Modeling and Reverse Engineering Genetic Networks -

Introduction to Systems Biology

Zoltan Szallasi M.D.

Children's Hospital
Informatics Program,
Harvard Medical School,
Boston, MA

Harvard-MIT Division of Health Sciences and Technology
HST.950J: Medical Computing

Peter Szolovits, PhD
Isaac Kohane, MD, PhD
Lucila Ohno-Machado, MD, PhD

Goals of science:

Predictive power

Understanding

Intellectual entertainment

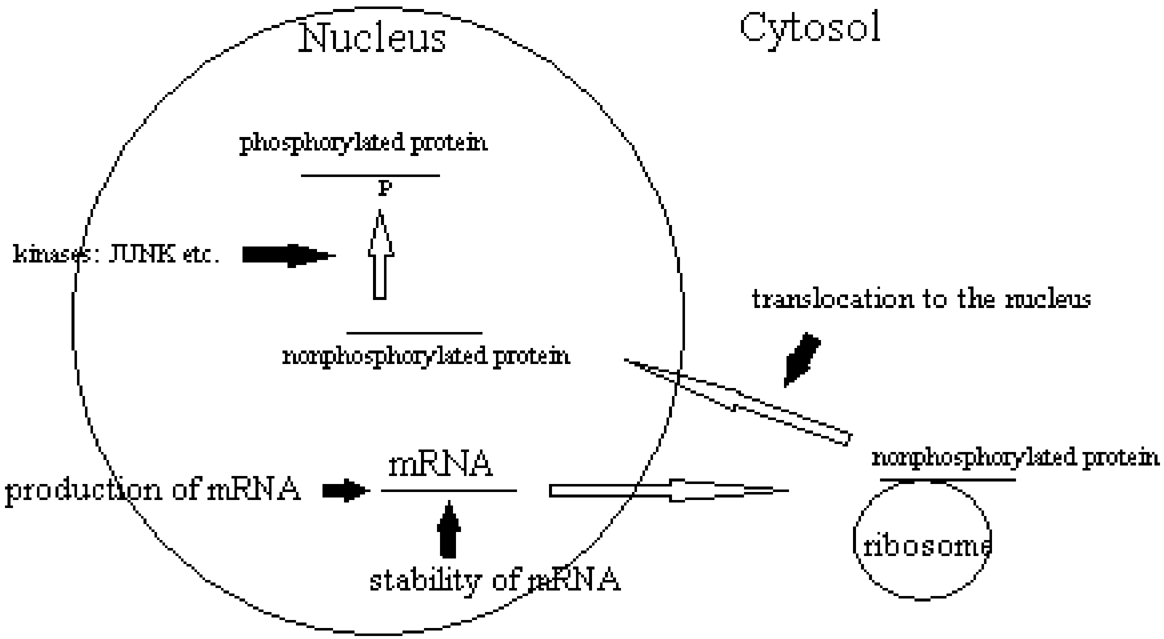
Playground for adults

Modeling:

1. **Conceptual framework/Data about the system**
2. **Model structure (mathematical)**
3. **Pick the best model - parameter fitting**
4. **Model validation**

the conceptual framework of genetic network analysis

Example: Independently regulated derivatives of the c-jun gene



Genetic network - heterogeneous network of interacting variables

Cell (experimental unit) is a network of “gene derivatives” (mRNA, protein) and other biochemical entities.

biological parameters:
They can be defined as a biochemical entity, that:
- can be measured
- is chemically (rather) homogeneous
-determines by itself or in combination with something else the state of another biological parameter.

How many biological parameters ?

