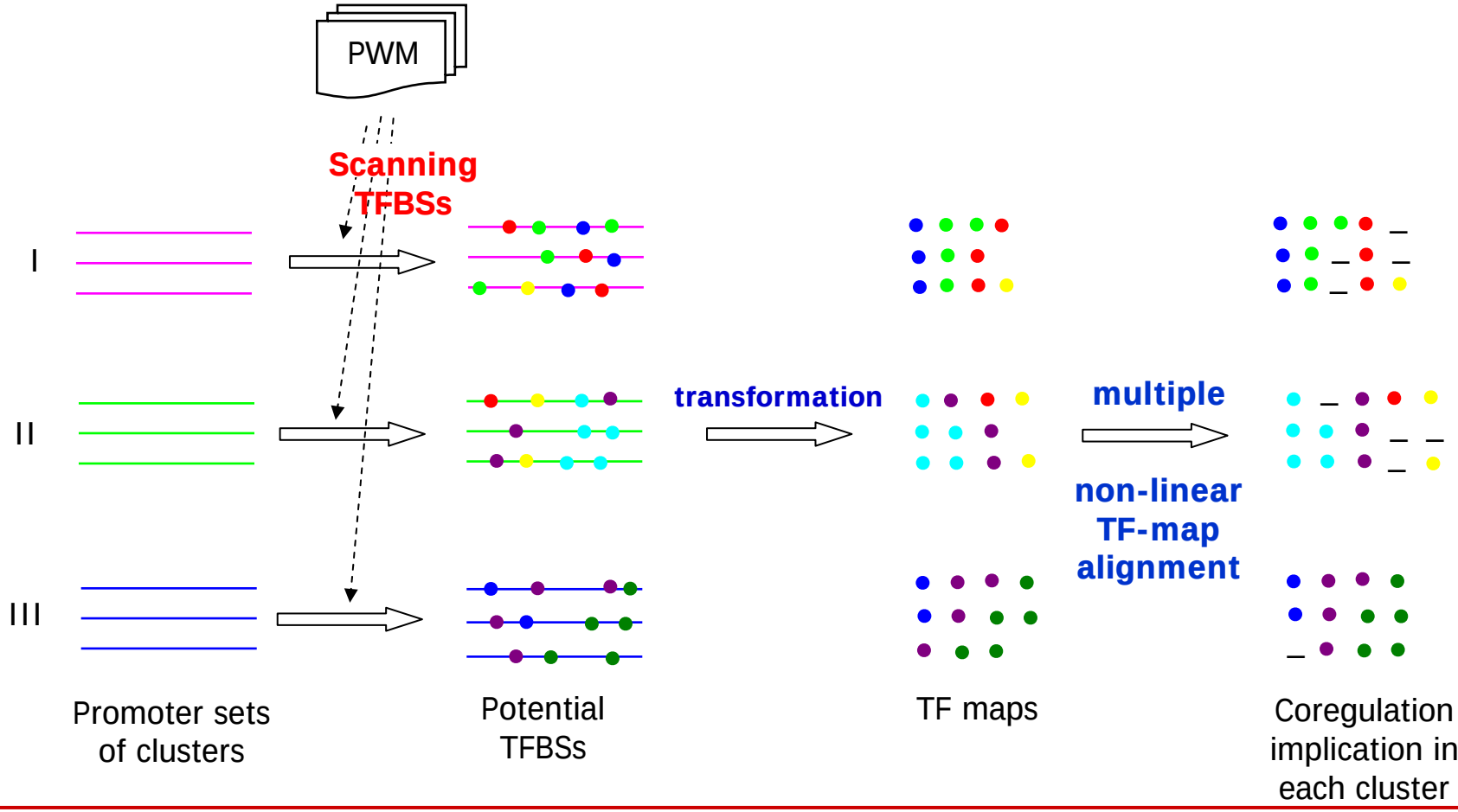
SGCA, CD99, AF1Q: known to be involved in carcinogenesis



CAV1: cell cycle dysregulation, key tumorigenesis component

Sample Selection to Improve Clusterability

- Hypothesis: the more similar the promoter regions, the higher the possibility of coregulation



Sample Selection to Improve Clusterability

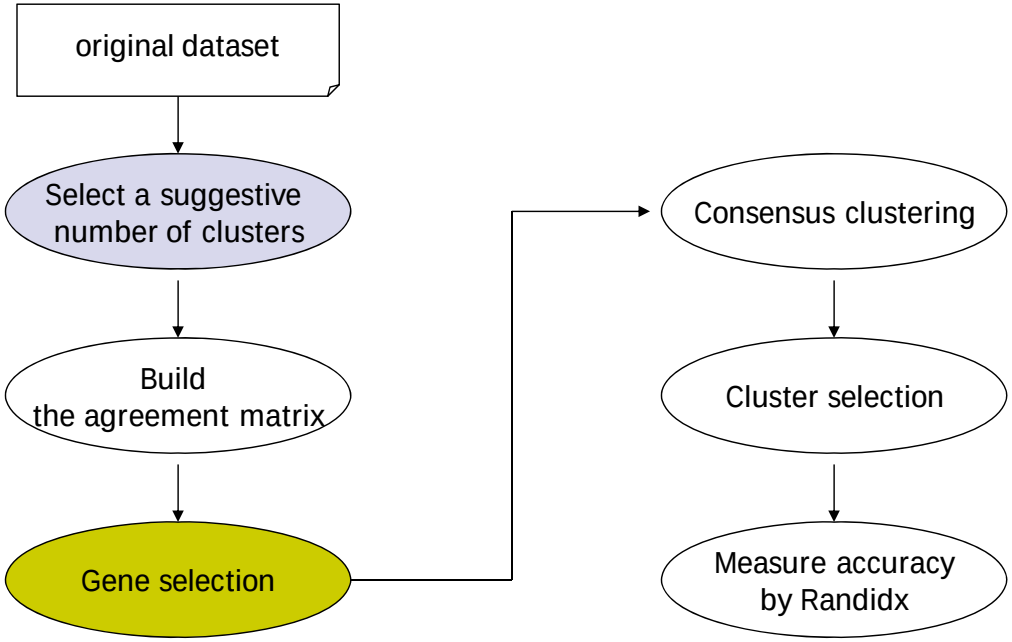
Consensus Clustering

Traditional clustering assumes that all samples must belong to classes

We explore the hypothesis that not all data should be clusterable but that a subset exists composed of all the pairs of samples that show a higher probability of either

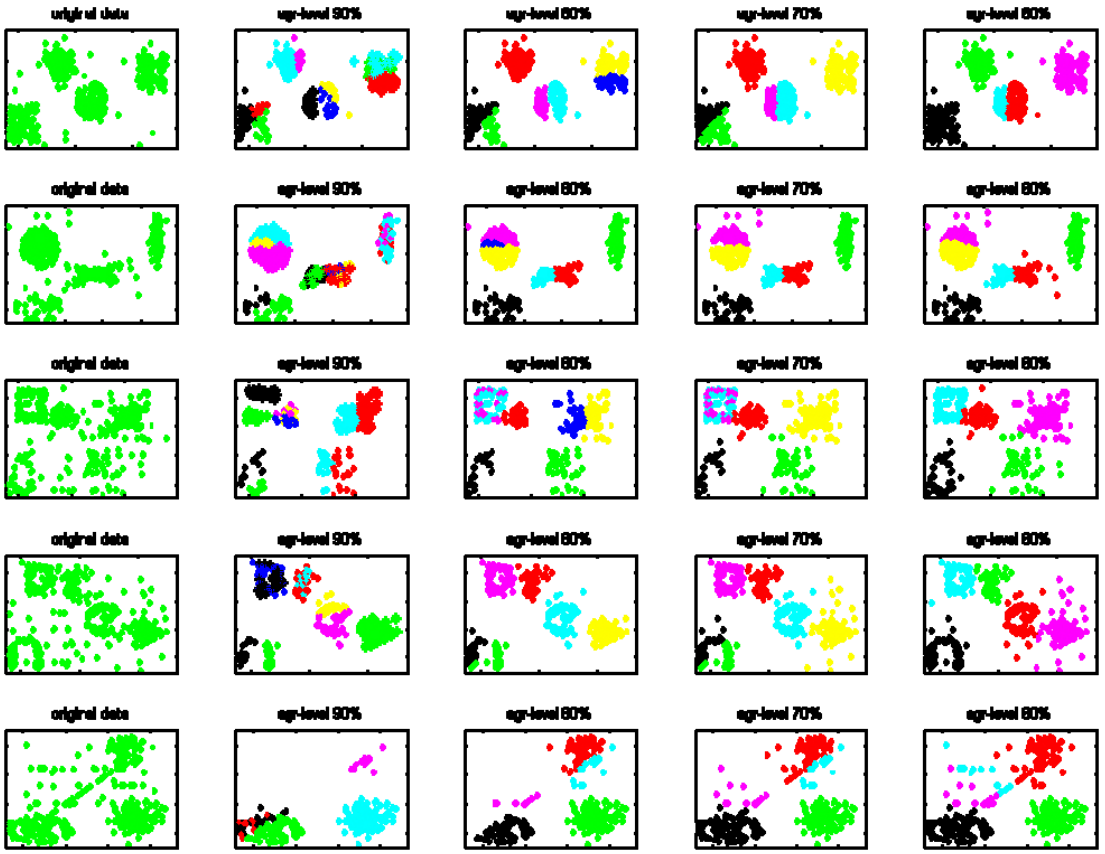
- Belonging to the same cluster, or
- not belonging to the same cluster

