Introduction to Microarray Data Analysis and Gene Networks lecture 8

Alvis Brazma

European Bioinformatics Institute

Lecture 8

- Gene networks – part 2
  - Network topology (part 2)
 - Network logics
  - Network dynamics

Gene Networks - four levels of hierarchical description

- **Parts list** – genes, transcription factors, promoters, binding sites, …

- **Topology** – a graph describing the connections between the parts

- **Control logics** – how combinations of regulatory signals interact (e.g., promoter logics)

- **Dynamics** – how does it all work in real time

Gene Networks - four levels of hierarchical description

- **Parts list** – genes, transcription factors, promoters, binding sites, …

- **Topology** – a graph describing the connections between the parts

- **Control logics** – how combinations of regulatory signals interact (e.g., promoter logics)

- **Dynamics** – how does it all work in real time

The arcs can have different meaning

G1 → G2
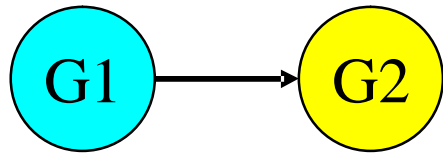
- The product of gene G1 is a transcription factor, which binds to the promoter of gene G2 (in Chip-chip experiment) – physical interaction network (direct network)

G1 → G2
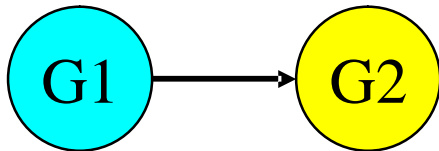
- The disruption of gene G1 changes the expression level of gene G2 – data interpretation network (indirect network)

How both networks compare

- How much the two networks have in common
- We can look at the intersection of the networks whether the common parts have evidence in our existing knowledge
- If the target sets of the transcription factors present in both networks are similar
- Are the network topology (e.g., connectivity) properties similar

How both networks compare

- How much the two networks have in common
- We can look at the intersection of the networks whether the common parts have evidence in our existing knowledge
- If the target sets of the transcription factors present in both networks are similar
- Are the network topology (e.g., connectivity) properties similar

A couple of simple notions

- Any gene (node in the graph) with outgoing edges is called a *source gene*
- Any gene with incoming edges is a *target gene*
- *Target set*

target node

source node

target set
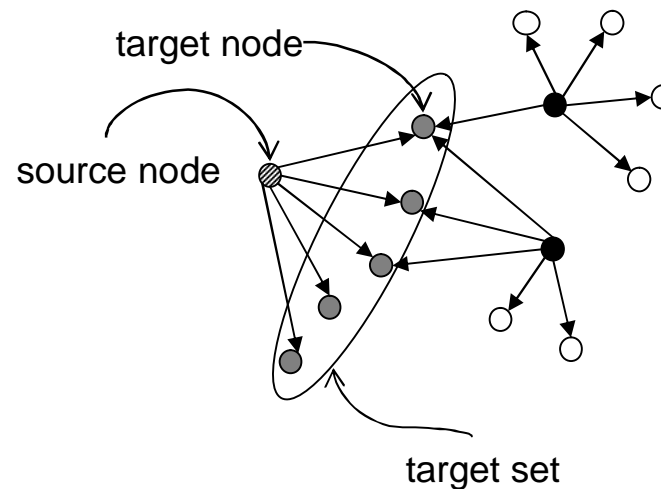
A problem:

- Both network depend on the chosen significance threshold - i.e., what level of microarray signal to use to draw and edge in the network

# The size of the networks for different significance thresholds

	ChIP network (p<0.01)	ChIP network (p<0.001)	mutant network (γ=2.0)	mutant network (γ=2.5)	mutant network (γ=3.0)
source genes	202	169	250	236	226
target genes	4939	2845	5396	4778	3920
genes	4980	2930	5654	4798	3959
edges	18842	6170	32017	17436	10356
edges where source gene and target gene have the same cellular role annotation in YPD (http://www.proteome.com)	3694 (19.6%)	857 (13.9%)	4096 (12.8%)	2425 (13.9%)	1507 (14.6%)
edges per source gene	93.3	36.5	135.7	73.8	45.6
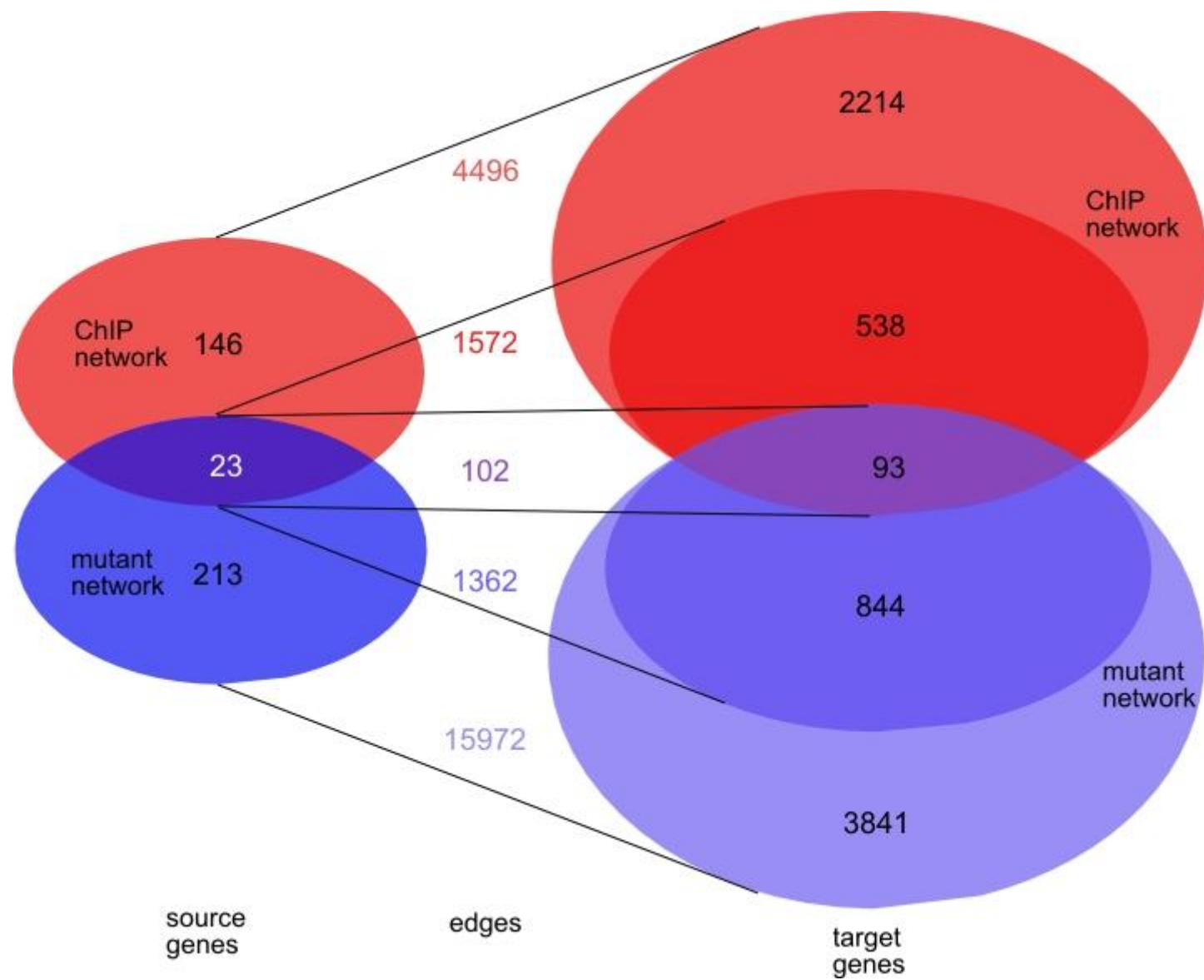
How both networks compare

- How much networks have in common
- We can look at the intersection of the networks whether the common parts have evidence in our existing knowledge
- If the target sets of the transcription factors present in both networks are similar
- Are the network topology (e.g., connectivity) properties similar