Cautious estimate: on the order of 1-2x10$_5$

10,000-20,000 active genes per cell

< 3 posttranslational modifications/protein in yeast

3-6 (?) posttranslational modifications/protein in humans

The number of biological parameters is probably less than 10 times the number of genes

Splice variants <   > modules

Compartmentalization (!!!)

reverse engineering of genetic regulatory networks

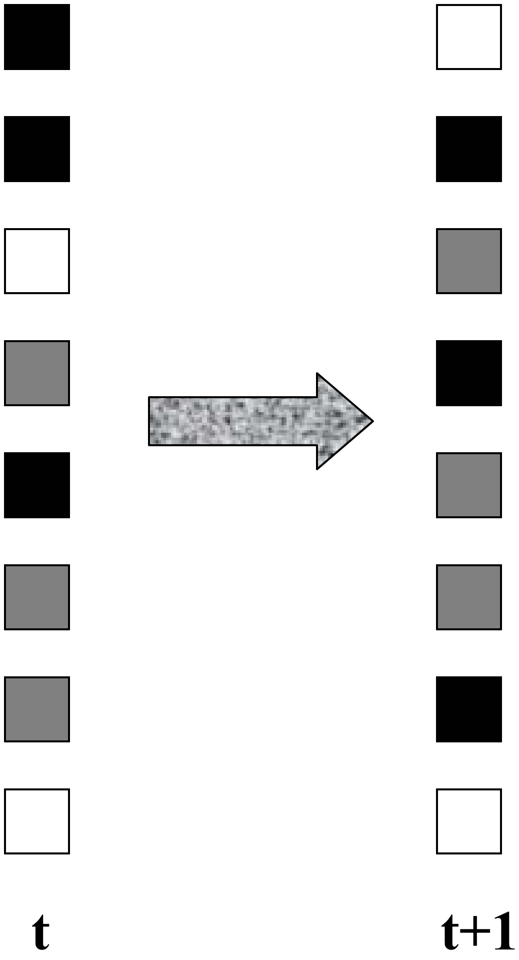
The more you know about the system

- **regulatory architecture/topology**
- **actual parameters etc.**

…..the easier it is

Even if you have a complete regulatory architecture you need to do some parameter fitting/testing

The Principle of Reverse Engineering of Genetic Regulatory Networks (Deterministic view):



Determine a set of regulatory rules that can produce the gene expression pattern at T_2 given the gene expression pattern at the previous time point T_1

$$x_i(t+1) = g\,(b_i + \Sigma w_{ij} x_j(t))$$

Continuous modeling: (variations on a theme)

$$x_i(t+1) = g\,(b_i + \Sigma_j w_{ij} x_j(t))$$

basic assumption of most continuous approaches

(Mjolsness et al, 1991 - connectionist model;
Weaver et al., 1999, - weight matrix model;
D'Haeseleer et al., 1999, - linear model;
Wahde & Hertz, 1999 - coarse-grained reverse engineering)

$$g(z) = \frac{1}{1 + e^{kz}}$$

The aim is to determine all the b_i and w_{ij} values.

- you need as many equations as variables

1. Genetic algorithms (Wahde & Hertz, 1999)
2. Solving weight matrices (singular value decomposition etc.)
(Weaver et al., 1999)
3. Least square fit for the linear modeling
(D'Haeseleer et al., 1999)

Correlation matrices:

(see Arkin, Shen & Ross, 1997)

If a chemical reaction takes 1 unit of time, then the B→A reaction will be a more likely candidate than the C→ A reaction to explain the time dependent changes in the figure above.

Correlation matrices:

(Arkin, Shen & Ross, 1997)

Time lagged correlation matrix can be prepared based on equations:

(1) $S_{ij}(\tau) = <[x_i(t)- x_i][x_j(t+\tau)- x_j]>$

(2) $$r_{ij}(\tau) = \frac{S_{ij}(\tau)}{\sqrt{S_{ii}(\tau)\, S_{jj}(\tau)}}$$

<…..> : time average over all the measurements

$x_i(t)$: t-th time point of the time series generated for species i

x_i : time average of the i-th time series.

How much does a change in the level of species i correlate with a change τ time later in the level of species j ?

How much information is needed for reverse engineering?