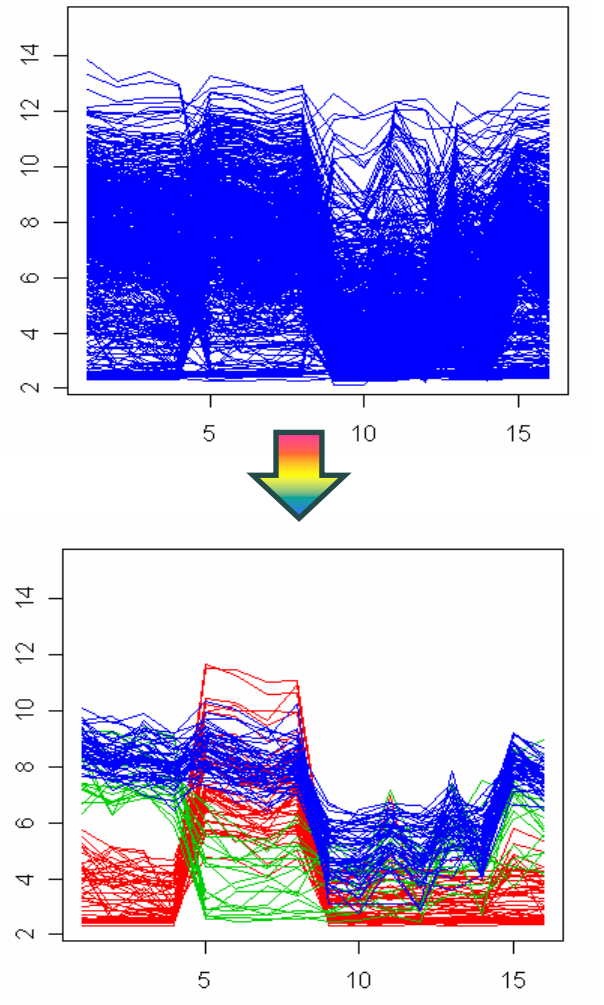
This “clusterable” subset of samples can potentially have a high probability of being relevant in terms of a coherent response



Sample Selection to Improve Clusterability

Consensus Clustering

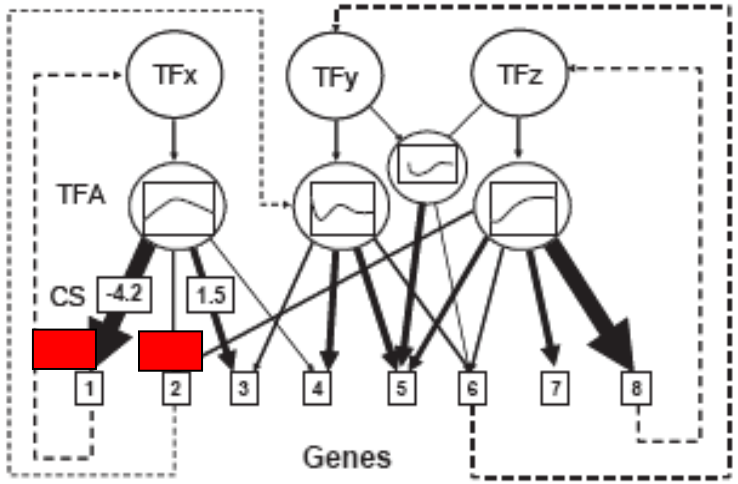




Sample selection

Synthesis and Analysis of Regulatory Networks

Transcriptionally regulated responses can be controlled by appropriate manipulation of critical putative targets



$$\min \sum_i \sum_t e^+(i,t) + e^-(i,t)$$

subject to

$$E(i,t) - \sum_j \pi(i,j) P^{eff}(i,j,t) = e^+(i,t) - e^-(i,t) \quad \forall i,t$$

$$\sum_j z(j) = m \le N_{TF}$$

$$\sum_j D(i,j) \cdot z(j) \ge 1 \quad \forall i$$

$$-r(i,j)M - P(j,t) \le P^{eff}(i,j,t) \le r(i,j)M - P(j,t) \quad \forall i,j,t$$

$$[r(i,j)-1]M + P(j,t) \le P^{eff}(i,j,t) \le [1-r(i,j)]M + P(j,t) \quad \forall i,j,t$$

$$z(j)P^{min} \le P(j,t) \le z(j)P^{max} \quad \forall j,t$$

$$\sum_{j \in N^k} z(j) - \sum_{j \in B^k} z(j) \le |N^k| - 1$$

$$N^k = \{j \mid z^k(j) = 1\}, \; B^k = \{j, \mid z^k(j) = 0\}$$

$$D(i,j) = \begin{cases} 1 & \pi(i,j) \ne 0 \\ 0 & \pi(i,j) = 0 \end{cases} \quad \forall i,j$$

$$P(j,t), P^{eff}(i,j,t) \in \Re$$

$$e^+(i,t), e^-(i,t) \in \Re^+ \quad \forall i,j,t$$

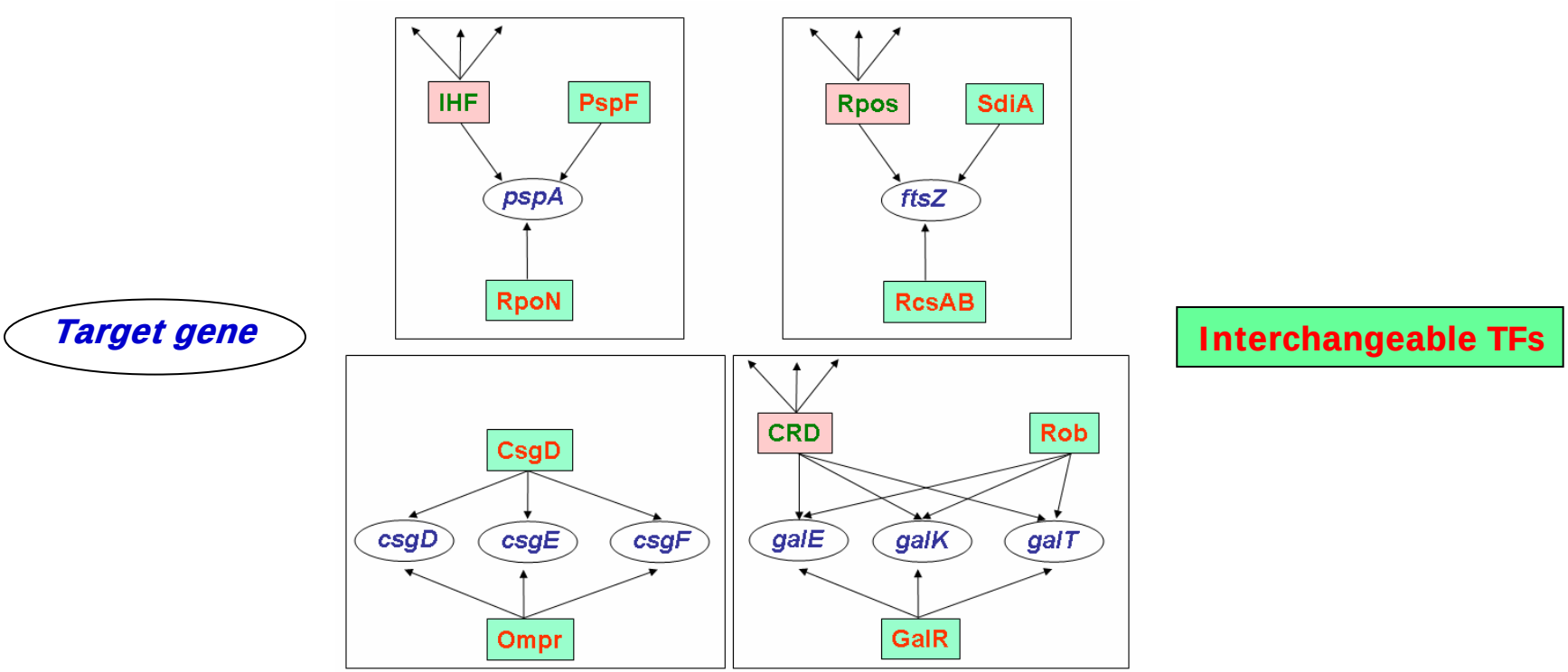
$$z(j), r(i,j) \in \{0,1\} \quad \forall i,j$$

$$i = 1,\ldots,N_g; j = 1\ldots,N_{TF}; t = 1\ldots N_T$$

Structurally Equivalent Modules of Regulatory Control

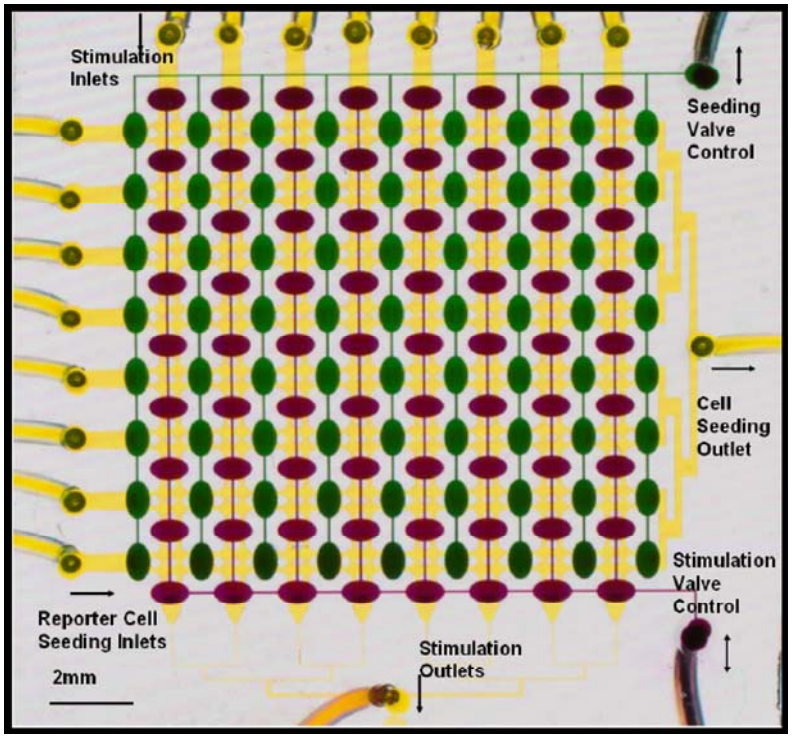
Knock-out experiments have demonstrated that equivalent structural alternatives are available to the cell largely contributing to the apparent robustness of biological systems, Kitano, *Nature* (2004)