Intersection of the networks – many connections are consistent with out a priori knowledge
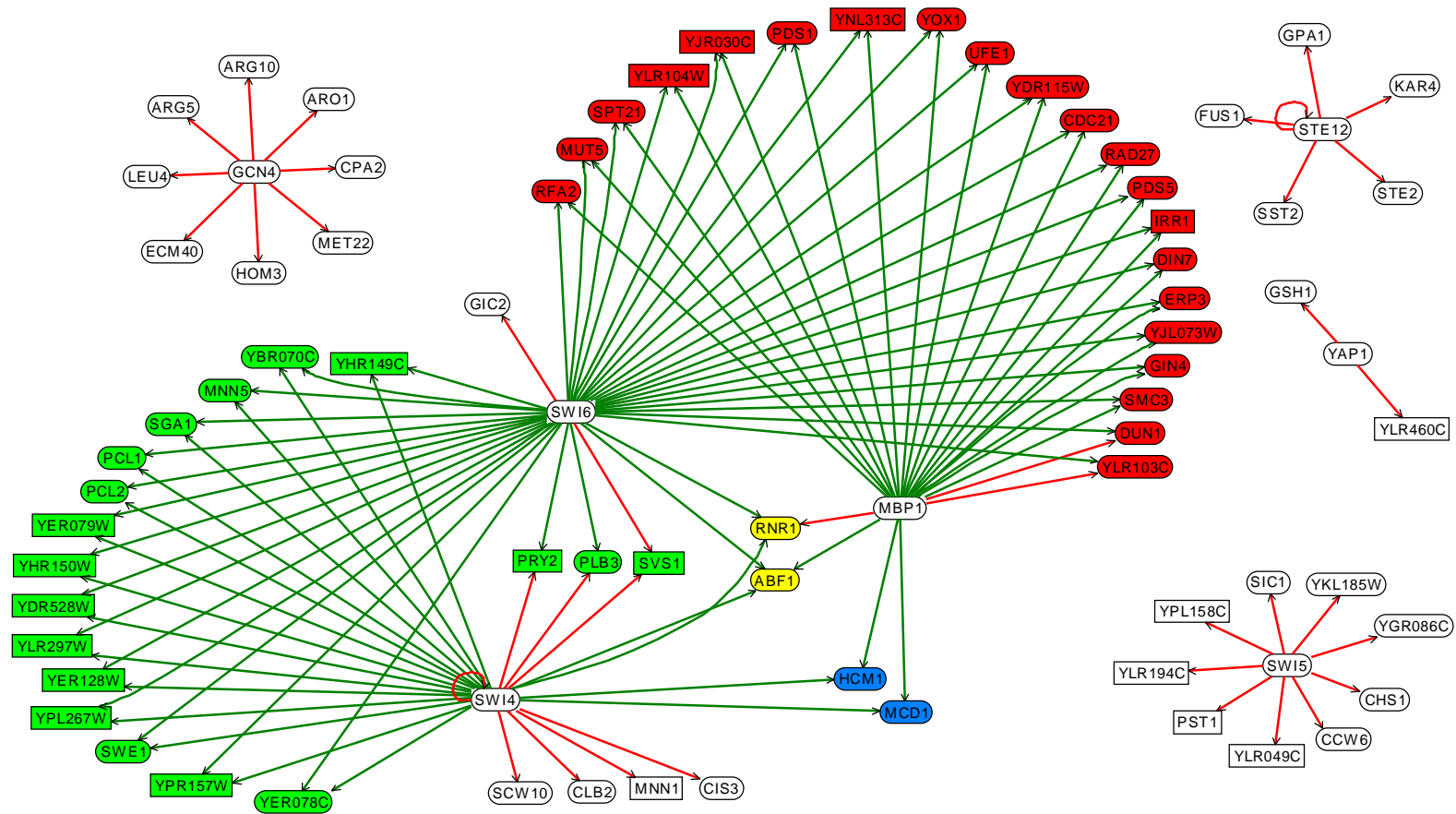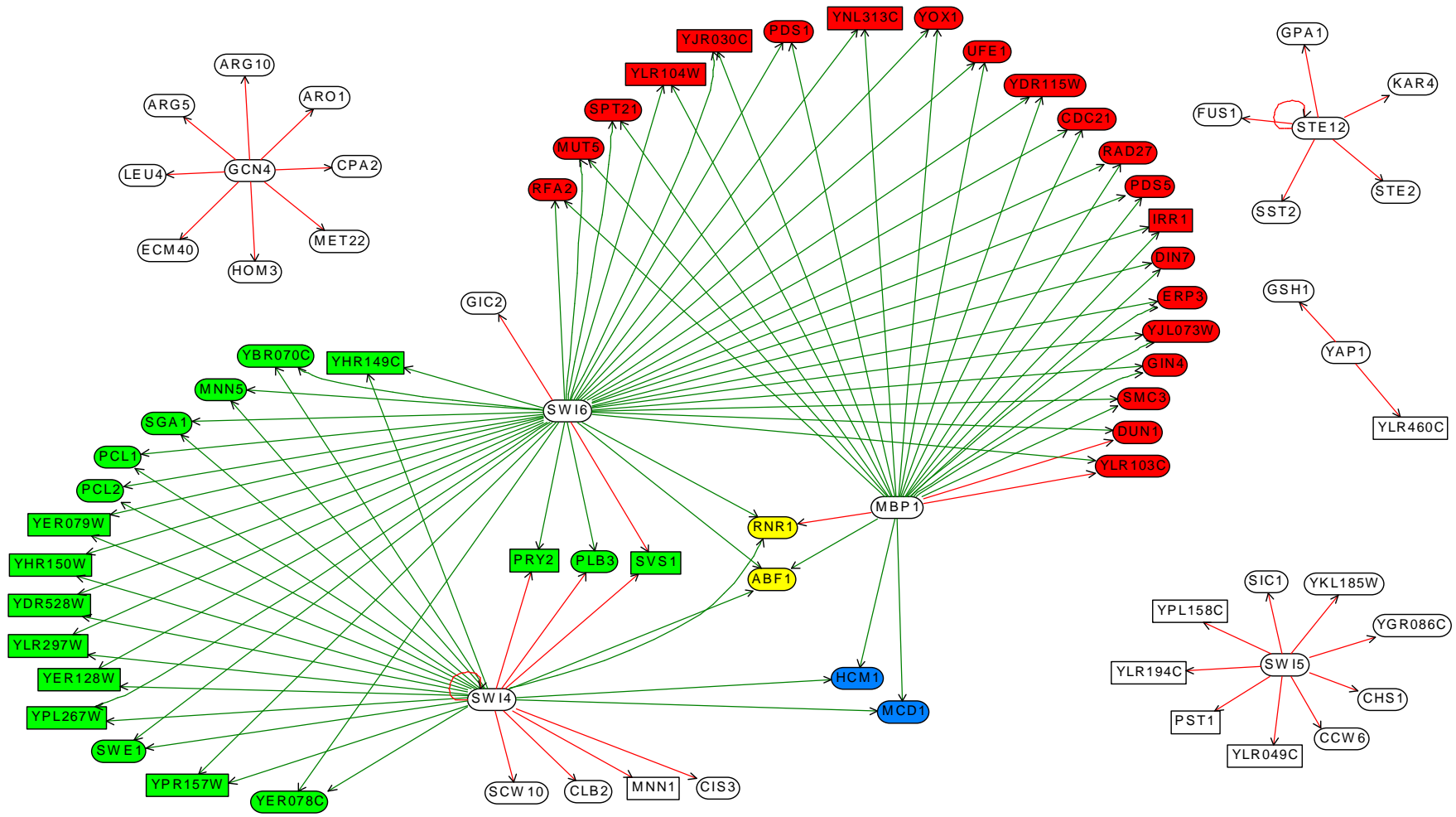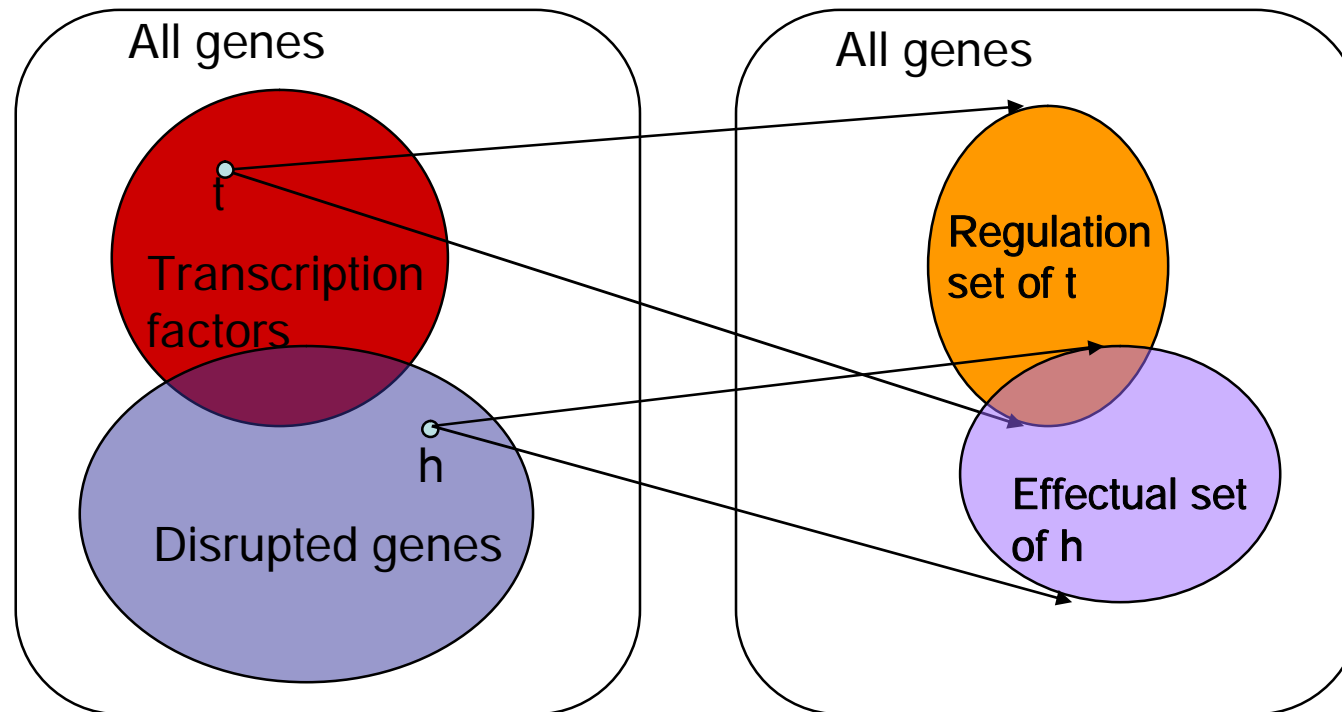
Figure 6

