Steps toward Models of Gene Regulatory Networks

John Grefenstette

George Mason University

Bioinformatics Colloquium

Feb 8, 2005

Outline

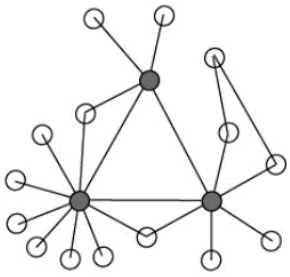
- Biological Network Models
- A biochemical model of gene regulation
- Simulation results from the model
  - regulatory network topology
  - regulatory rules
  - network dynamics
- Future directions

NIH Roadmap
http://nihroadmap.nih.gov

- Goals:
  - Earlier and more precise diagnosis, prevention and treatment of a wide variety of diseases
- Requirements:
  - Quantitative understanding of the many interconnected networks of molecules that comprise our cells and tissues, their interactions, and regulation
  - Models that can help predict the human body's response to disease, injury or infection

Biological Network Models
(cf. Alon, *Science*, 2003)

- Abstract representation of biological systems
- Molecules represented by nodes
- Interactions represented by edges
- May include:
  - protein-protein interactions
  - protein-DNA interactions
  - protein-metabolite interactions

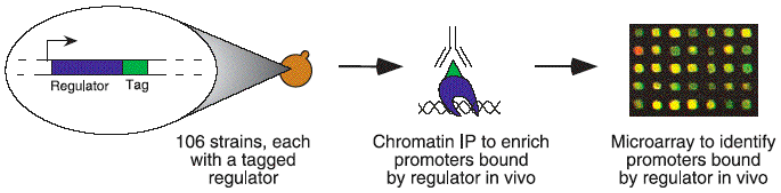


- Details suppressed
  - different mechanisms of transcription regulation represented by single type of edge
  - edges may not reflect strength of interactions

Toward a Quantitative Understanding of Networks

- Analysis of regulatory networks for specific organisms
  - in development
  - in normal cells
  - in disease states
  - in response to injury or environmental conditions
- Comparative analysis of regulatory networks
  - how do networks evolve?
  - how are networks related across species?
- Theory of regulatory networks
  - how might they have originated?
  - what regularities might be expected based on principles of complex systems and biochemistry

Regulatory Motifs in Yeast
(Lee et al, Science 298, 2002)



- Genome-wide binding analysis for 106 transcription regulators by Chromatin Immunoprecipitation (ChIP)

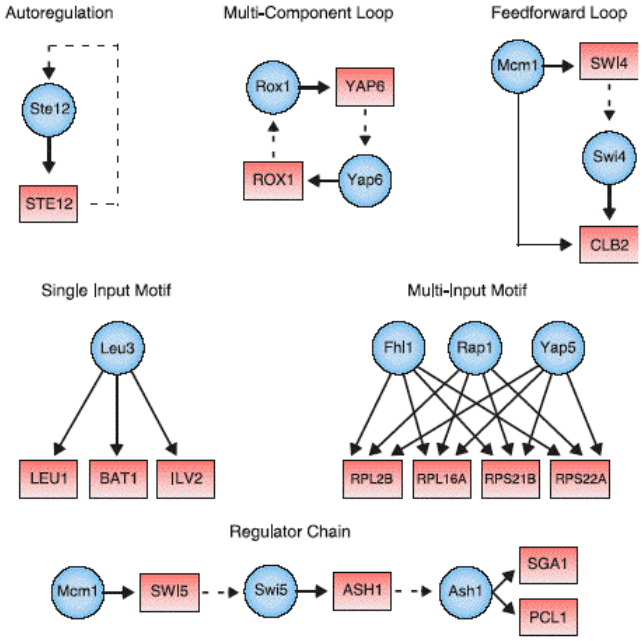
Regulatory Motifs in Yeast
(Lee et al, Science 298, 2002)

Results

- Observed about 4000 interactions between regulators and promoter sites (at P = 0.001)
- Identified common network structures (motifs)
- Model useful for suggesting further experiments