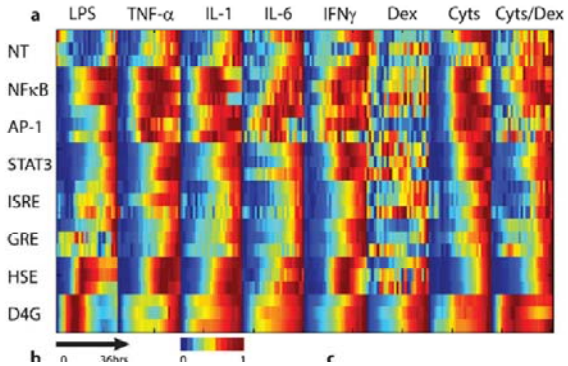
Integer cuts allow for the systematic generation of potentially equivalent structural alternatives

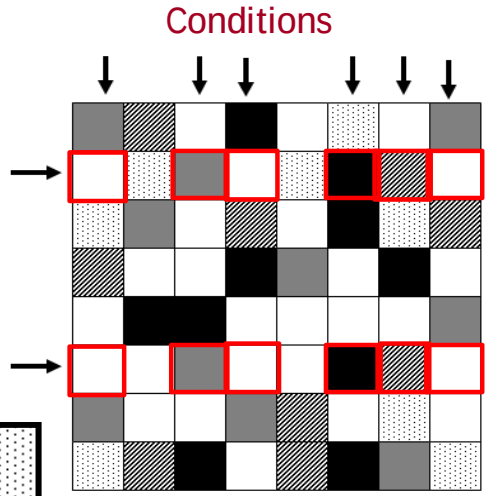


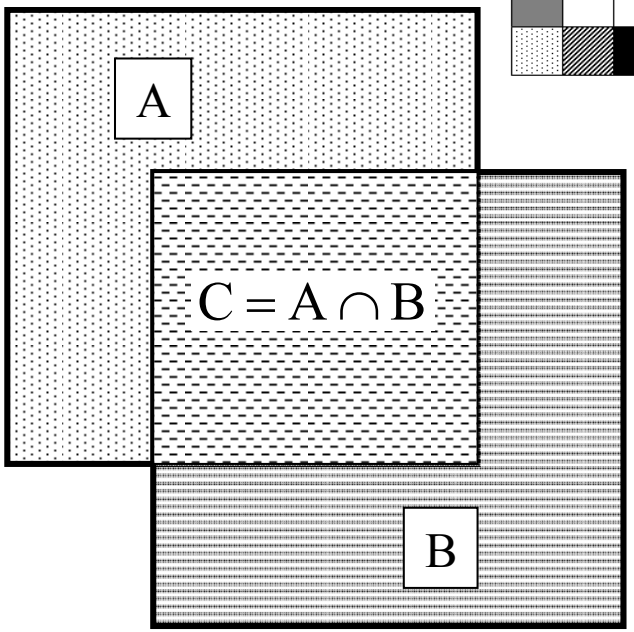
Network Reconstruction

Overlapping Biclustering









Network Reconstruction

Overlapping Biclustering

$$\max : \sum u_k$$

$$s.t : \sum \lambda_i = N$$

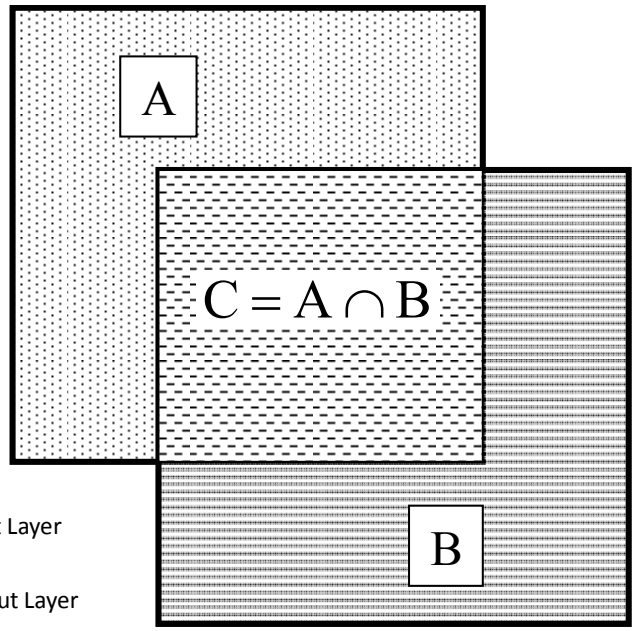
$$: [(\lambda_i + \lambda_j + \mu_k) - 3] * M \le (\lambda_i + \mu_k) * D(i,k) - (\lambda_j + \mu_k) * D(j,k)$$

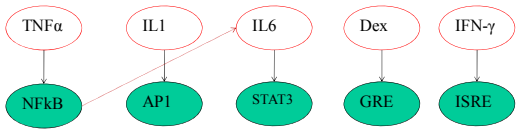
$$: [3 - (\lambda_i + \lambda_j + \mu_k)] * M \ge (\lambda_i + \mu_k) * D(i,k) - (\lambda_j + \mu_k) * D(j,k)$$

$$: \sum_Q u_k - \sum_P u_k > \sum u_k$$

$$: P = \{i \mid u_k^* = 1\}$$

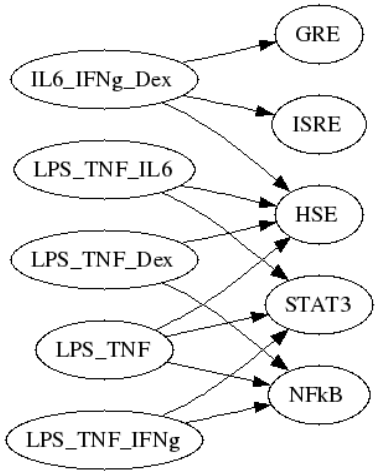
$$: Q = \{i \mid u_k^* = 0\}$$

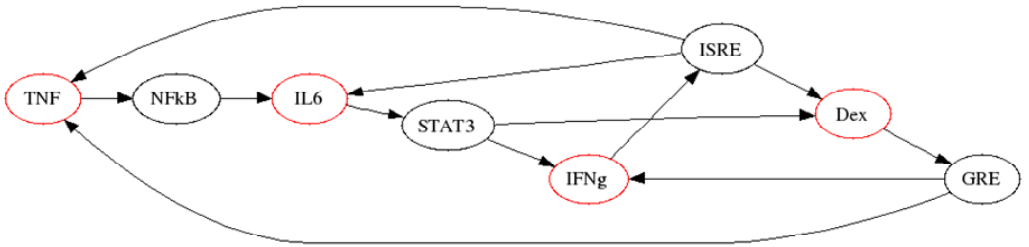




Input Layer

Output Layer





DAG Representation after post processing

Network Quantification

Deconvolution of Dynamics

Non-linear Strengths

$$\begin{bmatrix} dTF_1 & dTF_1 & dTF_1 & dTF_1 \\ dTF_2 & dTF_2 & dTF_2 & dTF_2 \\ dTF_3 & dTF_3 & dTF_3 & dTF_3 \\ dTF_4 & dTF_4 & dTF_4 & dTF_4 \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} & A_{13} & A_{14} \\ A_{21} & A_{22} & A_{23} & A_{24} \\ A_{31} & A_{32} & A_{33} & A_{34} \\ A_{41} & A_{42} & A_{43} & A_{44} \end{bmatrix} \begin{bmatrix} TF_1 & TF_1 & TF_1 & TF_1 \\ TF_2 & TF_2 & TF_2 & TF_2 \\ TF_3 & TF_3 & TF_3 & TF_3 \\ TF_4 & TF_4 & TF_4 & TF_4 \end{bmatrix} + \beta \begin{bmatrix} step & 0 & 0 & 0 \\ 0 & step & 0 & 0 \\ 0 & 0 & step & 0 \\ 0 & 0 & 0 & step \end{bmatrix}$$

Forcing Function

Conditions (different stimuli)

$$\min : \sum_{i=1}^{ng} \sum_{j=1}^{nc} \sum_{t=1}^{nt} \epsilon^{+}(i,j,t) + \epsilon^{-}(i,j,t)$$

$$s.t.$$

$$dTF(i,j,t) - \left[\sum_{i'=1}^{ng} f(i,i',t) * TF(i',j,t) + \beta(j) * s(i,j) \right] - \epsilon^{+}(i,j,t) + \epsilon^{-}(i,j,t) = 0$$