How both networks compare

- How much networks have in common
- We can look at the intersection of the networks whether the common parts have evidence in our existing knowledge
- **If the target sets of the transcription factors present in both networks are similar**
- Are the network topology (e.g., connectivity) properties similar
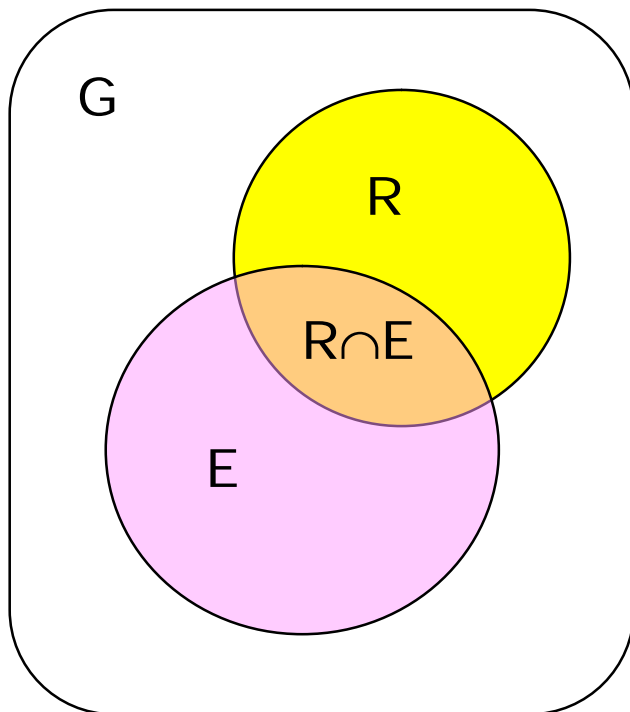
# How Chip-chip and disruption networks relate?
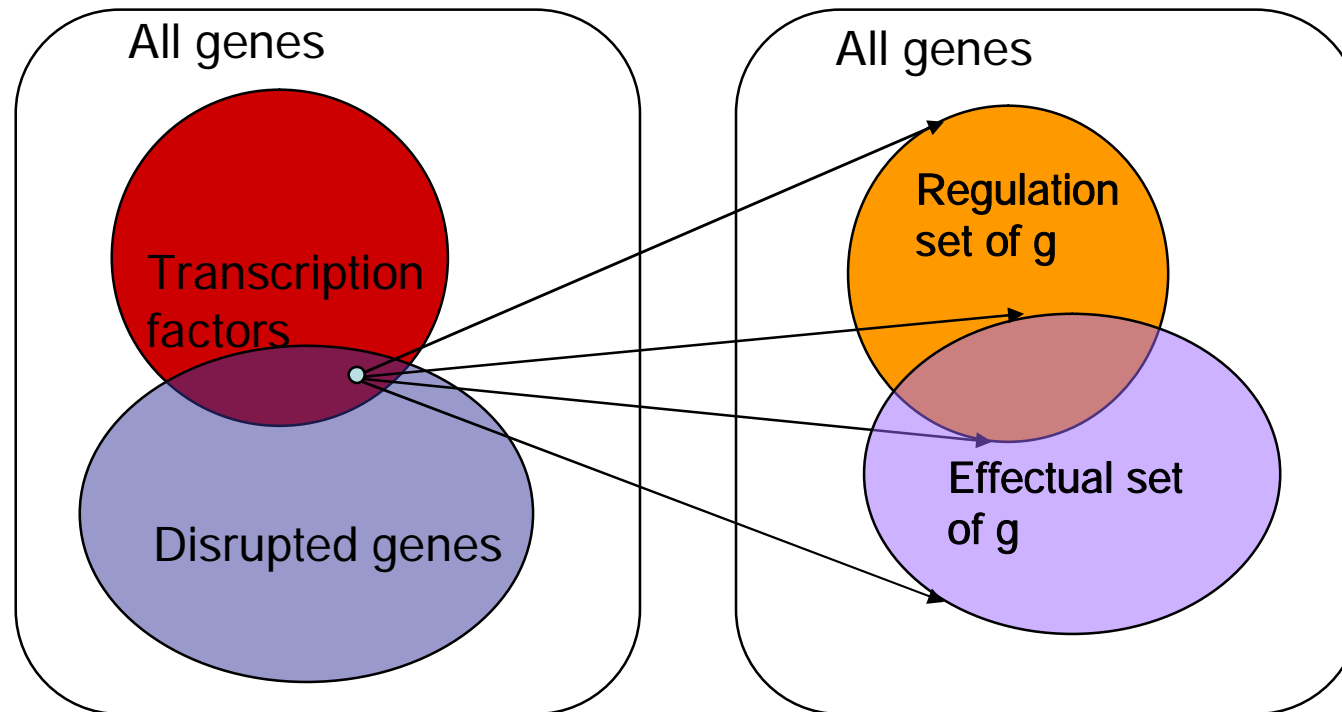
How to estimate that the overlap is more than expected by random?