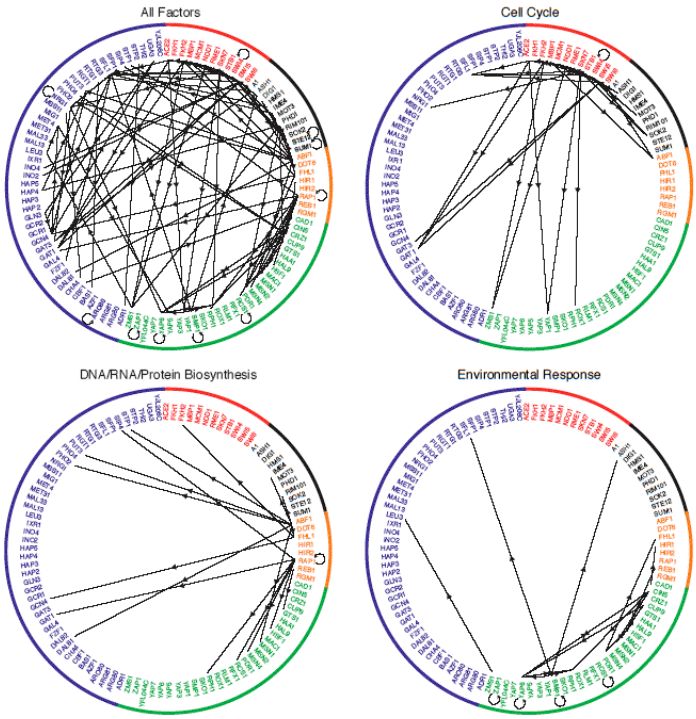
Open Issues:

- Origin of patterns?
- Statistical significance of patterns?



Regulatory Motifs in Yeast
(Lee et al, Science 298, 2002)

- Many regulators bind to genes that express other regulators
- Network substructures, e.g. cell cycle, metabolism, are coordinated at transcriptional level



Previous Work in Theory of Biological Network Models

- Interaction Types
  - Logical (Boolean) functions (Kauffman, 1969)
  - Continuous-time switching (Glass, 1973)
- Topology of interactions
  - Random graphs (Kaufman, 1969)
  - Scale-free (Barabasi, 1999)
  - Small-world (Jeong, 2000)
  - Modular (Alon, 2002)
- Dynamics
  - Ordered, complex, chaotic (Kauffman, 1993)
  - Oscillatory (Glass, 1979)

Specific Aims

- Previous work built models with specific topologies, interaction rules and network dynamics
- Our approach: construct a regulatory network model based on biochemical mechanisms and measure the resulting:
  - topologies
  - interaction rules
  - network dynamics
- Motivation:
  - Provide better understanding of how regulatory mechanisms results in system-level behavior
  - Provide more realistic "null models" to compare against experimental data

Boolean Regulatory Networks

- N Nodes (genes)
- Nodes have binary values: v = 0 or 1 (on or off)
- Each node i has k_i inputs (regulatory genes)
- Each node uses a deterministic Boolean (logical) function to update its value based on the values of its inputs

$$v_i = B_i (v_{i1}, v_{i2}, \ldots, v_{ik})$$

- State of system = current values of all nodes

$$S = (v_1, v_2, \ldots, v_N)$$

- Each node updates its value synchronously

B	C	A
0	0	0
0	1	1
1	0	0
1	1	0

A = C and (not B)

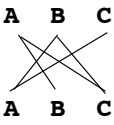
A	B
0	1
1	0

B = not A

A	B	C
0	0	0
0	1	1
1	0	0
1	1	0

C = A or B

wiring diagram



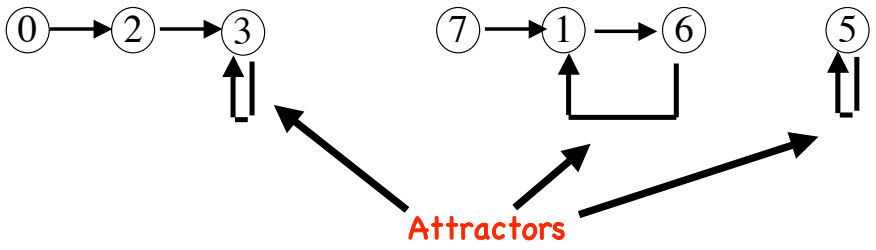
Trajectories:

	A	B	C
t=0	0	0	0
t=1	0	1	0
t=2	0	1	1
t=3	0	1	1

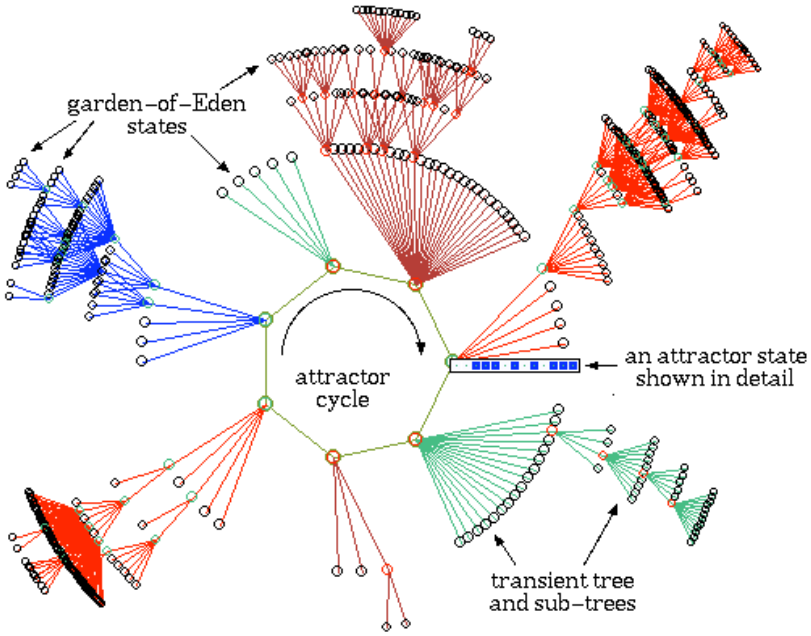
A	B	C
0	0	1
1	1	0
0	0	1
1	1	0

A	B	C
1	0	1
1	0	1

A	B	C
1	1	1
0	0	1



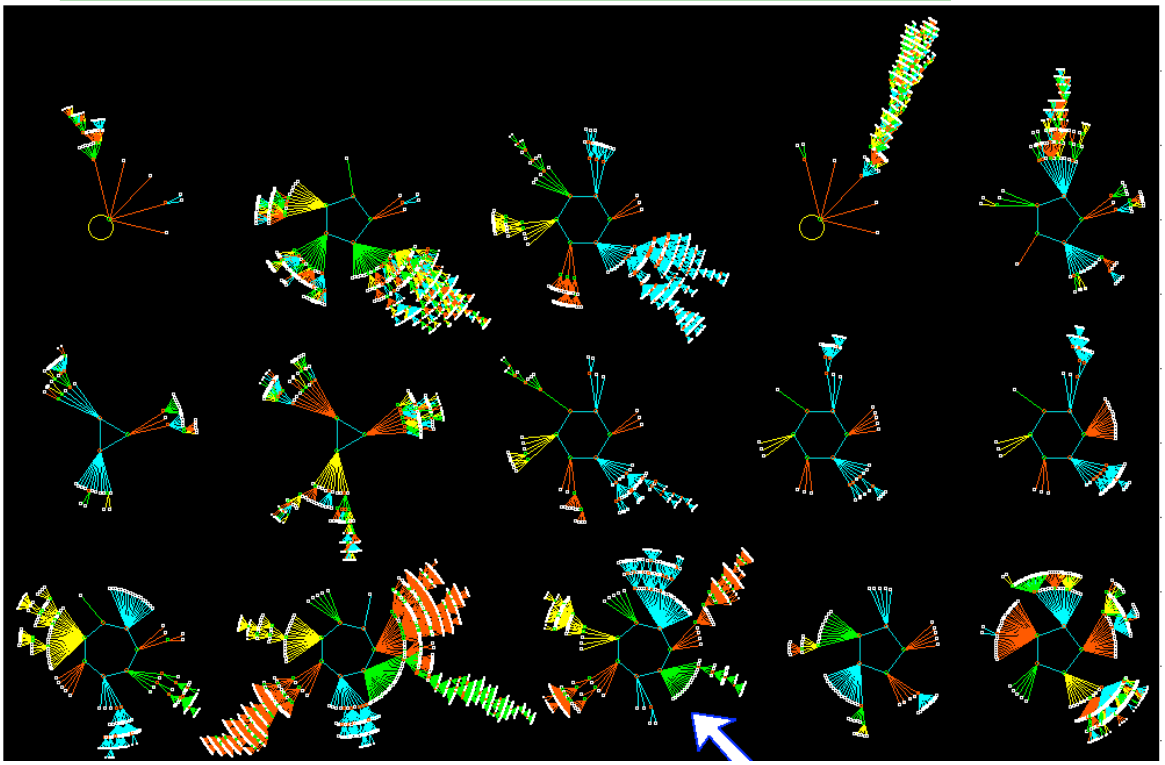
Basin of Attraction



Wuensche, SFI, 1998

Basins of attractions

Wuensche, SFI, 1998



Assumptions in the Regulatory Model

- Genes are associated with a cis site and a coding region
- Coding region may be expressed as proteins
- Proteins may form complexes
  - Monomers may form dimers, trimers, tetramers, etc
- Proteins may bind to cis sites
  - Competitive binding based on affinity between protein and cis-site
- Proteins may provide positive or negative control over transcription

Genes, Proteins, Regulatory Sites

The Model:

- Gene = (cis site, coding region)
- Coding region produces one monomeric protein
- Each protein has two templates (binary strings)
  - a protein-binding template
  - a dna-binding template
- Each cis-site has a protein-binding template
- Templates are used to form protein complexes and to binding proteins to cis site