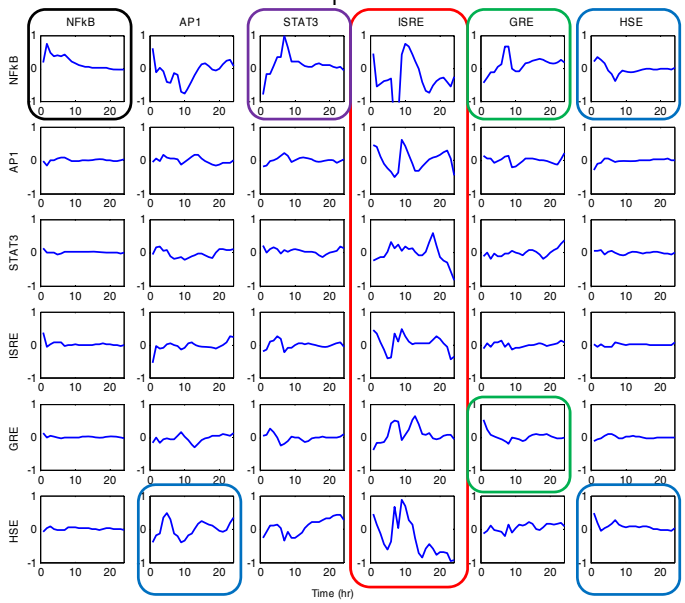
$$f(i,i',t) \le M * N(i,i')$$

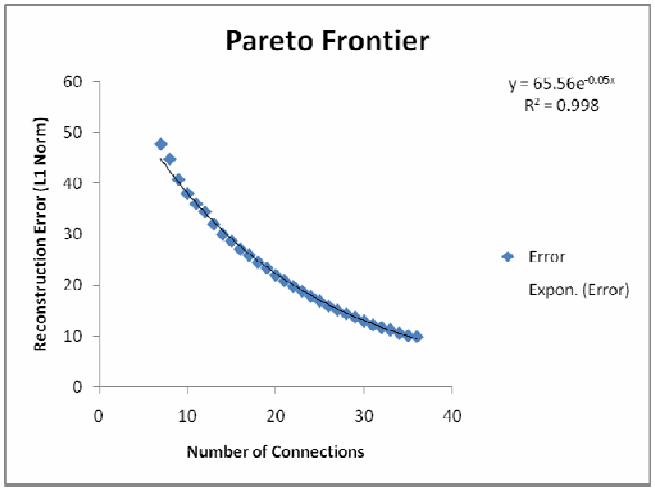
$$f(i,i',t) \ge -M * N(i,i')$$

$$\sum_{i=1}^{ng} N(i,i') \ge 1$$

$$\sum_{i'=1}^{ng} N(i,i') \ge 1$$

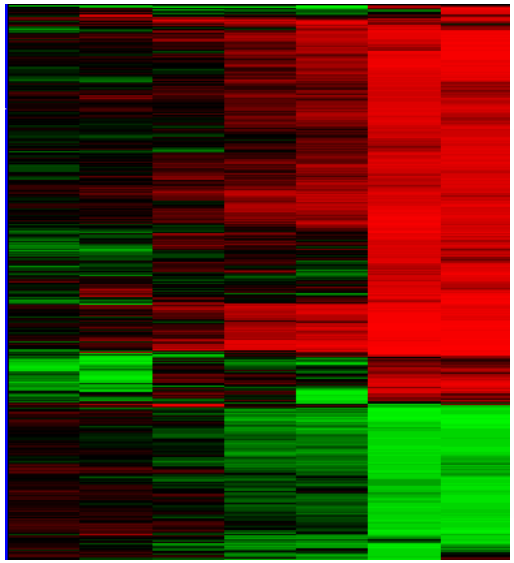
$$\sum_{i}^{ng} \sum_{i'}^{ng} N(i,i') = T$$

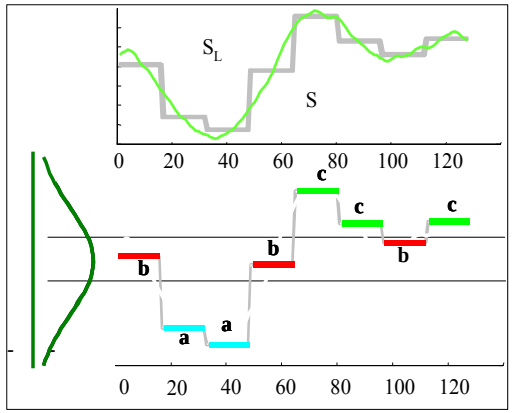


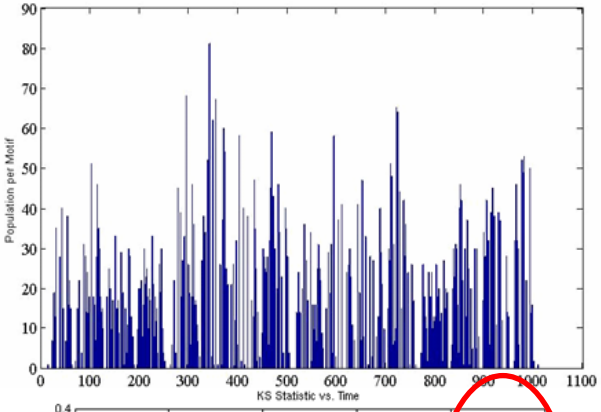


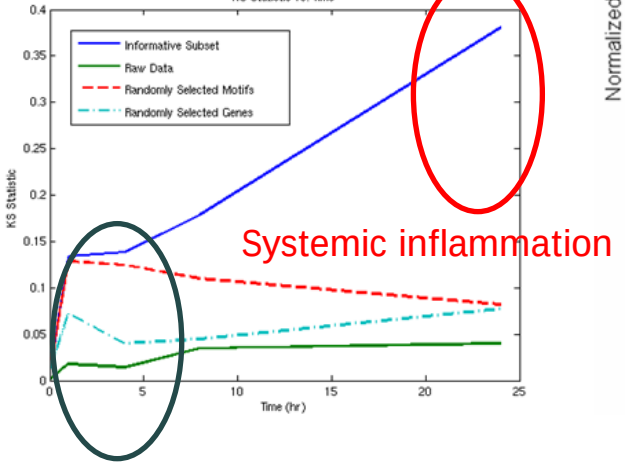
Intrinsic Dynamics and Essential Responses