G

R

R∩E

E

We assume that the elements of the set E are marked, and pick the set of size |R| at random. Then the size x=|R∩E| of the intersection are distributed according to *hypergeometric* distribution.
The probability of observing an intersection of size k or larger can be computed according to formula:

$$P(x \geq k) = 1 - \sum_{i=0}^{k} \frac{\binom{|E|}{i}\binom{|G|-|E|}{|R|-i}}{\binom{|G|}{|R|}}$$
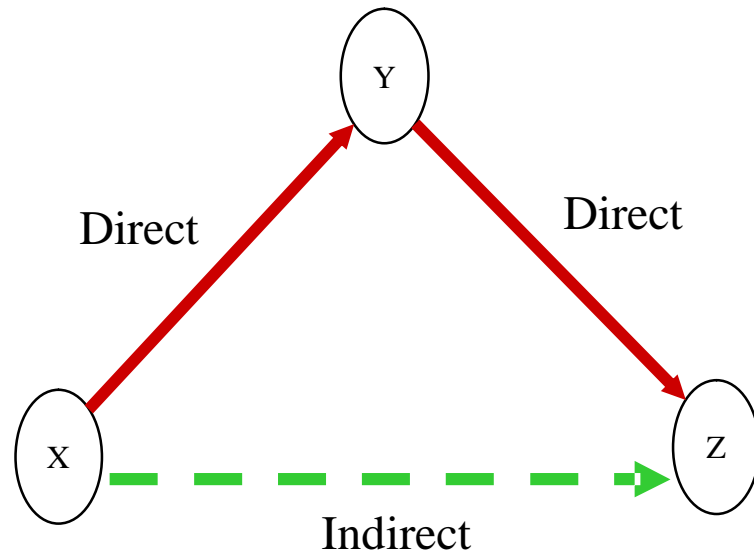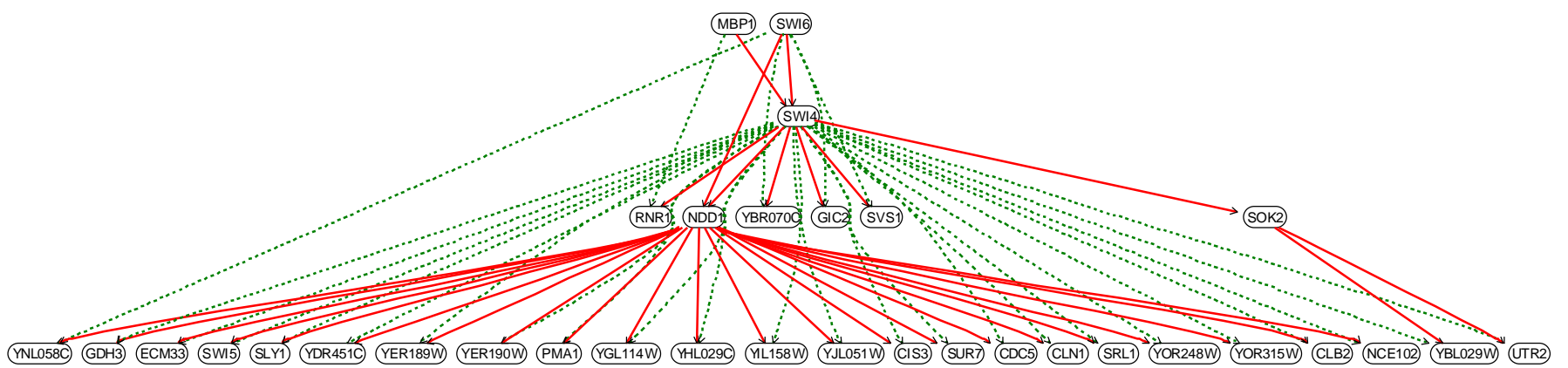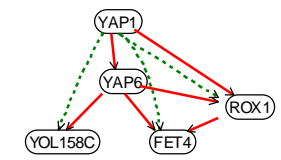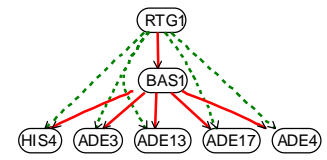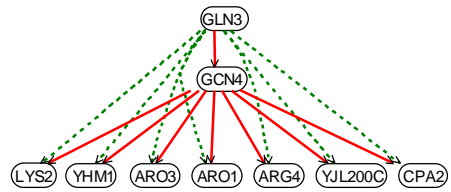
How Chip-chip and disruption networks relate?
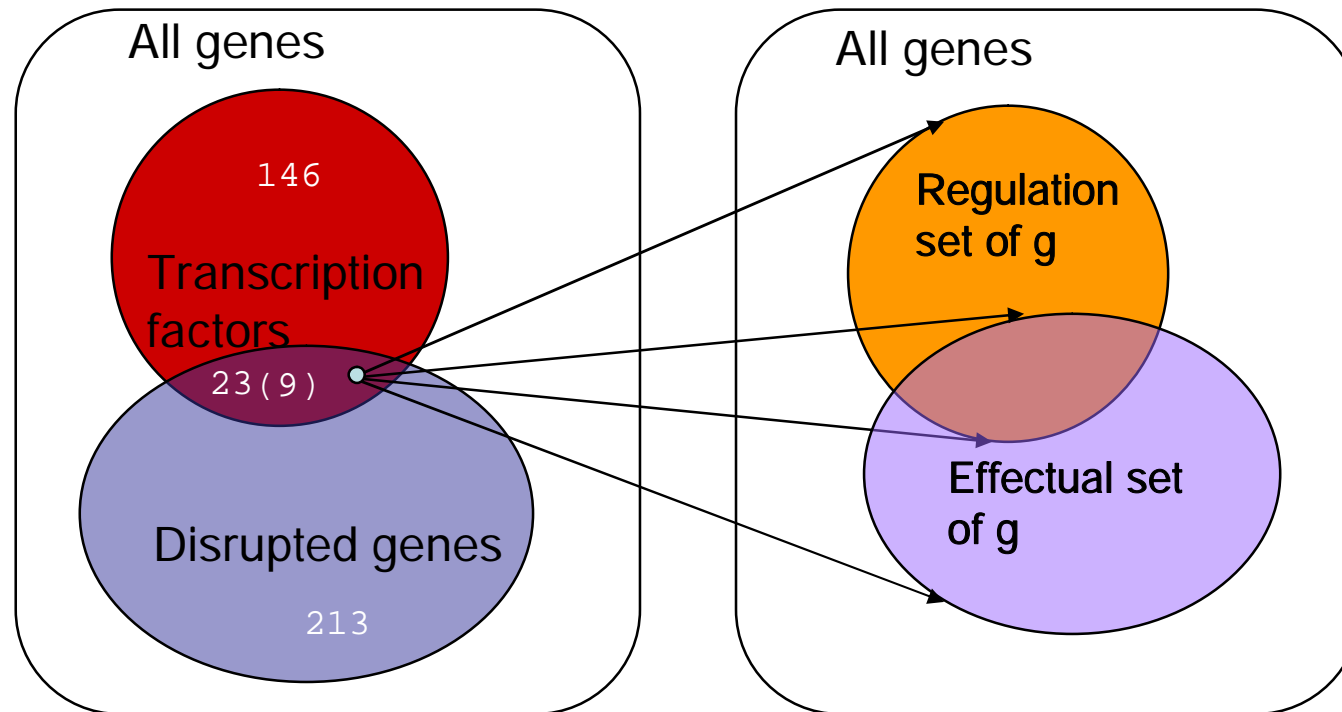
From 23 transcription factors studied in both networks only 9 have their target sets overlapping more than expected by chance L

From 23 transcription factors studied in both networks only 9 have their target sets overlapping more than expected by chance

- Is it as bad as my look?
  - We will expect many indirect connections in the disruption network that are not present in Chip network – is this the case?

