Genetics Networks

Russ B. Altman
BMI 214
CS 274

(see special supplement to Nature Genetics, January, 1999, Vol 21, No. 1, available free to Stanford community)

What is a genetic network?

Individual genes have a function (e.g., transforming a substance or binding to a substance)

Sets of functions when sequenced can produce pathways (e.g., output of one transformation is the input to another)

Sets of pathways, as they interact with other pathways, create a genetic network of interactions.

The emergent properties of these networks constitute the “observables” when we study cells.

Genetic Networks, Genetic Regulatory Networks, Metabolic Networks

Genetic Network: used to denote general interactions of genes, gene products, and small molecules.

Genetic Regulatory Network: used to denote the network of control decisions used to turn genes on/off. A subset of the entire genetic network.

Metabolic Network: used to denote the network of proteins that synthesize and breakdown cellular molecules (enzymes, cofactors)

Interactions & Modulations

A Gene Regulatory Network

Connections Map
Pathway: PFK

Legend:
- gene
- gene regulation
- active regulatory gene
- stimulated gene
- inactive gene
- metal ion
- metabolite
Studying Gene Regulatory Networks

Until recently, difficult because

(1) Data about interactions and timing of expression was very difficult to collect.
(2) No good methods to analyze these networks

But now:

(1) Methods for measuring the “network” are available.
(2) Progress made in thinking about computing on these network.

Reconstructing Genetic Regulatory Networks

Hard problem.

Given N genes, there are an exponential number of connections between the genes.

Relationships are not generally +/- but are continuous valued (e.g. concentration of molecule varies smoothly).

Must use knowledge about expected function and membership in pathways to prune the list of possible network interactions.
Two problems for decoding regulatory networks

1. Getting sufficient data so that you can build and test models for a regulatory network.
   - Finding co-regulated genes from among 35,000+ genes
   - Finding the features of these genes in the sequence that allow them to be co-regulated (shared motifs?)

2. Choosing mathematical formalisms for describing the network

Availability of Relevant Data

Expression microarrays

1. Fix pieces of DNA of known sequence on a 2-dimensional array (e.g. all 6000 yeast genes)
2. Gather samples from cells under two conditions (e.g. starved vs. not starved)
3. Label the mRNA of the two samples with fluorescent dyes of different colors (green/red)
4. Mix the samples and pour on the 2D array. mRNA that is complementary (in Watson-Crick sense) will anneal to pieces of DNA.
5. Measure fluorescence to see if levels of mRNA in the one sample is same, more, less than the other.

How do microarrays work?

Cells of Interest

Known DNA sequences

Isolate mRNA

Glass slide

0.25 0.01 0.30 0.70
0.73 0.89 0.92 0.67
0.14 0.15 0.60 0.23
0.12 0.12 0.07 0.02
0.01 0.05 0.14 0.12

Typical DNA array for Yeast

What does this tell you?

mRNA is a proxy for how much protein is available in the cell.

DNA array gives a snapshot of the level of mRNA expression (and thus protein) in the cell at a particular time.

Can take multiple snapshots to watch the evolution of availability of mRNA over time, or in response to stimuli.

For example: in response to starvation, evolution of cancer, normal vs. disease, etc...

Can build trees from cluster analysis, groups genes by common patterns of expression.
What to do with array data?

Some obvious choices…

1. Cluster genes based on similar gene expression = a new metric for “similarity” separate from sequence measures.

2. Try to infer genetic interactions based on timing and co-regulation of genes.

Microarray data to study genetic networks

Control of Cell Cycle in Caulobacter

Average of clustered wave forms

Typical “wave forms” observed (note: not lots of bumps)
How can we infer genetic regulatory information?

Look at genes in the genome.

Look for signals around the gene that may be used to tell them to turn ON/OFF.

--look for shared sequence motifs near gene
--perfectly conserved?

Looking for signals two ways:
--looking for commonly occurring shared words (using techniques of motifs discussed previously)
--looking for multiple alignments that have high scores, but imperfect matches

Discovering co-regulation

Excellent recent review...