Clustering & Selection in Multidimensional Temporal Data



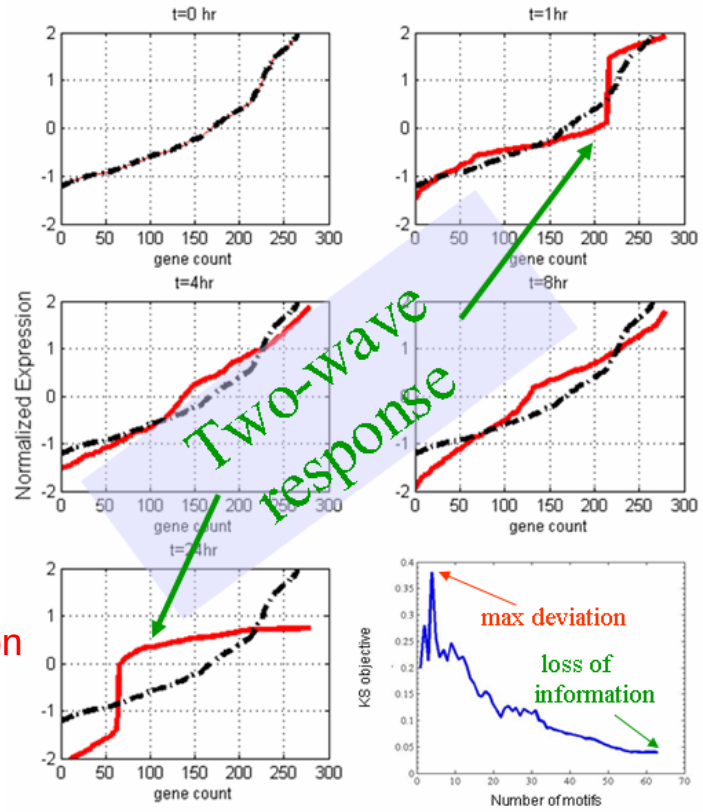






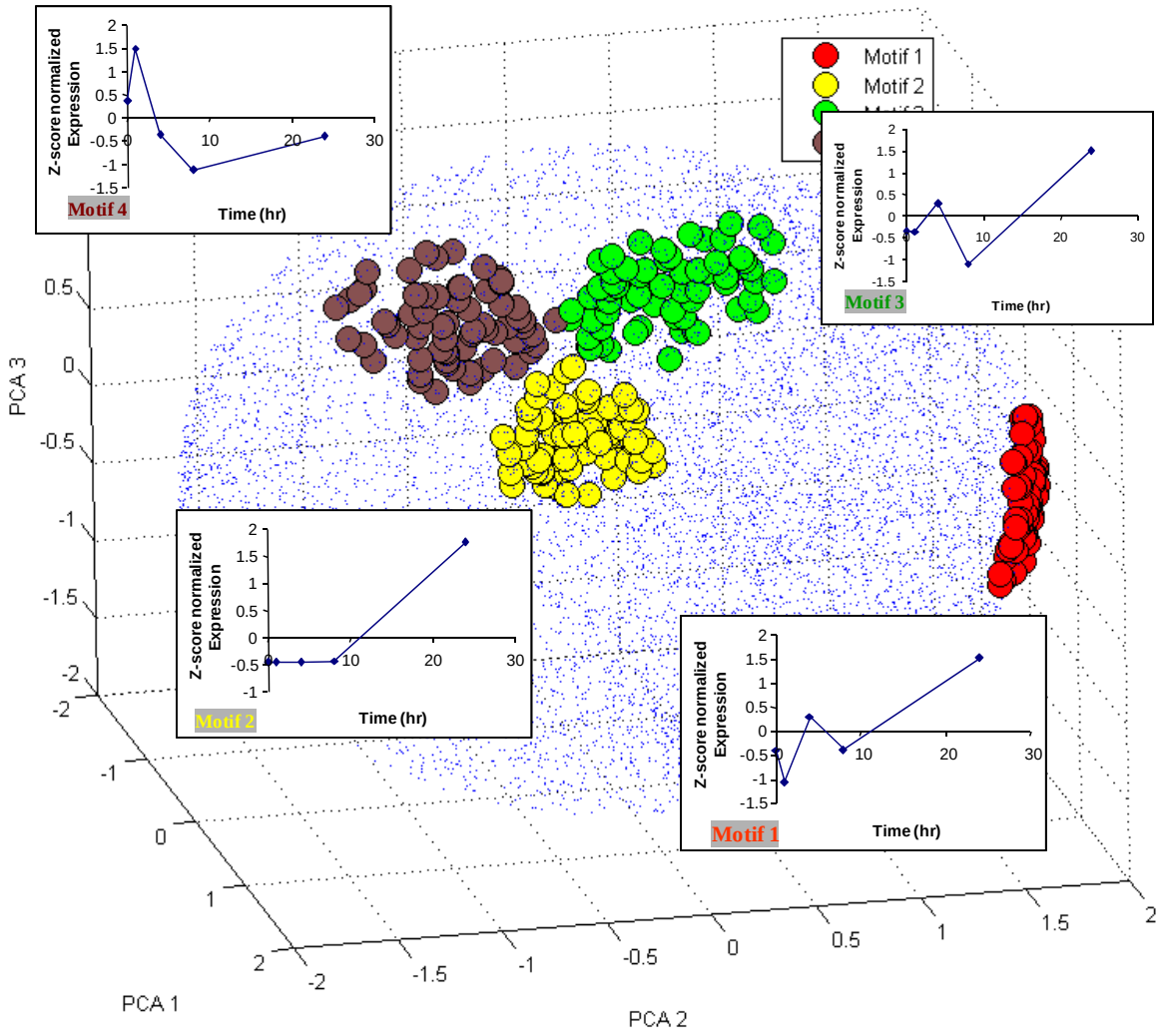
Systemic inflammation

Acute response



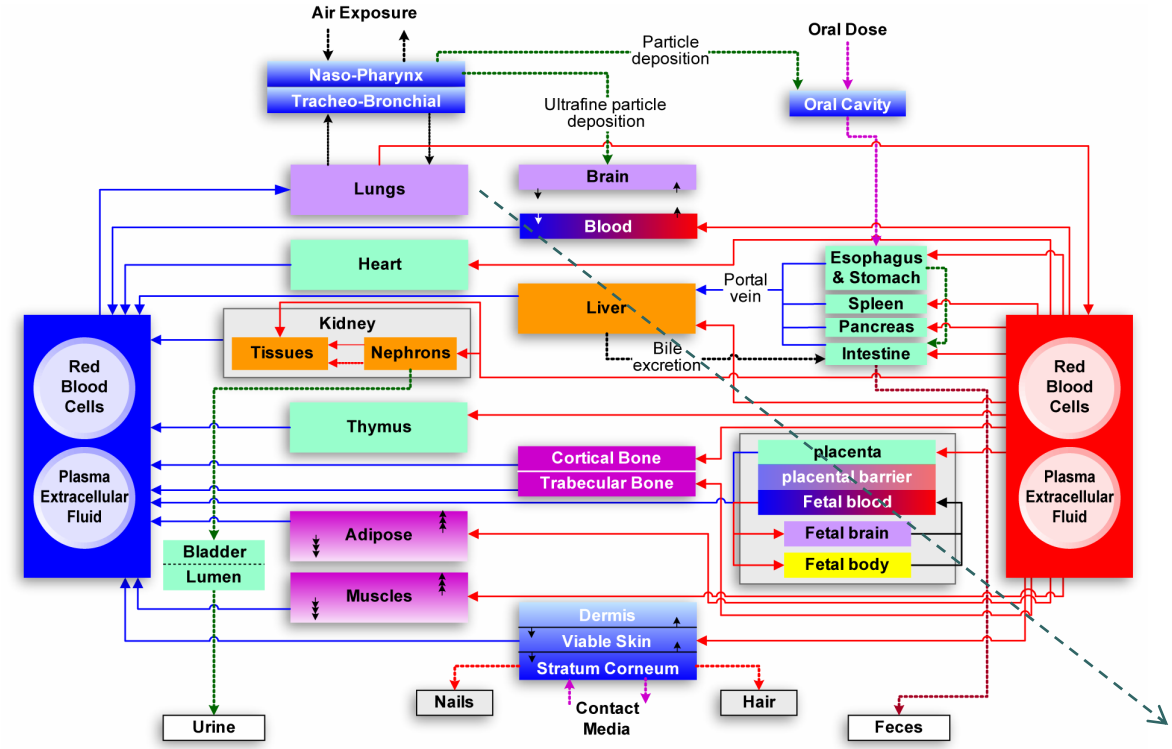
Intrinsic Dynamics and Essential Responses