Forming Protein Complexes

Each protein P_i has protein-binding template b_i

Dimerization rule:

If hamming(b_A , b_B) > *dimer_threshold*

then proteins A and B form dimer AB

Example:

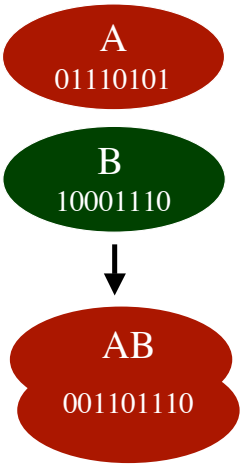
Suppose A has b_A = 01110101

B has b_B = 10001110

and *dimer_threshold* = 0.8

Then hamming(b_A , b_B) = 0.875

Therefore, A and B form dimer AB



Similar rules applies to creation of trimers, tetramers, etc

Binding to Regulatory Sites

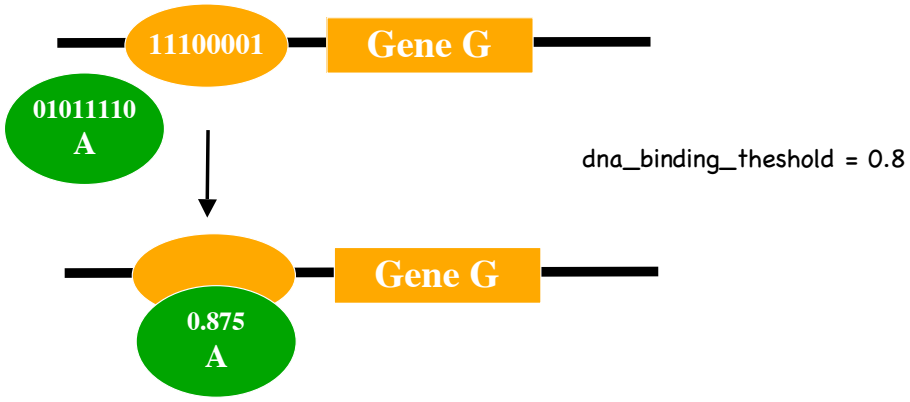
Each protein P_i has dna-binding template d_i

Each cis site C_j has protein-binding template b_j

Protein-DNA binding rule: If hamming(d_i , b_j) > *dna_binding_threshold*

then protein P_i may bind to cis site C_j

Protein-DNA binding affinity: $B(P_i , C_j)$ = hamming(d_i , b_j)



Transcription Control

Each cis site may act as a promoter (or not)

- Probability of being a promoter = p_{site}
- Probability of requiring a positive transcription factor = $1 - P_{site}$

Each protein may exert positive or negative transcription control:

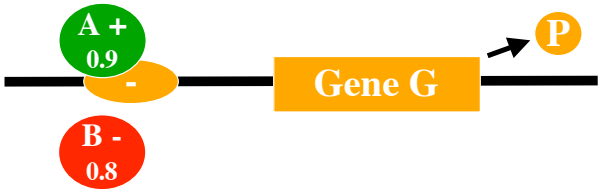
- Probability of being an activator = p_{prot}
- Probability of being a repressor = $1 - p_{prot}$

Gene expression rules:

1. If a gene's regulatory site is a promoter, the gene is expressed unless an repressor protein is bound to the regulatory site

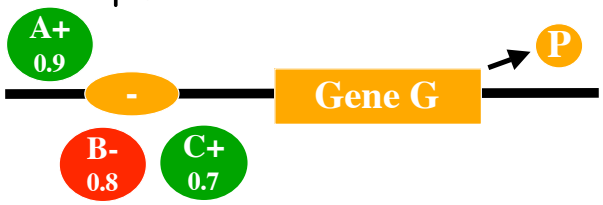
2. If a gene's regulatory site is not a promoter, the gene is expressed only if a activator protein is bound to the regulatory site

3. If more than one protein is available to bind to a given regulatory site, the protein with the highest affinity binds to the site



Generating Boolean Regulatory Functions

Example: suppose a cis site for gene G requires an activator and that three proteins bind:



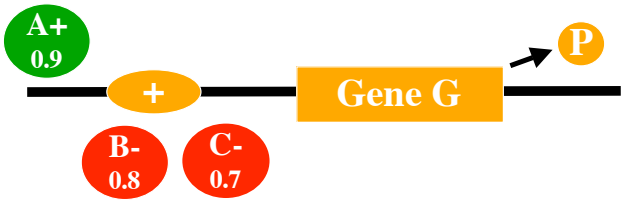
where (A+ 0.9) means that A is an activator protein (+) and binds to this cis site with affinity 0.9

Then the Boolean function for this gene would be:

G = A or (C and not B)

Generating Boolean Regulatory Functions

Example: suppose a cis site for gene G is a promoter and that three proteins bind:



The Boolean function for this gene would be:

G = A or (not B and not C)

Generating Boolean Regulatory Functions

1. Generate N genes and associated monomers
2. Generate all dimers, trimers, tetramers, etc
3. For each gene
   a. find all proteins that bind to its cis site
   b. sort the list by the binding affinity
   c. for each activator protein in the list, add to the Boolean function a disjunct that includes the activator and the negation of all higher affinity repressor proteins;
   d. If the cis site is a promoter, include a disjunct that includes all repressors that bind to the cis site

Methods

- Generate many networks of size N
  - varying model parameters
- Record regulatory interaction functions
- Analyze the resulting regulatory motifs
  - Cluster regulatory functions by number of inputs (k)
  - Identify regulatory functions that occur more often than expected by chance (regulatory motifs)
  - Characterize common classes of functions
    - random? canalyzing? other?
- Characterize network topology and dynamics as function of model parameters

Methods

- Model parameters:

  site_promoter_prob = 0.5
  protein_activator_prob = 0.5
  binding_template_length = 20 bits
  dna_binding_threshold = 0.80
  dimer_threshold = 0.80
  trimer_threshold = 0.85
  tetramer_threshold = 0.90

- Vary number of genes N = 250, 500, 750, 1000
- Generate 1000 networks of each size

Classes of Boolean Functions

- Random (Kauffman, 1969)
  - ordered, complex and chaotic dynamics
- Canalyzing (Kauffman, 1993)
  - A function is canalyzing if there is an input variable such that one of its values determines the output
    - e.g. G = A or (B and C) is canalyzing on input A
    - e.g. G = (A and not B) or (B and not A) is not canalyzing
  - appear to help prevent chaotic behavior
  - appear to be prominent in eukaryotic regulation
    - (Harris et al, 2002)
- Post functions (Shmulevich et al, 2003)

Results: Distributions of Boolean Functions

- Networks display strongly biased sets of Boolean functions (Boolean Motifs)

			N = 250		N = 500		N = 750		N = 1000	
K	All Truthtables	Distinct Truthtables	Observed	Samples	Observed	Samples	Observed	Samples	Observed	Samples
3	256	68	17	2050	36	16394	51	49418	57	90508
4	64k	3904	42	430	63	7013	110	32608	146	81487
5	4B	?	46	80	118	2187	180	13706	278	43107

Table 2. Results from 1000 simulation runs with N=250, 500 and 1000 genes. If Boolean functions were assigned as random, the number of observed functions should approximate the number of distinct functions (column 3). However, even in 1000 instances of simulations with N = 1000, an extremely biased set of of functions has been observed. For example, only146 distinct functions with K=4 have been observed in all simulations, compared with 3904 possible functions.