Boolean fully connected 2^N

Boolean, connectivity K $K \, 2^K \log(N)$

Boolean, connectivity K, linearly separable rules $K \log(N/K)$

Pairwise correlation $\log(N)$

N = number of genes
K = average regulatory input/gene

r unknown parameters in a set of ODEs $2r+1$
(Sontag, 2002)

P = K log(N/K) (John Hertz, Nordita)

P : gene expression states
N: size of network
K: average number of regulatory interactions

1. Stochasticity (??????)
2. Size of network $N_{bic} < 10$ x N_{gen}
   about 1.2-fold increase in P (but definitely less than 2)
3. Connectivity (compartmentalization)- it will make thing easier (it can reduce P)
4. Information content is 1-2 order of magnitude less: 10-100 fold increase in P.

The useful information content of a gene expression matrix will depend on:

1. Measurement error (conceptual and technical limitations, such as normalization)
2. Kinetics of gene expression level changes (lack of sharp switch on/off kinetics - stochasticity ?)
3. Number of genes changing their expression level.
4. The time frame of the experiment.

Applying all this to cell cycle dependent gene expression measurements by cDNA microarray one can obtain 1-2 orders of magnitude less information than expected in an ideal situation. (Szallasi, 1998)

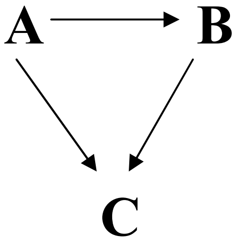
Reverse engineering using perturbations

Perturbations on time and population averaged measurements

Wagner, A. (2001)
Ideker, T. ….Hood, L. (2001)

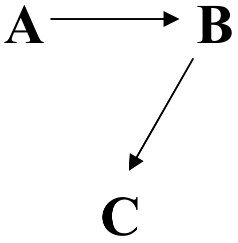
perturbation matrix

Knock-out		A	B	C
Gene	A	0	1	1
expression	B	0	0	1
	C	0	0	0



accessibility matrix

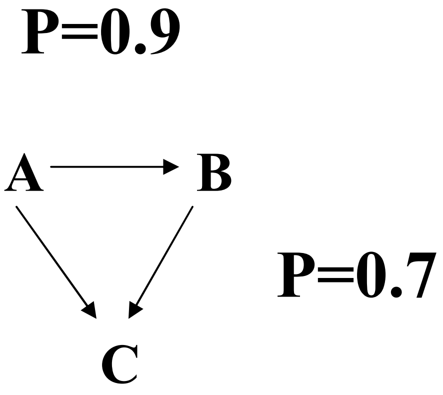
	Regulator	A	B	C
	A	0	0	0
Regulated	B	1	0	0
	C	1	1	0



Start with an already known topology if you can:

Ideker et al (2001) – update the knowledge

N. Friedman, Hartemink : Bayesian view of the network
- May work well on subnetworks,
- USE prior knowledge of topology !!!!



genetic network modeling
&
systems biology

The size of the network:

Small scale - a few genes (N=1-3)

Intermediate scale (N=10-100)

Ensemble approaches (N=1000-100000)

Principle of interactions between genes

- **stochastic**
- **continuous differential eq.**
- **step functions/Boolean networks**

Small-scale genetic networks:

Detailed computational and experimental analysis of a few genes

Becskei & Serrano, (2000) stability of feedback loops

Elowitz &Leibler (2000) synthetic oscillatory network

Gardner…. J. Collins (2000) - genetic toggle switch

Does a feedback loop stabilize gene expression levels ?

Becskei & Serrano

Intermediate-scale genetic networks:

Computational analysis of a 5 to 100 gene network

(protein networks)

Schoeberl et al, (2002) EGF receptor pathway

Smith et al. (2002) analysis of the Ran regulated nucleocytoplasmic transport

1) **Overall topology of the network**

2) **Kinetic and other parameters**

Virtual cell

Does the model produce time series results that fit the data ?

Is the model robust ?
How sensitive to the initial setting of parameters ?

Can the model produce useful and testable hypothesis ?