Clustering & Selection in Multidimensional Temporal Data

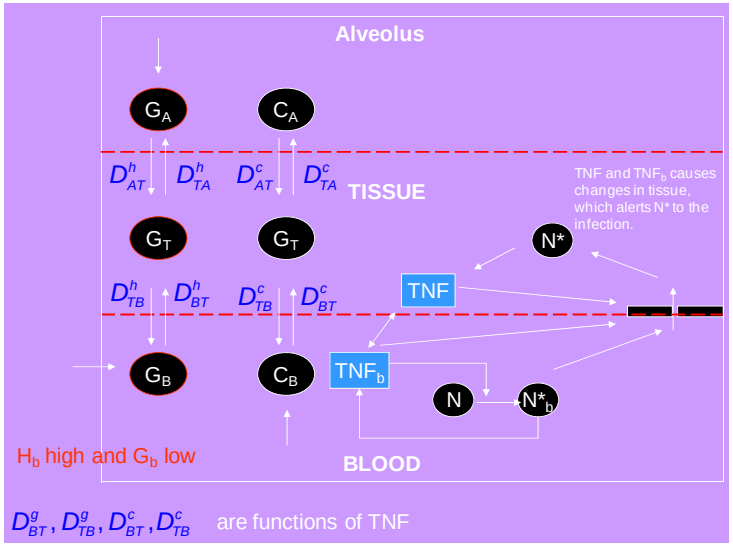


Global Dynamic Models

Exploring Global Transcriptional Dynamics

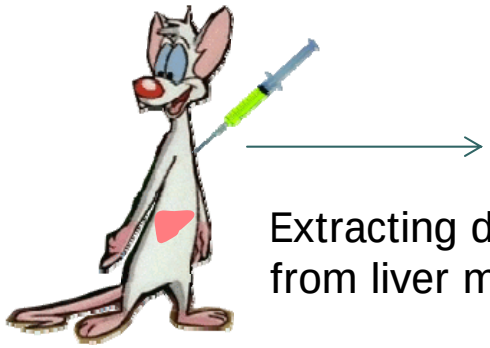


Combining tissue -specific transcriptional dynamics and PBPK models

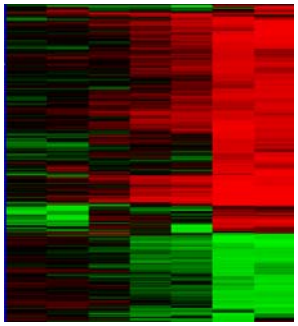


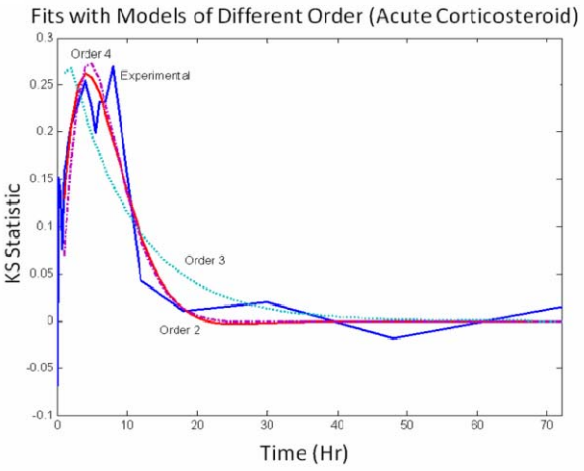
Global Dynamic Models

Exploring Global Transcriptional Dynamics



Extracting dynamics
from liver microarray



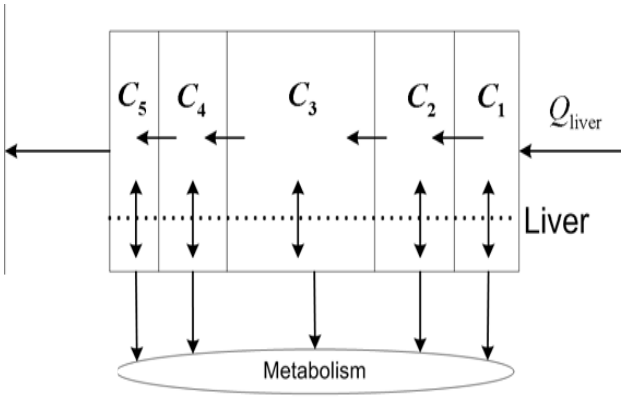


$$\sum_{n=1} K_n \frac{d^{(n)}C}{dt} = f(x)$$

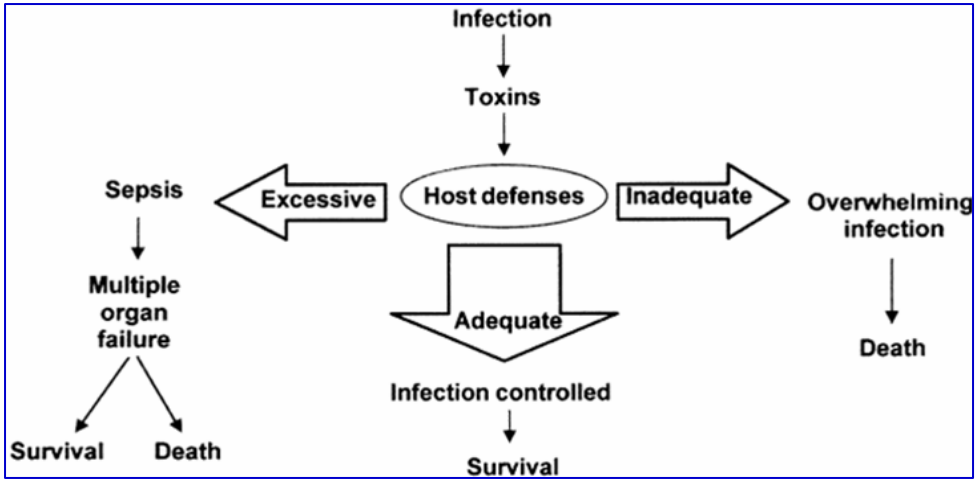
C = Compartment

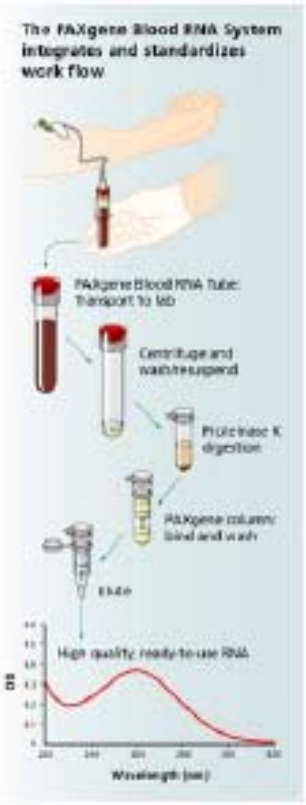
K = Reparameterized Constants

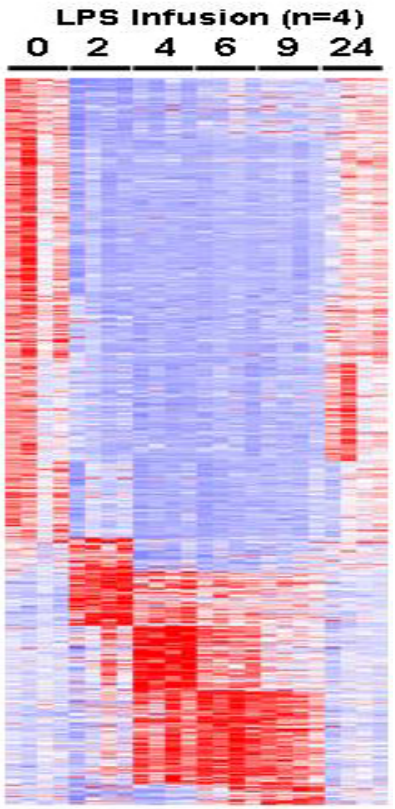
$f(x)$ = stimulus profile



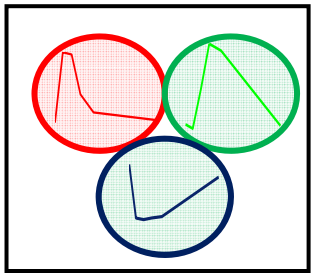
Reverse Engineering of Mechanistic-based Models







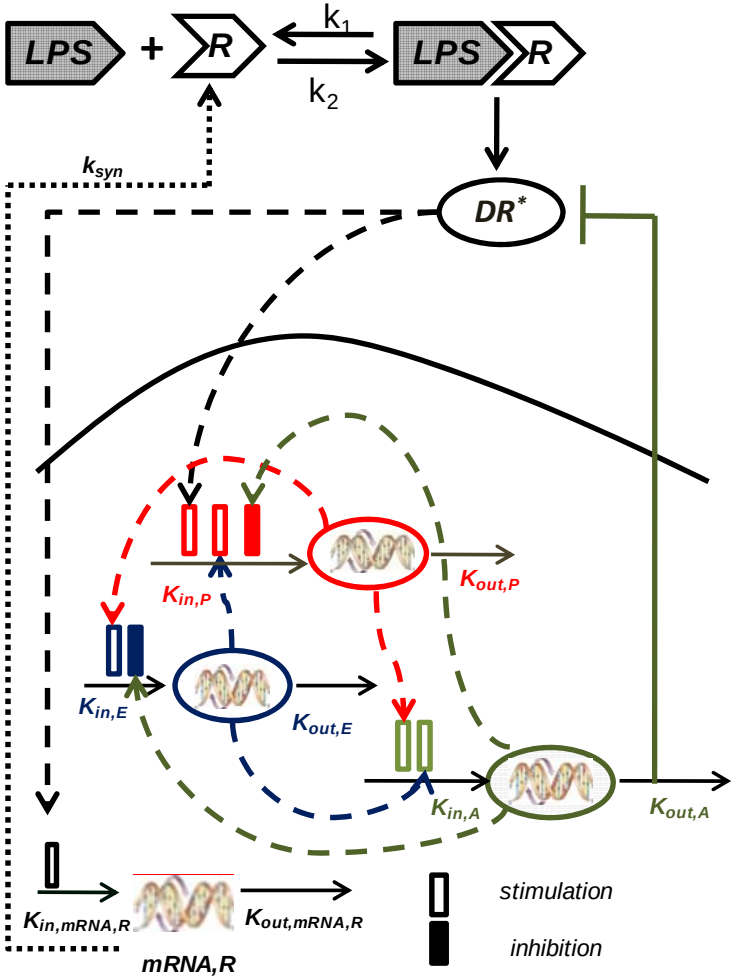
Pro–inflammatory Response



Energetic Response
Cellular bioenergetics

Anti–inflammatory Response
Resolution of inflammation

Reverse Engineering of Mechanistic-based Models



$$\frac{dLPS}{dt} = k_{lps,1} \cdot LPS \cdot (1 - LPS) - k_{lps,2} \cdot LPS$$

$$\frac{dR}{dt} = k_{syn} \cdot mRNA,R + k_2 \cdot (LPS - R) - k_1 \cdot LPS \cdot R - k_{syn} \cdot R$$

$$\frac{dmRNA,R}{dt} = K_{in,mRNA,R} \cdot (1 + H_{mRNA,DR^*}) - K_{out,mRNA,R} \cdot mRNA,R$$

$$\frac{d(LPS - R)}{dt} = k_1 \cdot LPS \cdot R - k_3 \cdot (LPS - R) - k_2 (LPS - R)$$

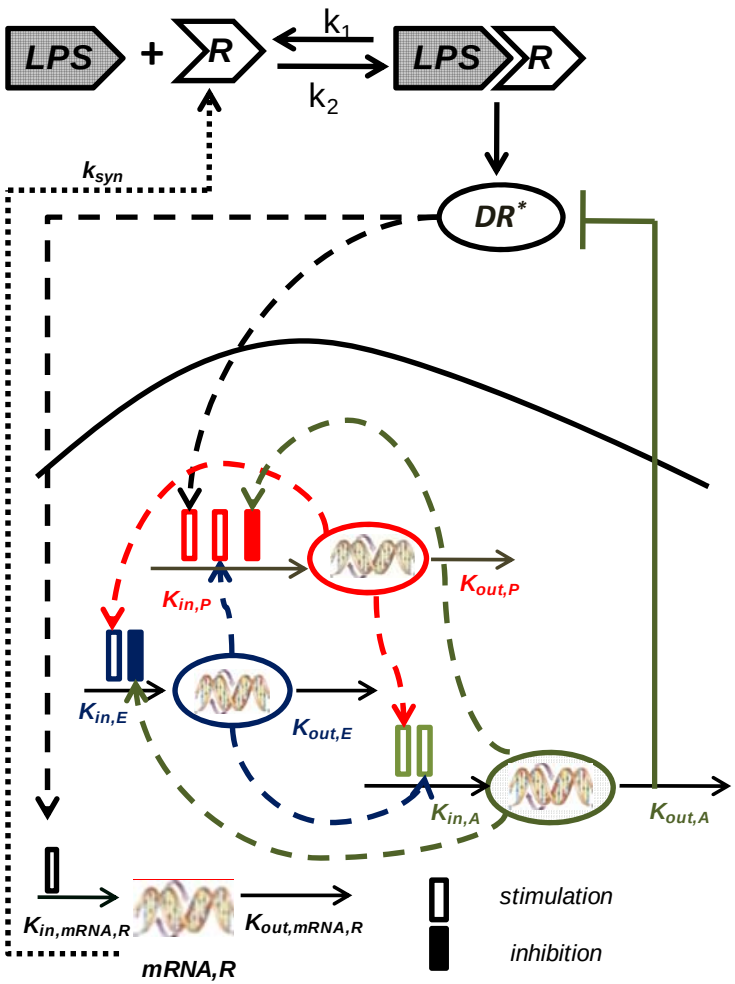
$$\frac{dDR^*}{dt} = k_3 \cdot (LPS - R)/A - k_4 \cdot DR^* + k_c \cdot \left(\frac{DR^{*5}}{1 + DR^{*5}} \right)$$