Direct vs. indirect interactions

From 23 transcription factors studied in both networks only 9 have their target sets overlapping more than expected by chance

- Is it as bad as my look?
  - We will expect many indirect connections in the disruption network that are not present in Chip network – is this the case? There is an anecdotal evidence that this is the case
  - What about the connections present in the Chip network, but not in the disruption network? – can be explained by nonfunctional relationships in the chip network and combinatorial regulatory effects

Conclusions

- We want to think that networks share enough in common both to be meaningful, but at the same time apparently there is a lots of noise in at least one of them present

How both networks compare

- How much networks have in common
- We can look at the intersection of the networks whether the common parts have evidence in our existing knowledge
- If the target sets of the transcription factors present in both networks are similar
- **Are the network topology (e.g., connectivity) properties similar – and what are they**
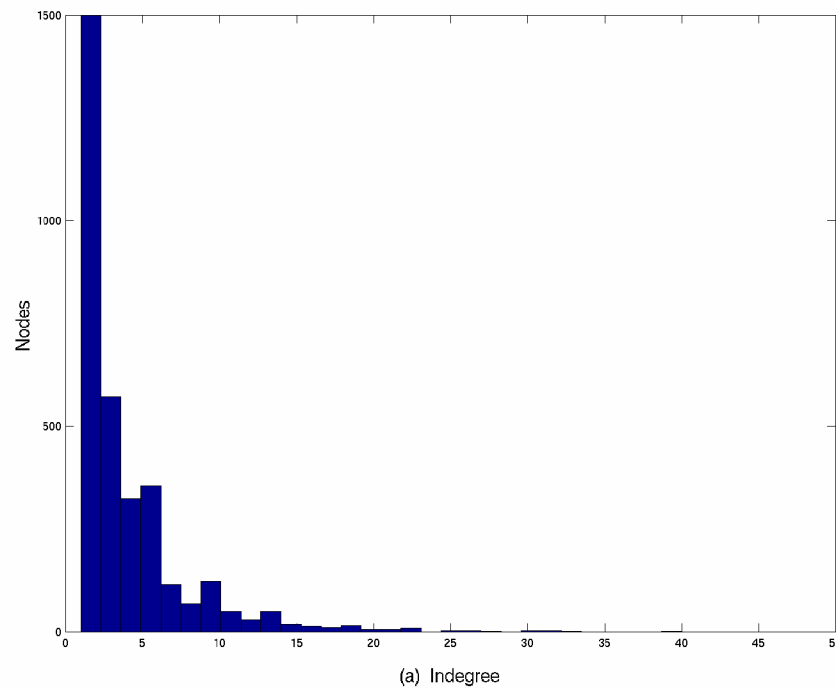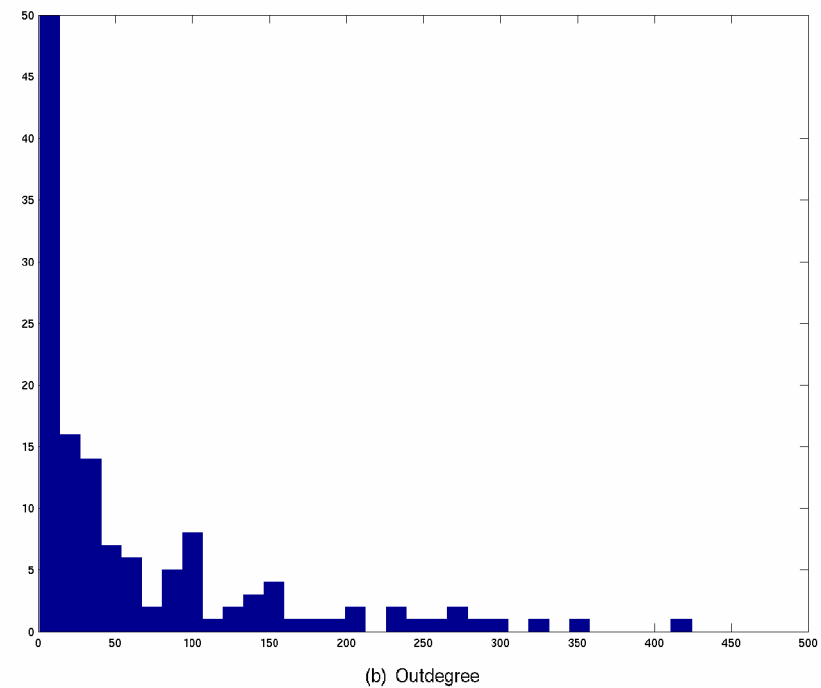
Degree of a node in a graph
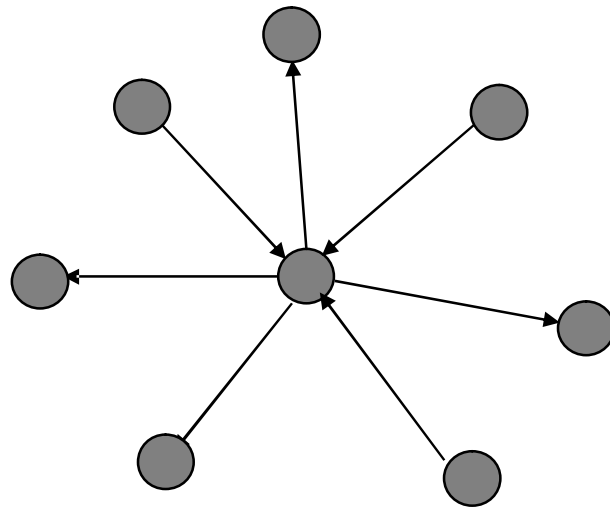
The central node has
degree = 7
indegree = 3
outdegree = 4

Important genes and genes with complex regulation

Most genes have only a few incoming / outgoing edges, but some have high numbers (>500)

Indegree

Outdegree

Genes with highest in- and out-degree

γ	outdegree	m	n	indegree	m	n
2.0	Carbohydrate metabolism	363	4	Amino-acid metabolism	9	194
	RNA turnover	353	4	Nucleotide metabolism	6	82
	Meiosis	244	3	Energy generation	5	242
	Cellstress	207	9	Small molecule transport	5	343
	Protein translocation	197	3	Other metabolism	5	148
2.8	RNA turnover	110	4	Amino-acid metabolism	4	167
	Cellstress	62	8	Nucleotide metabolism	3	67
	Meiosis	54	3	Energy generation	2	184
	Proteinsynthesis	53	7	Differentiation	2	43
	Cellwallmaintenance	47	6	Small molecule transport	2	286
3.6	RNA turnover	48	4	Small molecule transport	2	230
	RNA processing/ modification	41	4	Other metabolism	2	96
	Cellstress	27	8	Nucleotide metabolism	2	58
	Small molecule transport	19	8	Matingresponse	2	57
	Cellwallmaintenance	19	6	Amino-acid metabolism	2	133

Cellular role table showing the top 5 groups with the highest median degrees for the networks with γ=2.0, 2.8 and 3.6 with a minimum group size of 3 for outdegree and 40 for the indegree (m median degree, n number of genes per group)