Boolean motifs with k = 4

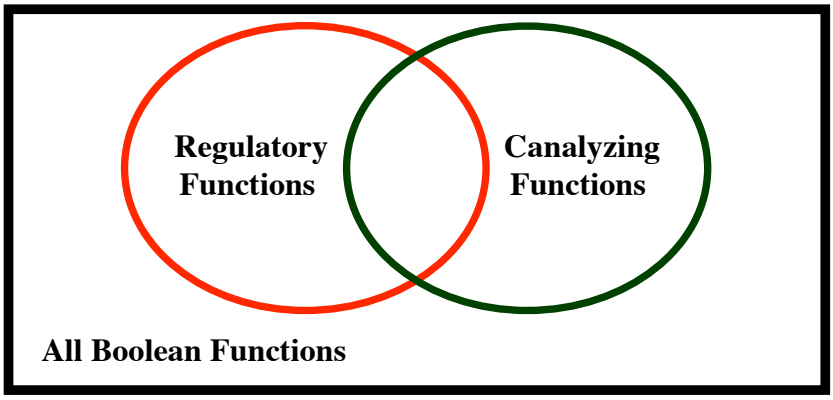
- Six functions each appear in more than 5% of regulatory interactions in which k = 4 (for N = 500, N = 1000)
  - Together, these 6 functions account for over 40% of 4-input regulatory interactions
  - Includes 2 non-canalyzing functions where gene is regulated by two dimers with same sign

- G = (A and B) or (C and D)
  - cis site for G is not a promoter

- G = not (A and B) and not (C and D)
  - cis site for G is a promoter

Observations

- Several non-canalyzing functions appear among the most prevalent regulatory motifs for k=4, 5, 6
- Many canalyzing functions never occur in simulations
- Suggests that canalyzing functions, while occurring much more often than expected, may not be best characterization of Boolean motifs



An alternative class of functions

Analysis of the observed Boolean motifs suggests the following definition:

A Boolean function B is in the class of Activator-Repressor (AR) functions iff each variable that appears in B's disjunctive normal form appears as either a positive term or a negative term, but not both.

Examples of AR vs Canalyzing:

AR but not canalyzing: G = (A and B) or (C and D)

Canalyzing but not AR: G = A or (B and not C) or (C and not B)

neither AR nor canalyzing: G = (A and not B) or (B and not A)

Non-AR functions seem to be rare

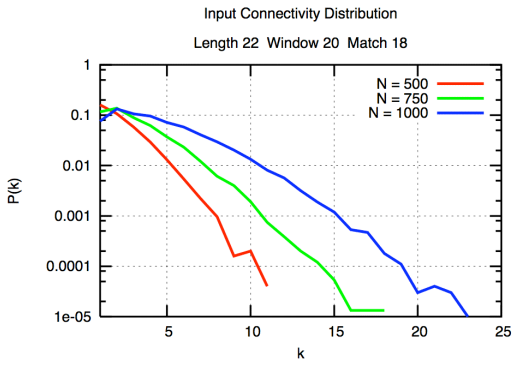
- Most Boolean function observed in all simulated networks are AR
- Some non-AR functions are observed (freq < 0.0005)

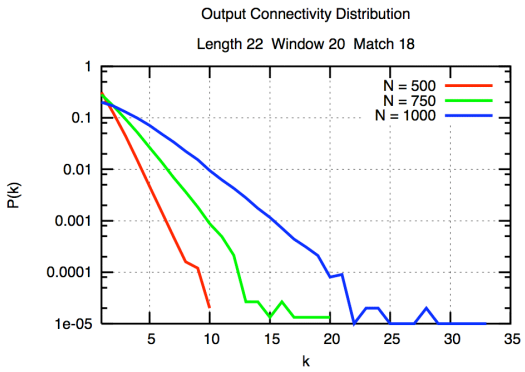
Logical Formula:	Realization:
(A and B) and not (A and C)	(AC)- (AB)+ [-]
(A and B) or not (A and C)	(AB)+ (AC)- [+]