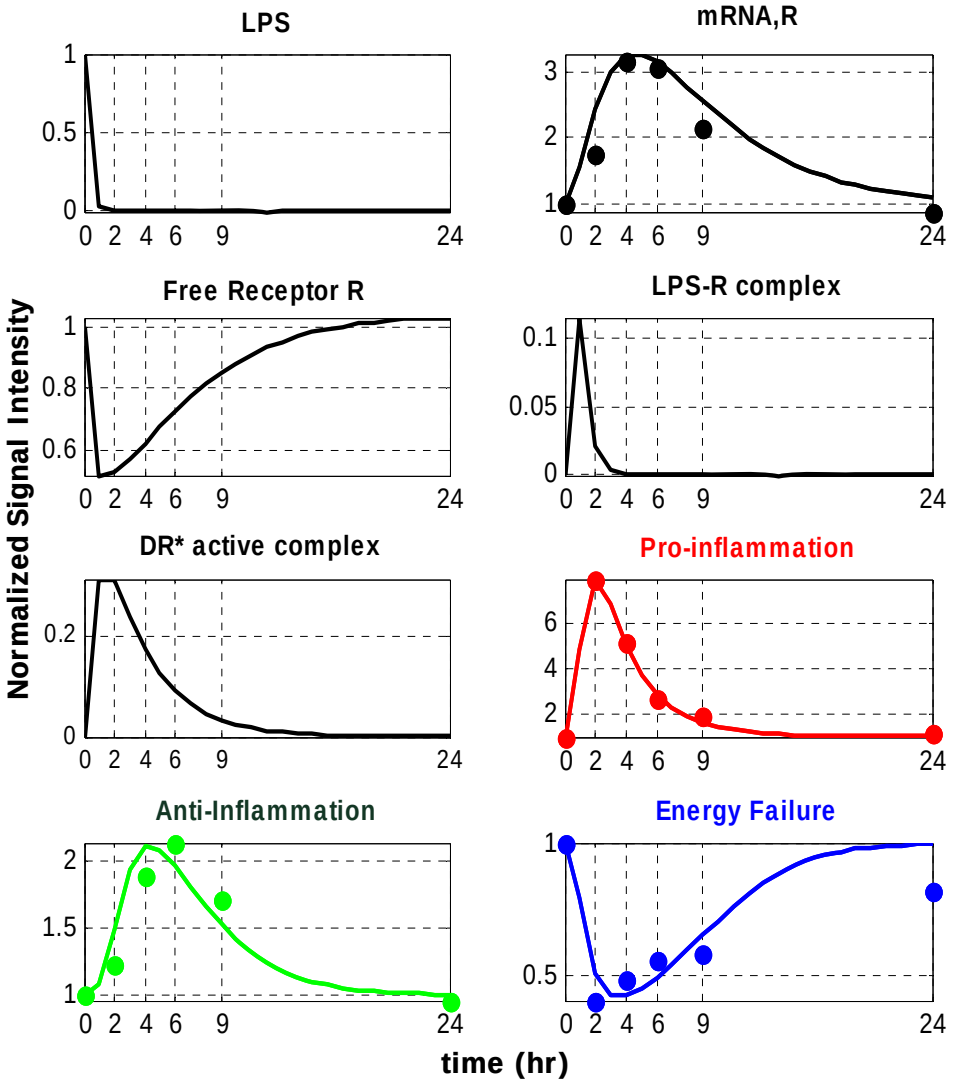
$$\frac{dP}{dt} = (K_{in,P}/A) \cdot (1 + H_{P,DR^*}) \cdot (1 + H_{P,E}) - K_{out,P} \cdot P$$

$$\frac{dA}{dt} = K_{in.A} \cdot (1 + H_{A,P}) \cdot (1 + H_{A,E}) - K_{out,A} \cdot A$$

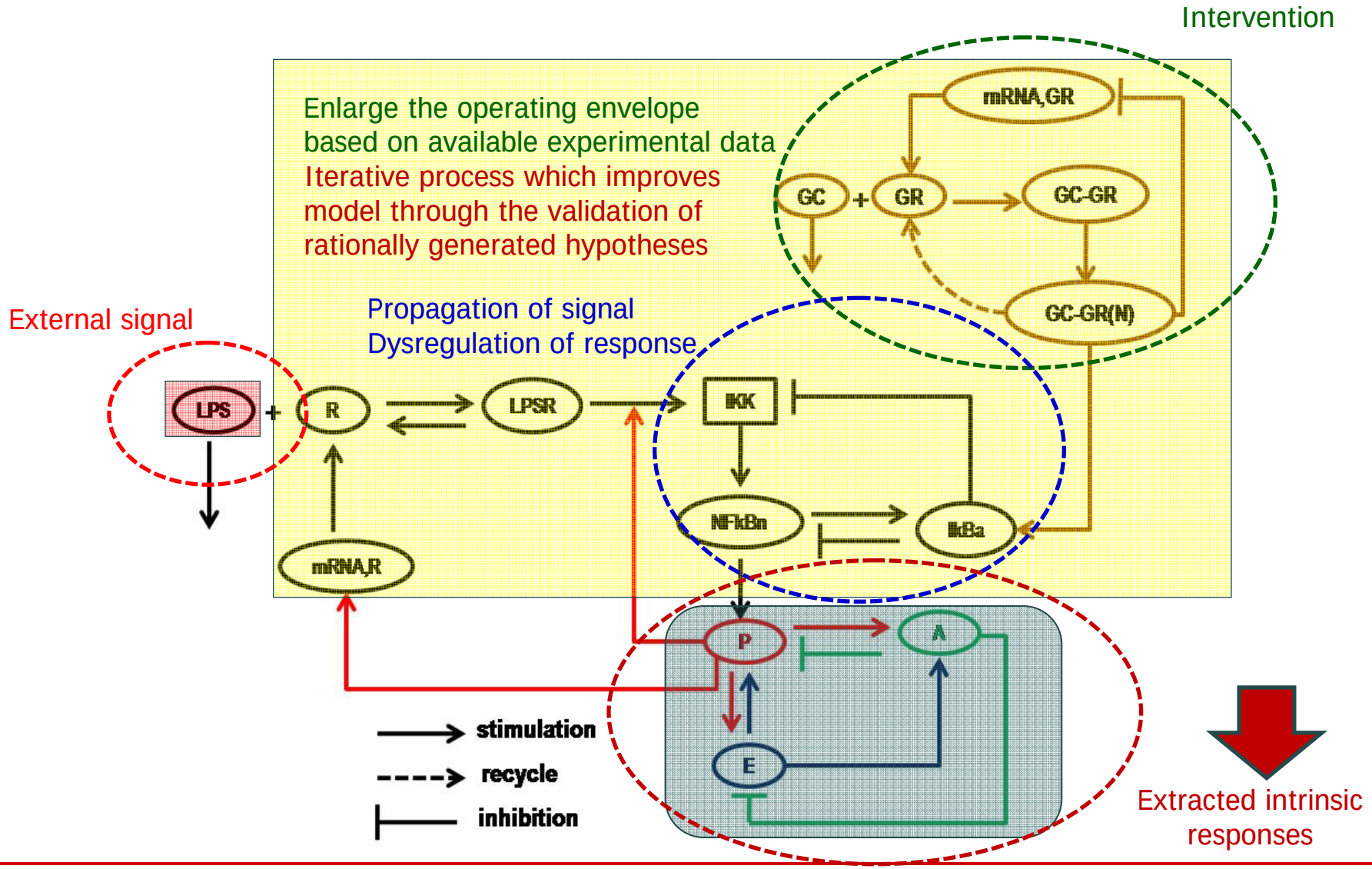
$$\frac{dE}{dt} = (K_{in,E}/A) \cdot (1 + H_{E,P}) - K_{out,E} \cdot E$$

Reverse Engineering of Mechanistic-based Models





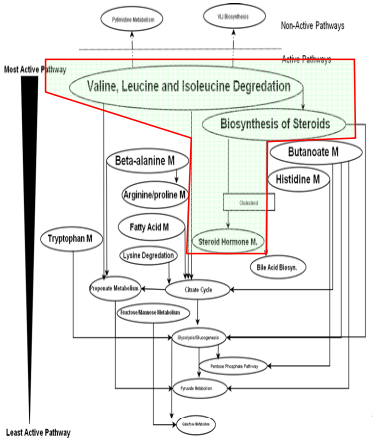
Reverse Engineering of Mechanistic-based Models

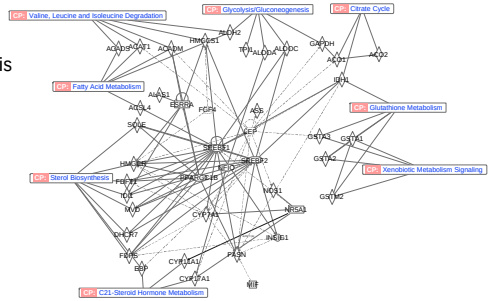


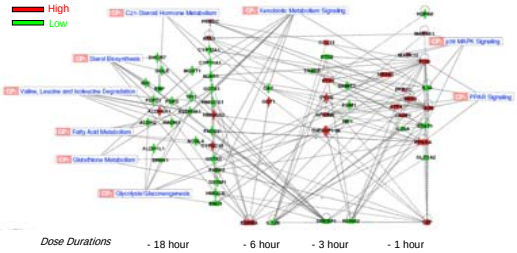
Case Study I: *in utero* exposure to Dibutyl Phthalate (DBP)

Metabolic pathways grow into metabolic networks; DBP affects cholesterol biosynthesis before steroid hormone biosynthesis

Gene networks allow us to predict putative TF's



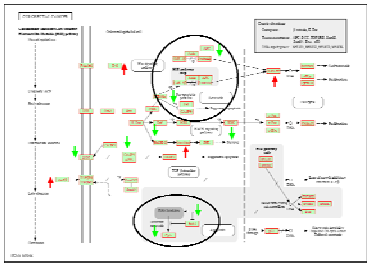




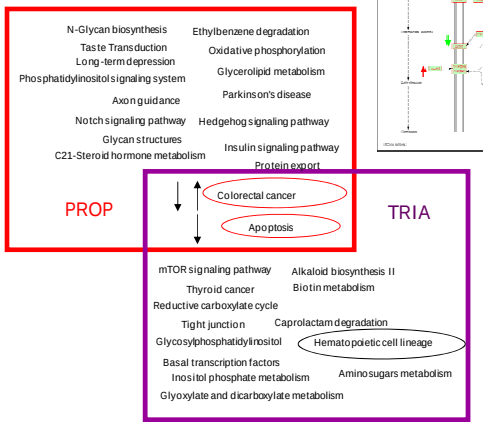
Case Study II: Triazole Conazole Fungicides