Further uses of studying robustness:
Eldar …. Barkai, (2002)

Comments – Suggestions:

1) **Organize the model in a flexible way: libraries, automatic equation generators**

Modeling of biological systems is not new !

Large-scale modeling:

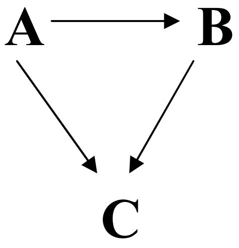
“the entire network” - 1000-100,000 genes

Boolean models: (Stu Kauffman, Leon Glass)

Principles of the "ensemble approach" of genetic network modeling:

1. Define a set of genes and their interactions - e.g. by a directed graph in which each gene is a node and each directed vertex denotes a regulatory interaction + define the function that describe the regulatory interaction (Boolean, continuous, stochastic)

Boolean: E.g. If A and C is on then B is on



$$[B]= g(w_1[A]+w_2[C])$$

$$g(z) = \frac{1}{1 + e^{-kz}}$$

2. Take an initial value set of genes and determine computationally how the system will behave when left alone? In a Boolean network for instance:

A	B	C	D	E	F	. . . .	
■	■	■	□	□	■		T i
□	■	■	□	□	□		T i + 1
■	□	■	■	■	□		T i + 2
□	□	■	□	□	■		T i + 3

A is ON IF B AND C is ON

B is OFF IF A OR D is ON

E is ON IF A OR D is ON and B is ON.
F is ON IF B OR D is ON and A is OFF

Under appropriate (!!!!) conditions the gene network will display organized behavior e.g. gene expression trajectories of reasonable length lead to an earlier state of the same trajectory, thus forming an attractor. SELF-ORGANIZING network. (???????????)

The network is not micromanipulated by e.g. feedback loops etc. but left alone to develop certain properties.

-Boolean models :
computationally (sort of) tractable (OR at least representatively sampled even for a network of 100,000 genes*)
<u>BUT the interpretation of results is not easy, gene regulatory</u> interactions are very "crudely" modeled
(5000 genes, Bhattacharjya & Liang,1996).

2. Building the modeling environment

continuous differential equations (Entelos, Physiome e-cell, Gene Network Sciences)

3. Robustness of forward modeling

fitting the kinetic constants to actual time-series measurements (searching for best fits, local or overall minimae, in a rugged landscape)

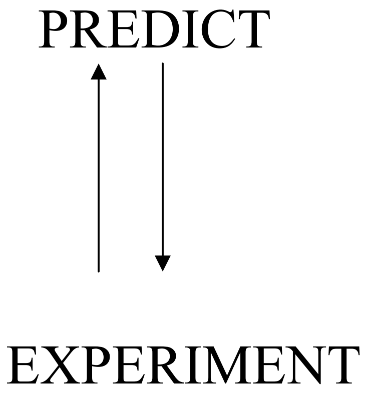
testing the robustness by perturbing the constants

Constraint-based models

Metabolic nets – flux balance analysis

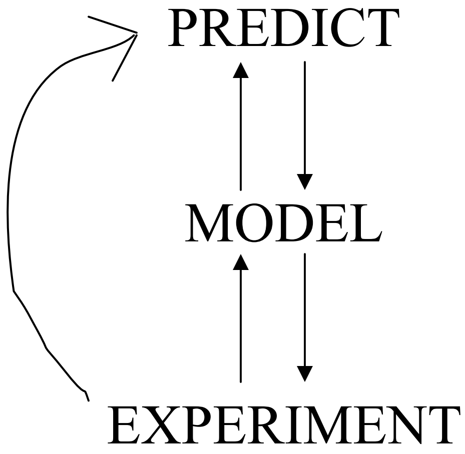
B. Palsson’s group (Nature Biotech, 19:125-130)

Kirchhhoff’s first law



PREDICT

EXPERIMENT



PREDICT

MODEL

EXPERIMENT

update your prediction, reduce number of experiments