- Boolean functions derived for 86 genes in TransCOMPEL database:
  - 77 of 86 (90%) are Canalyzing
  - 82 of 86 (95%) are AR
  - only one case identified in which a regulatory protein appears in both an activator and a repressor for the same gene
  - 3 of 4 non-AR functions based on concentration effects

Network Topology

- Biochemistry determines connectivity distributions, e.g.:
- N increases => higher connectivity
- Template length = 22

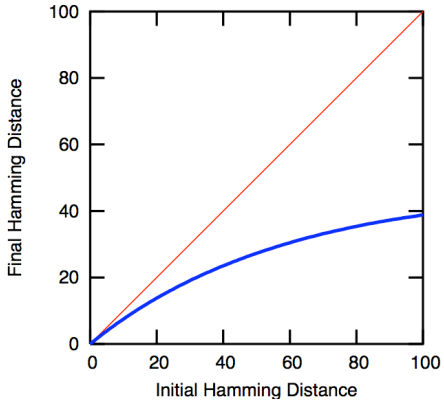




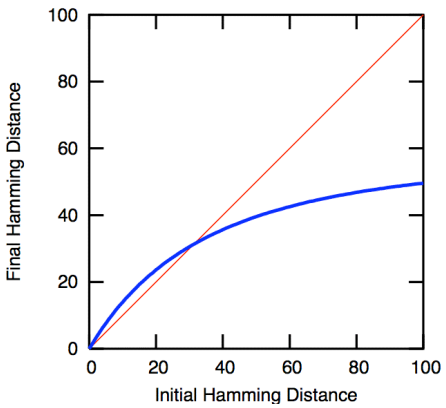
Network Dynamics

- Biochemistry determines dynamics, e.g.:
- Longer templates => higher connectivity => earlier transition to chaotic regime

Template length 22



Template length 24



Summary

- A model of regulation has been developed that includes protein-protein and protein-dna interactions
- Model provides more realistic null-model that previous models:
  - Assumption of uniform random interaction rules is not plausible
  - While canalyzing functions appear more often than expected, the model indicates other classes may be relevant
- Biochemistry parameters affects topology and dynamics
  - tune model to reflect biological system
  - explore selective pressures

Future Directions

- Analyze effects of model parameters
  - Topologies of AR nets
  - Dynamics of AR nets
- Inference complexity
- Evolutionary models
- Continue to validate model via experimentally derived transcription regulation databases

Acknowledgements

- Stuart Kauffman, U New Mexico
- Sohyoung Kim, Ph.D.
  - School of Computational Sciences, GMU

Selected Bibliography

Barabasi, A. L. and E. Bonabeau (2003). "Scale-free networks." Sci Am 288(5): 60-9.

D'Haeseleer, P., S. Liang, et al. (2000). "Genetic network inference: from co-expression clustering to reverse engineering." Bioinformatics 16(8): 707-26.

Fox, J. J. and C. C. Hill (2001). "From topology to dynamics in biochemical networks." Chaos 11(4): 809-815.

Grefenstette J, Kim S, Kauffman S (2005) . "An analysis of the class of gene regulatory functions implied by a biochemical model." (submitted to BioSystems).

Harris, S. E., B. K. Sawhill, et al. (2002). "A model of transcriptional regulatory networks based on biases in the observed regulation rules." Complexity 7(4): 23-40.

Kauffman, S., C. Peterson, et al. (2003). "Random Boolean network models and the yeast transcriptional network." Proc Natl Acad Sci U S A 100(25): 14796-9.

Kauffman, S. A. (1993). The origins of order: self-organization and selection in evolution. New York, Oxford University Press.

Lee T.I et al (2002). "Transcriptional regulatory networks in Saccharomyces cerevisiae." Science 298: 799-804.

Shmulevich, I., H. Lahdesmaki, et al. (2003). "The role of certain Post classes in Boolean network models of genetic networks." Proc Natl Acad Sci U S A 100(19): 10734-9.