Directed graphs & hypergraphs

Bayes' Networks

Boolean Networks
Simplification: Boolean Network

1. All genes are either in on state or off state.
2. State of a gene at time T determined by the logical combination of states of regulatory genes at time T-1.
3. Can propagate the states over time to simulate the evolution of the network.

Very simplified, but still useful for seeing the kinds of properties that can emerge from these networks.

Sample network

\[ \begin{align*}
A' &= B \\
B' &= A \text{ or } C \\
C' &= (A \text{ and } B) \text{ or } (B \text{ and } C) \text{ or } (A \text{ and } C)
\end{align*} \]

equivalent to:

<table>
<thead>
<tr>
<th>INPUT</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>A'</th>
<th>B'</th>
<th>C'</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 0 0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0 0 1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0 1 0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0 1 1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1 0 0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1 0 1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1 1 0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1 1 1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Some sample state transitions

\[
\begin{align*}
1 0 0 -& 0 1 0 - 1 0 0 - 0 1 0 - 1 0 0 \\
1 0 1 -& 0 1 1 - 1 1 1 - 1 1 1 - 1 1 1 \\
0 0 0 -& 0 0 0 - 0 0 0 - 0 0 0 - 0 0 0 \\
0 0 1 -& 0 1 0 - 1 0 0 - 0 1 0 - 1 0 0 \\
0 1 0 -& 1 0 0 - 0 1 0 - 0 1 0 - 1 0 0 \\
0 1 1 -& 1 1 0 - 0 1 0 - 0 1 0 - 1 0 0 \\
0 1 0 -& 1 0 0 - 0 1 0 - 0 1 0 - 1 0 0 \\
0 1 1 -& 1 1 0 - 0 1 0 - 0 1 0 - 1 0 0 \\
1 0 0 -& 0 1 0 - 1 0 0 - 0 1 0 - 1 0 0 \\
1 0 1 -& 0 1 1 - 1 1 1 - 1 1 1 - 1 1 1 \\
0 1 0 -& 1 0 0 - 0 1 0 - 0 1 0 - 1 0 0 \\
0 1 1 -& 1 1 0 - 0 1 0 - 0 1 0 - 1 0 0 \\
1 1 0 -& 1 1 1 - 1 1 1 - 1 1 1 - 1 1 1 \\
1 1 1 -& 1 1 1 - 1 1 1 - 1 1 1 - 1 1 1 \\
\end{align*} \]

Finite State Automata

![Finite State Automata Diagram]

Things to notice

Attractors are can be static (only one node) or dynamic (sequence of nodes that repeat).

Different equilibriums are possible, even for very simple network. These can correspond to:

- disease state
- resting state
- perturbed state

Transitions from one attractor to another require external events (starvation, food supply, heat, etc...) to create mutation in state, and move to different attractor.

Correspondences

<table>
<thead>
<tr>
<th>Genetic Network</th>
<th>Boolean Network</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genotype/DNA</td>
<td>Wiring and rules</td>
</tr>
<tr>
<td>Gene</td>
<td>Element of state</td>
</tr>
<tr>
<td>Expression pattern</td>
<td>State</td>
</tr>
<tr>
<td>Development</td>
<td>Trajectory</td>
</tr>
<tr>
<td>Mature cell</td>
<td>Attractor</td>
</tr>
</tbody>
</table>
Generalized Logical Networks

\[ \bar{x}_i = 0, \text{if } x_i < \sigma_i^{(1)} \]
\[ \bar{x}_i = 1, \text{if } \sigma_i^{(1)} < x_i < \sigma_i^{(2)} \]
\[ \sigma_i^{(1)} < \sigma_i^{(2)} < \ldots < \sigma_i^{(p)} \]

\[ \bar{x}_i = p, \text{if } x_i > \sigma_i^{(p)} \]

Different sigmas denote different cutoff levels for important concentration effects. Thus, \( x_i \) has \( p \) different concentrations that are important for determining its effect on other genes...

Sigmas can be thought of as "critical concentration" levels where effects change.