Node degree distributions for both networks – roughly follow power-law
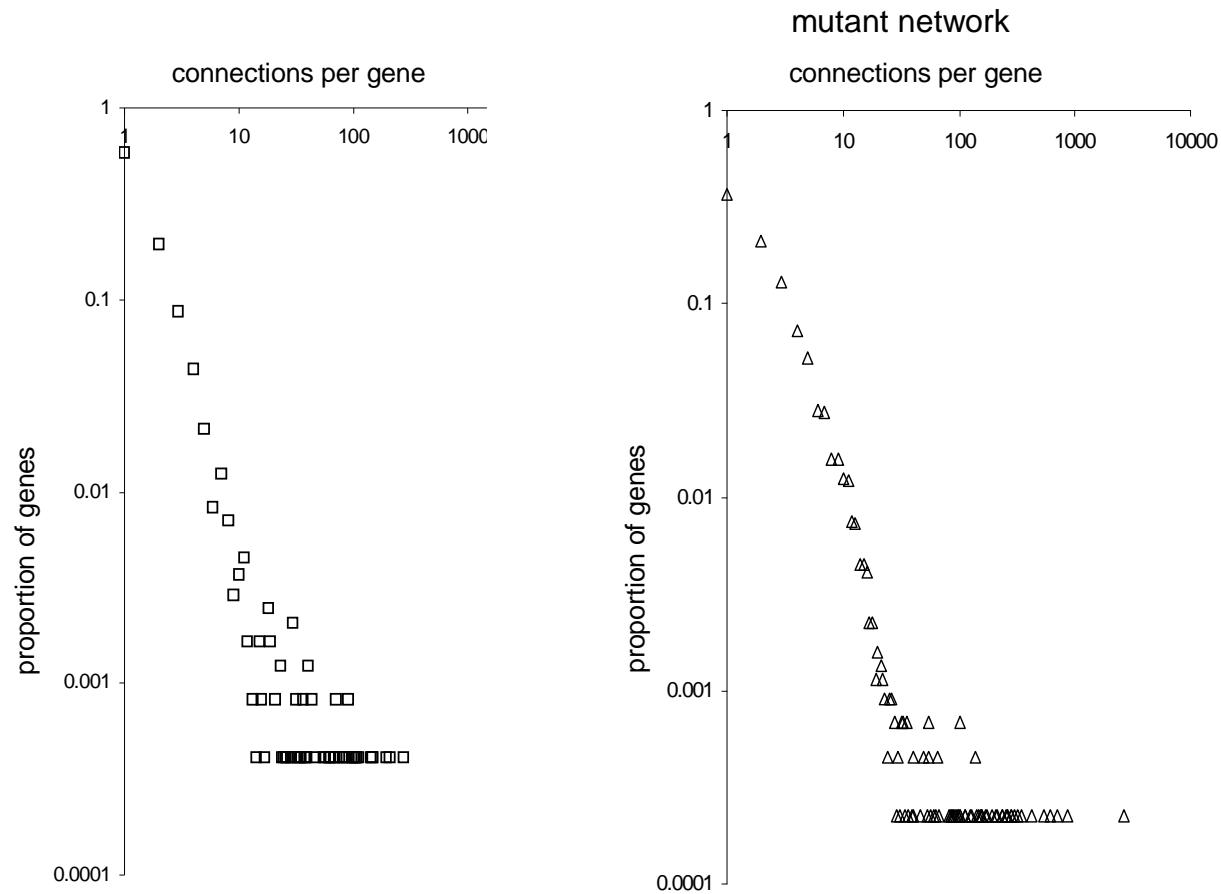
Yeast network topology properties
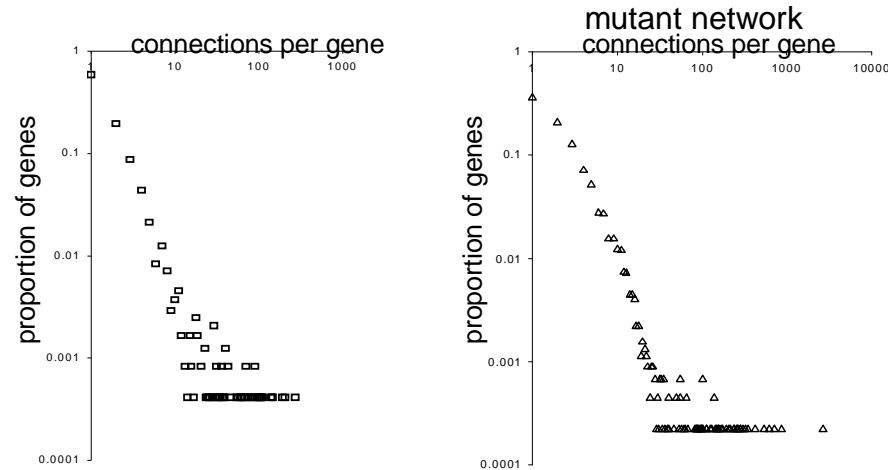
- Power-law property – on logarithmic scale approximately linear relationship
- Whether this is so is still open to debate – what is clear however is that most genes have relatively few connections, a few have many

Why topology is important?

- Reduce hypothesis space when analysing next layers of model complexity – instead of default – all genes depend on all, topology tells us which genes are independent
- What is the complexity of gene regulation
  - Given a transcription factor T – how many genes does T regulate?
  - Given a gene A, how many transcription factors regulate A?
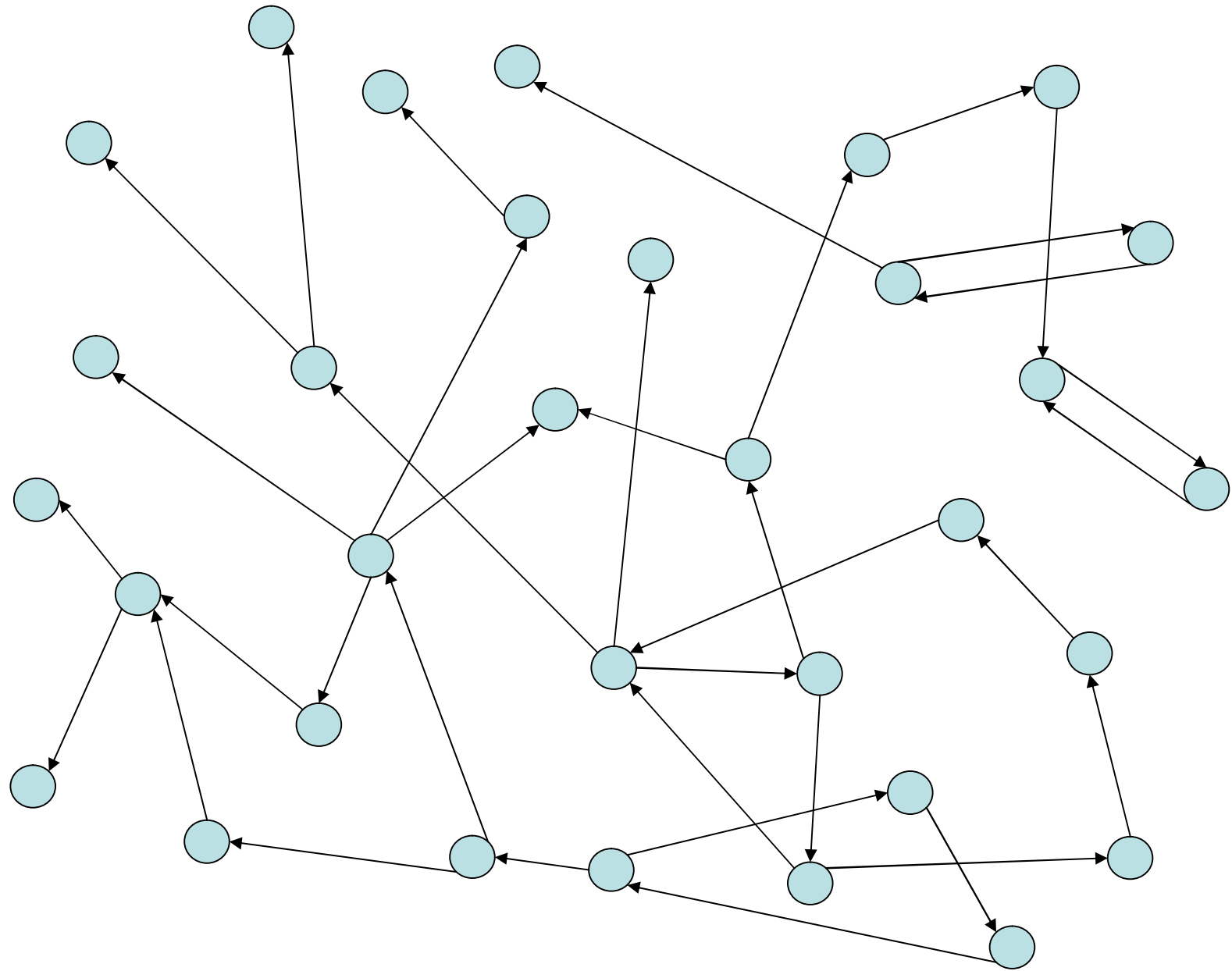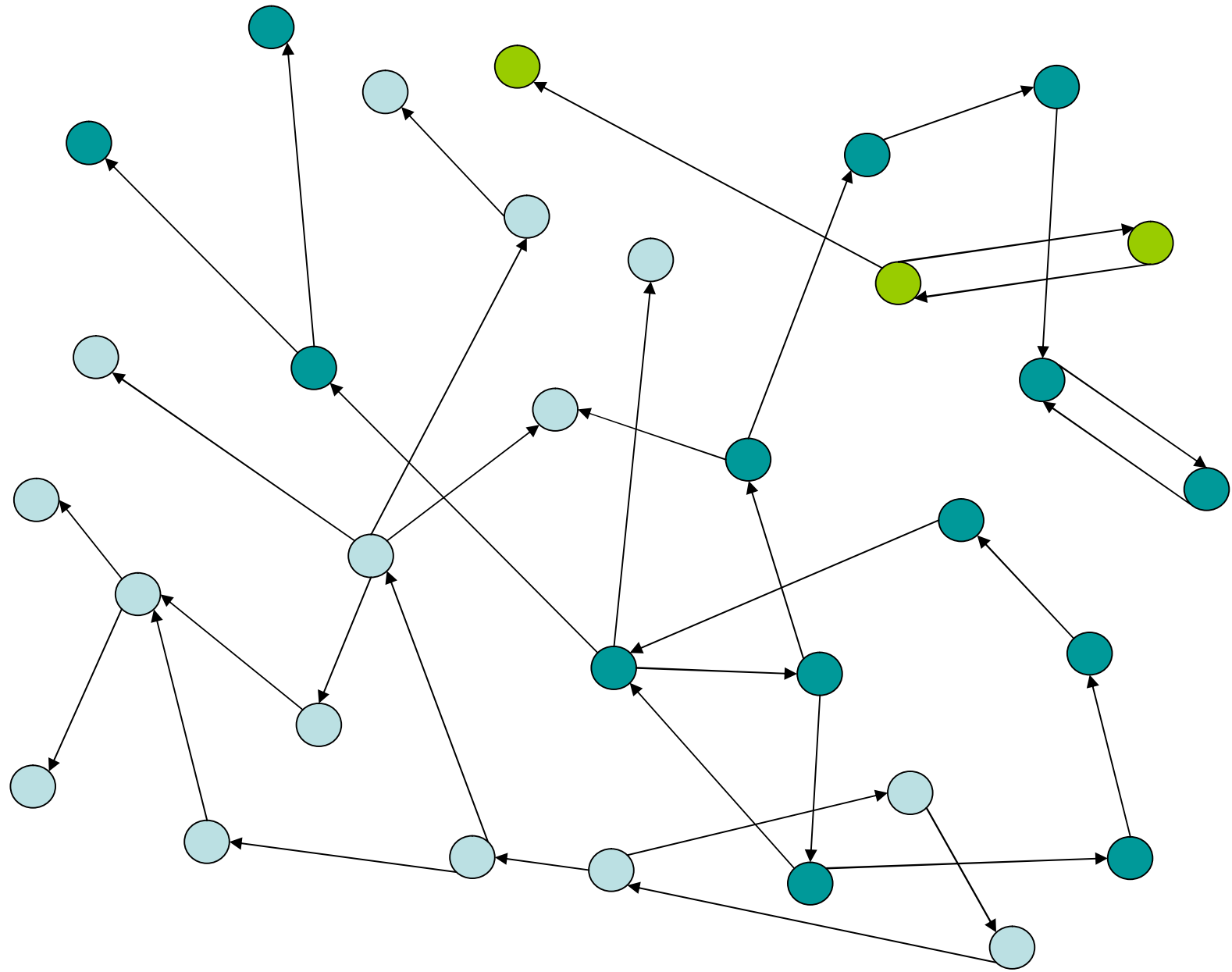- Are networks modular?