Myclobutanil vs. Triadimefon

Myclobutanil vs. Propiconazole

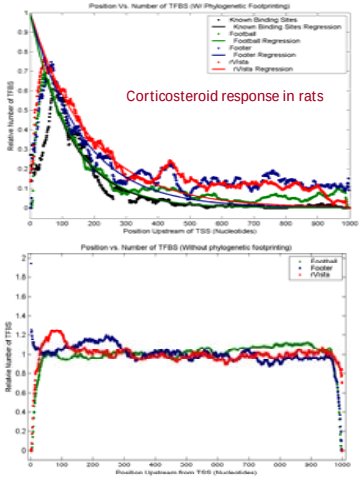


Non-Tumorigenic vs Tumorigenic



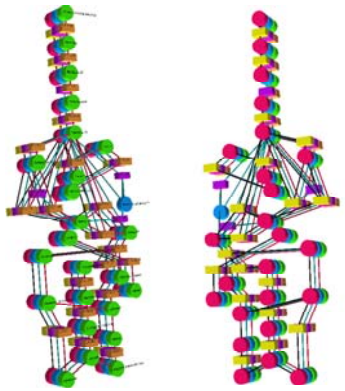
Phylogenetics: Cross-species Extrapolation of MoA

Cross-species promoter conservation



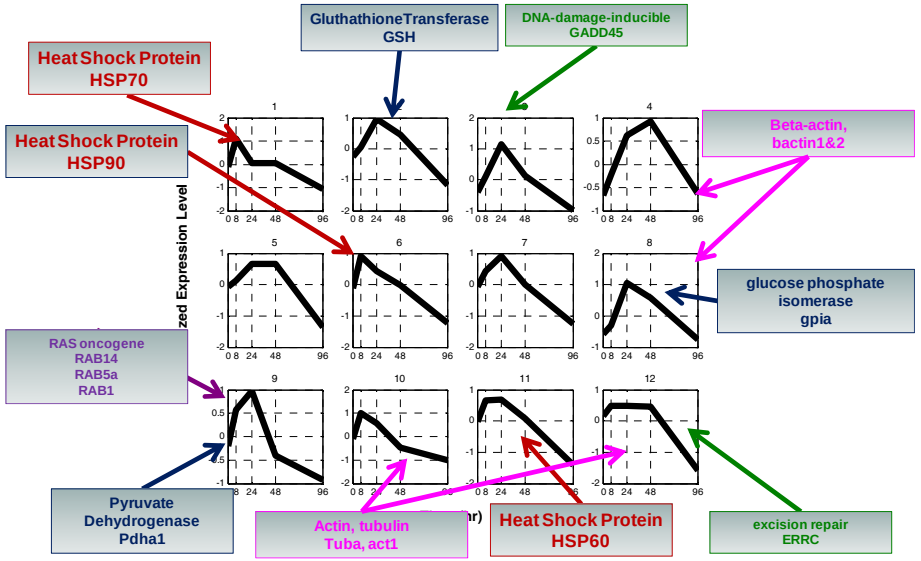
Cross-species pathway similarities

Steroidogenesis in rat, mouse and human



Annals of Biomedical Eng., 2006; unpublished data, 2007

Arsenic Exposure – Zebra Fish

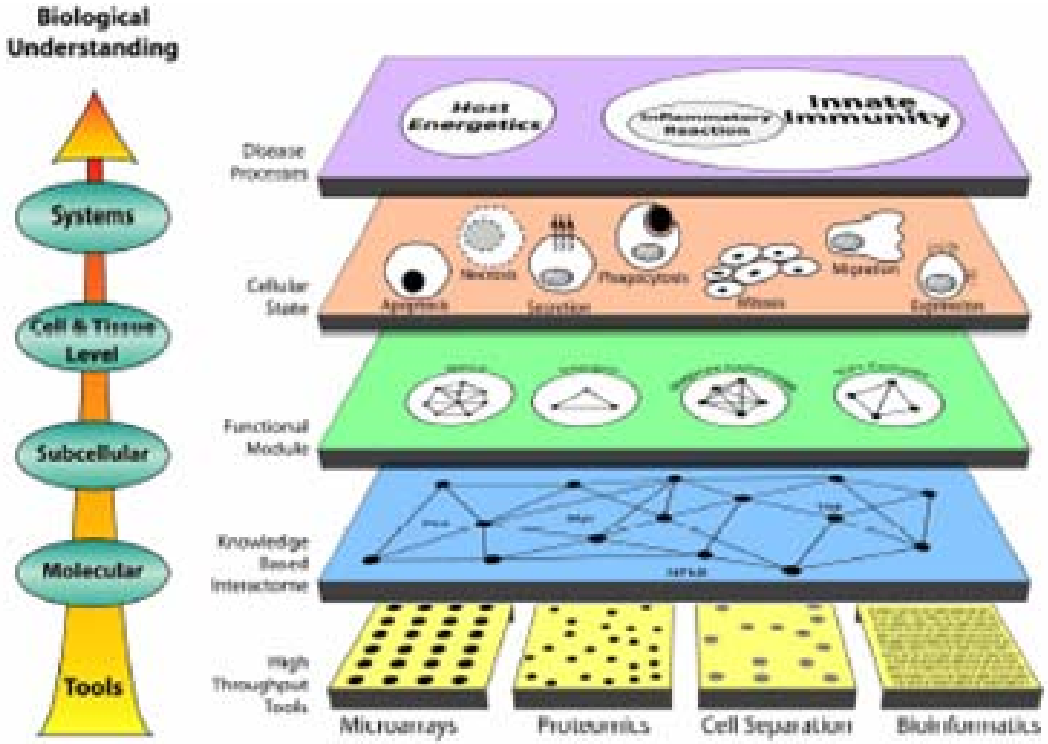


ebCTC – An Integrative Approach

The focus today was on only one aspect of the activities taking place at ebCTC

We have a well-integrated network of interactions

- Physiomics, Toxicokinetics and Toxicodynamics – Georgopoulos
- Systems Level – Androulakis
- Proteomics , Metabolomics and Metabolic Engineering – Floudas, Ierapetritou
- Bio-network Modeling and Dynamics – Rabitz
- Receptors and Molecules - Welsh



Summary and Outlook

A systems approach to toxicogenomics allows the integration of multiple data sources in an attempt to place interpretation of experimental observations in a reasonable context

One the most challenging, yet promising, outcomes would be higher level models that allow developing associations and hypotheses

The examples and methodologies presented emphasized:

- Essential responses and PBPK models
- Context-specific regulators and controls
- Combining expression and relational data
- Cross-species extrapolations of MoA
- Metabolic context of expression data

Main conclusions: Significant opportunities related to optimization and modeling of complex systems and need for high-throughput data generation (multiple disturbances, time course data)

The wish list is well defined (data, promoters, annotations etc.). What we need to promote is the attitude that systems biology is a hypothesis generation framework closely interacting with and guiding experimental design rather that a test bed for algorithm development or software development

Possibilities and Limitations

Despite being in the genomics-era we are still seriously data limited

- We may have more analytical and computational capabilities that we have data ...

Initiatives such as ToxCast™ (www.epa.gov/ncct/toxcast) can have significant impact

Relevant data is a critical enabler for any future success

- Relevant in terms of significance
- Relevant in terms of resolution

These activities should embrace and foster close collaboration between scientists and engineers with diverse background

Acknowledgments

http://rci.rutgers.edu/~yannis/publications.html
http://ccl.rutgers.edu/ebCTC/publications.html

Rutgers University/UMDNJ

- Eric Yang, Pegy Foteinou, Meric Ovacik, Kai He, Nguyen Tung

Harvard Medical School & Massachusetts General Hospital

- Prof. Maish Yarmush (RU/CEM), Prof. François Betrhiaume

Biomedical Sciences, SUNY Buffalo

- Prof. Richard Almon, Prof. Bill Jusko, Prof. Debra Dubois

Department of Surgery, University of Medicine & Dentistry of New Jersey

- Prof. Stephen Lowry, Prof. Steve Calvano

EPA/RTP

- Dr. S. Euling, Dr. K Gaido, Dr. S. Hester, Dr. B. Sen

$$$ Funding $$$

EPA ebCTC environmental bioinformatics and Computational Toxicology Center

Consortium Members

UMDNJ Robert Wood Johnson Medical School — University of Medicine & Dentistry of New Jersey

Computational Chemodynamics Laboratory, Environmental & Occupational Health Sciences Institute
Department of Environmental & Occupational Medicine
Department of Pharmacology
Informatics Institute

The State University of New Jersey RUTGERS

Department of Biomedical Engineering
Department of Chemical & Biochemical Engineering
Department of Environmental Sciences
Department of Statistics

Princeton University

Computer Aided Systems Laboratory, Department of Chemical Engineering
Department of Chemistry
Program in Applied and Computational Mathematics

FDA U.S. Food and Drug Administration

Center for Toxicoinformatics, National Center for Toxicological Research

NSF, ONR, Busch Biomedical Research Award

Thank you!


© Original Artist
Reproduction rights obtainable from
www.CartoonStock.com
HARDIN
"I was close to a breakthrough when the grant money ran out."


© Original Artist
Reproduction rights obtainable from
www.CartoonStock.com
BAR
"I gave up hope of receiving a Genius Grant years ago. Now I'm basically shooting for the less impressive Mediocrity Grant."