Nonlinear Ordinary Differential Equations

\[ \frac{dx_i}{dt} = f_i(x), \quad 1 \leq i \leq n, \]

\[ \frac{dx_i}{dt} = f_i(x_1 t - \tau_1), \ldots, x_n (t - \tau_n), \quad 1 \leq i \leq n, \]

\[ \frac{dx_1}{dt} = k_1 r(x_n) - \gamma_1 x_1, \]

\[ \frac{dx_i}{dt} = k_{i, i-1} x_{i-1} - \gamma_i x_i, \quad 1 < i \leq n, \]

Sample \( r(x) = \text{Hill Function} \)

\[ h^+(x_j, \theta_j, m) = \frac{x_j^m}{x_j^m + \theta_j^m} \]

Piecewise Linear DE

\[ \bar{z}_1 = \kappa_1 r(x_1) - \gamma_1 z_1 \]
\[ \bar{z}_2 = \kappa_2 z_1 - \gamma_2 z_2 \]
\[ \bar{z}_2 = \kappa_3 z_1 \]

Copyright Russ B. Altman
Other methods…

1. Qualitative differential equations
   - non-numerical approximations
   - “as x1 increases, x2 decreases…a lot”
   - “as x2 decreases, x3 increases…a little”

2. Spatial models
   - distribute molecules in a grid of boxes representing different parts of cell
   - in each cycle of a simulation, reactions occur in cells, products diffuse to neighboring cells, and repeat
   - watch evolution of the simulation…

Stochastic Master Equations

\[ p(X(t + \Delta t)) = p(X, t) \left( 1 - \sum_{j=1}^{m} \alpha_j \Delta t \right) + \sum_{j=1}^{m} \beta_j \Delta t. \]

Probability distribution over gene X

Alpha = probability that reaction j will occur (consuming X)

Beta = probability that reaction j will create more X

This equation can be rearranged…

\[ \frac{\partial}{\partial t} p(X, t) = \sum_{j=1}^{m} (\beta_j - \alpha_j p(X, t)). \]


Phage (virus) lambda inserts its DNA into the host bacteria genome. It decides when to stay in the DNA and passively be transmitted vs. when to express its proteins & DNA and kill the cell.

This “decision” was simulated using a master equation approach.

Able to make quantitatively accurate predictions about aspects of the decision.

Rule-based Simulation of networks

1. Encode interactions as rules that allow transformations (generalization of Boolean networks…)
2. Start network in state 0
3. Apply rules in forward direction to see how system evolves
4. Limited by difficult-to-predict interaction between rules that may not be truly independent…

Old Premise (from Somogyi et al)

1. Gene for every function, function for every gene.
2. Complete reduction of organism
3. Determination of protein structures/activities
4. Mapping molecular gene product interactions
5. Assembly of database of molecular mechanisms
6. Synthesis is sum-of-parts computer model.

IF (AND (temperature-range of Exp-conditions is 0-to-45)
   (ionic-strength-range of Exp-conditions is 0.001-to-.3)
   (pH-range of Exp-conditions is 6.0-to-9.6))
THEN (activity of DNA-polymerase-I is DNA-binding)
New Premise

1. Gene function distributed across parallel network
2. Identify genes and genetic network elements
3. Determine states of network (e.g. expression patterns)
4. Map out alternative trajectories/attractors
5. Reverse engineer the network
   • parallel trajectories suggest shared input
   • temporal links determined by wave shapes

Efforts in cell simulation: E-Cell

http://e-cell.org/

Input:
Molecules, locations, concentrations, reaction rules

Output:
Concentration changes over time

Contribution: Simulation of minimal 127 gene “cell” at this level of detail.

Efforts in cell simulation: Virtual Cell

http://www.arcam.uchc.edu/

Input:
Molecules, compartments, concentrations, reaction rules, diffusion rates,

Output:
Concentration changes over time, spatial distribution

Contribution: Calcium waves in cells, dynamics of IP3.

Spatial distribution of Ca++
Conclusions

Increasing data availability relevant to reconstruction of genetic networks, regulatory networks.

Many representations possible, usually chosen based on nature of system and type of data.

Despite apparent plethora of data, there are many parameters to be estimated, and so this is not yet a “data rich” modeling effort (relatively speaking).

There have been some high profile successes using mathematical modeling to re-create experimental observations. That’s the ultimate test.