What does 'modular' mean?

- Are there only one connected component or several

- In scale free graphs there are hub nodes (nodes with high degree keeping everything together) and there is a theory that networks fall into pieces (modules) if the hub nodes are removed – is this the case for real netoworks
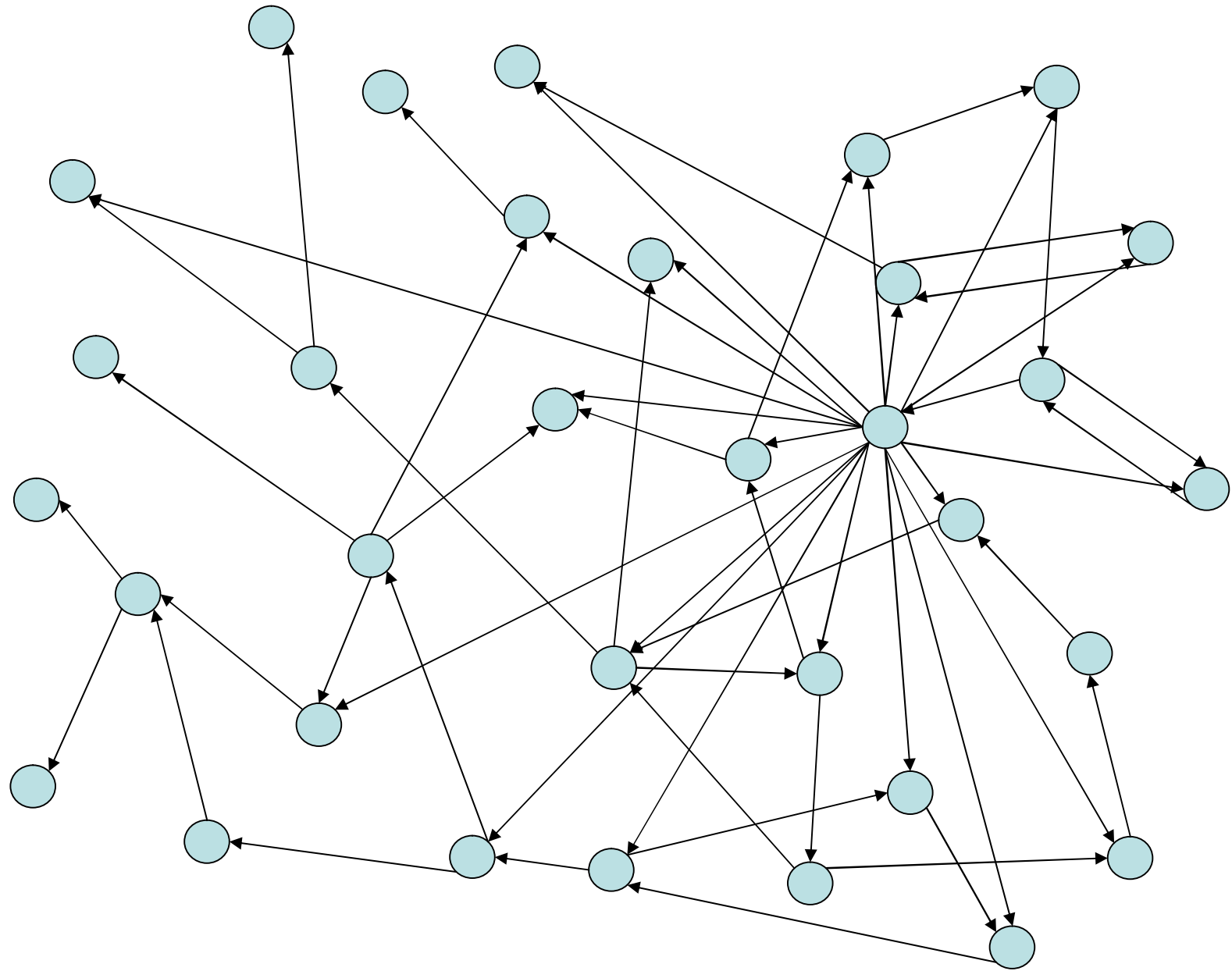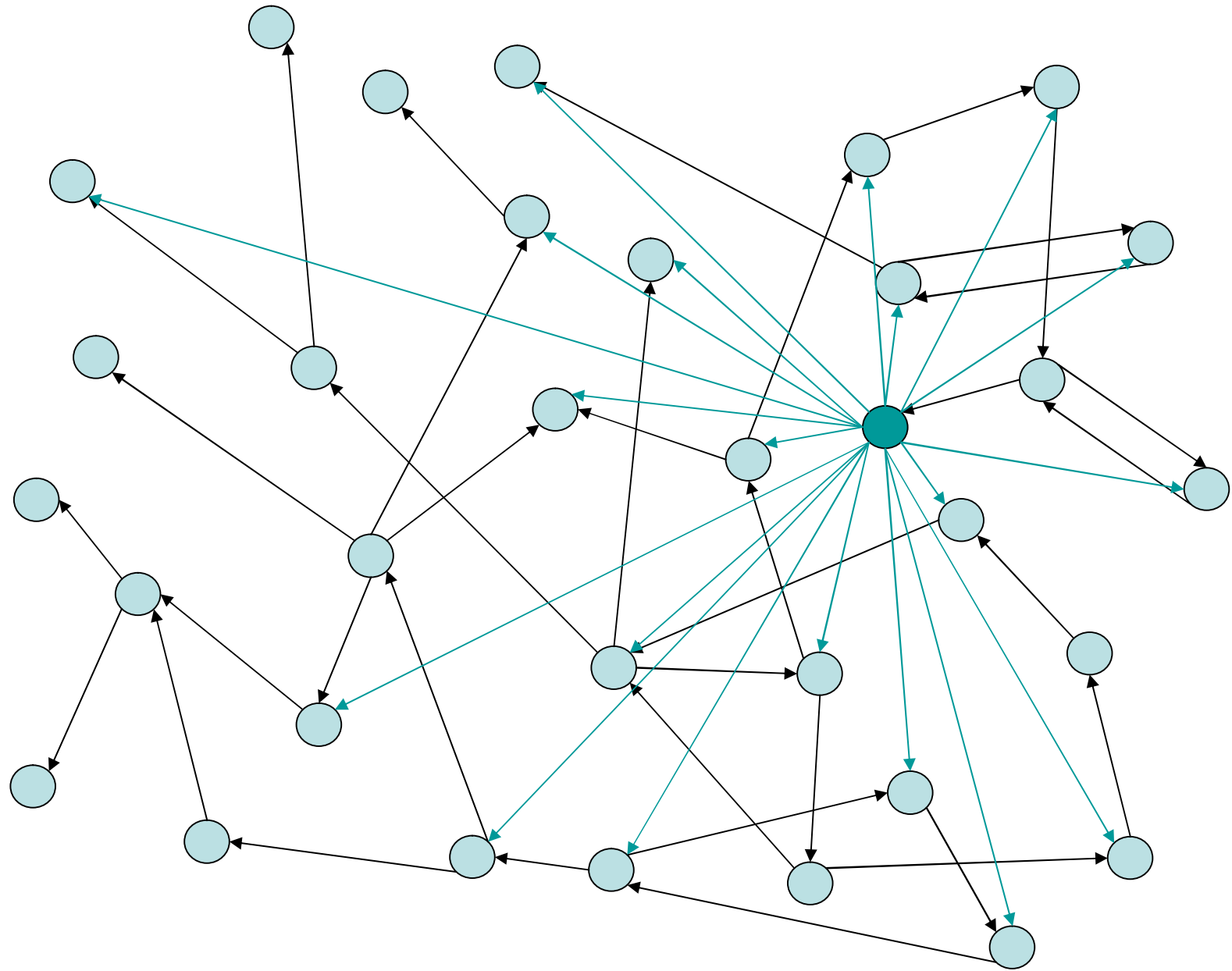
Looking for modules

		full	removed		
			1%	5%	10%
ChIP network	largest	2403	2201	1859	1721
	second	11	11	11	11
	total number	3	3	4	6
in-silico network	largest	5583	5416	4988	4259
	second				
	total number	1	1	1	1
mutant network	largest	4095	3209	2301	1815
	second	2	2	3	3
	total number	2	2	3	8

Network modularity

- On static topology level there are no obvious modules in yeast transcription regulation network

- This does not mean however that there are no modules?
  - there is evidence for modules
  - More subtle methods may be needed to find them

What have we learned on the topology level?

- Comparison of different networks shows that we have some idea of what the true topology is like, but it is far from perfect
- The network topology is roughly scale free
- There are not obvious modules in these networks on the topology level – one giant component

Gene Networks - four levels of hierarchical description

- **Parts list** – genes, transcription factors, promoters, binding sites, …
- **Topology** – a graph describing the connections between the parts
- **Control logics** – how combinations of regulatory signals interact (e.g., promoter logics)
- **Dynamics** – how does it all work in real time

Gene Networks - four levels of hierarchical description

- **Parts list** – genes, transcription factors, promoters, binding sites, …
- **Topology** – a graph describing the connections between the parts
- **Control logics** – how combinations of regulatory signals interact (e.g., promoter logics)
- **Dynamics** – how does it all work in real time

More complex interactions
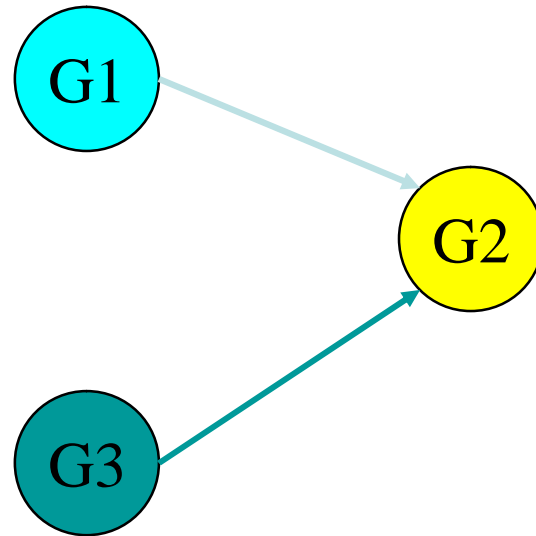
Logics
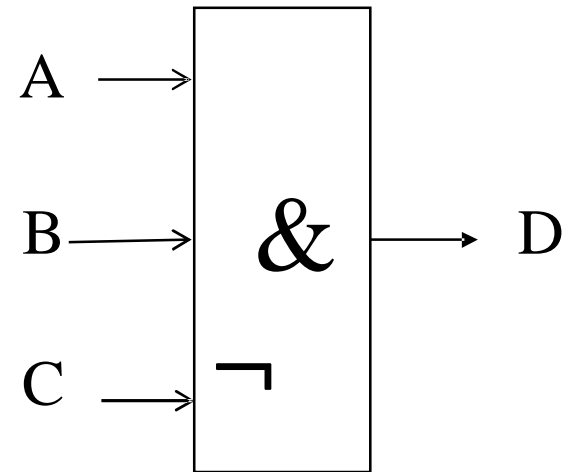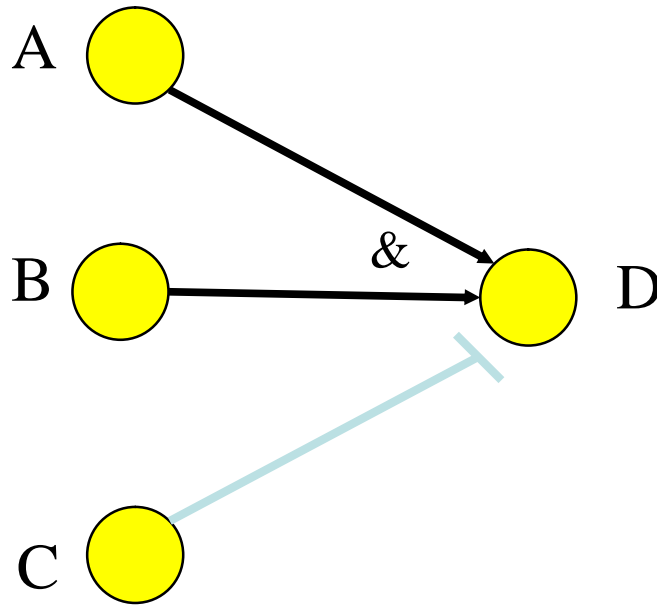
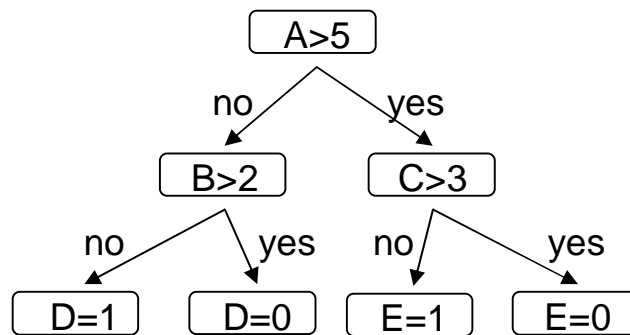$$D = A \mathbin{\&} B \mathbin{\&} \neg C$$

Control functions

- Discrete vs. continuous

$$D = A \,\&\, B \,\&\, \neg C$$

$$D = w_1 A + w_2 B + w_3 C$$

Decision trees

A>5

no — yes

B>2   C>3

no — yes   no — yes

D=1   D=0   E=1   E=0

if A>5 and B<=2 then D=1
if A>5 and B>2 then D=0
if A<=5 and C<=3 then E=1
if A<=5 and C>3 then E=0

- Decision tree for CLN2 gene in yeast
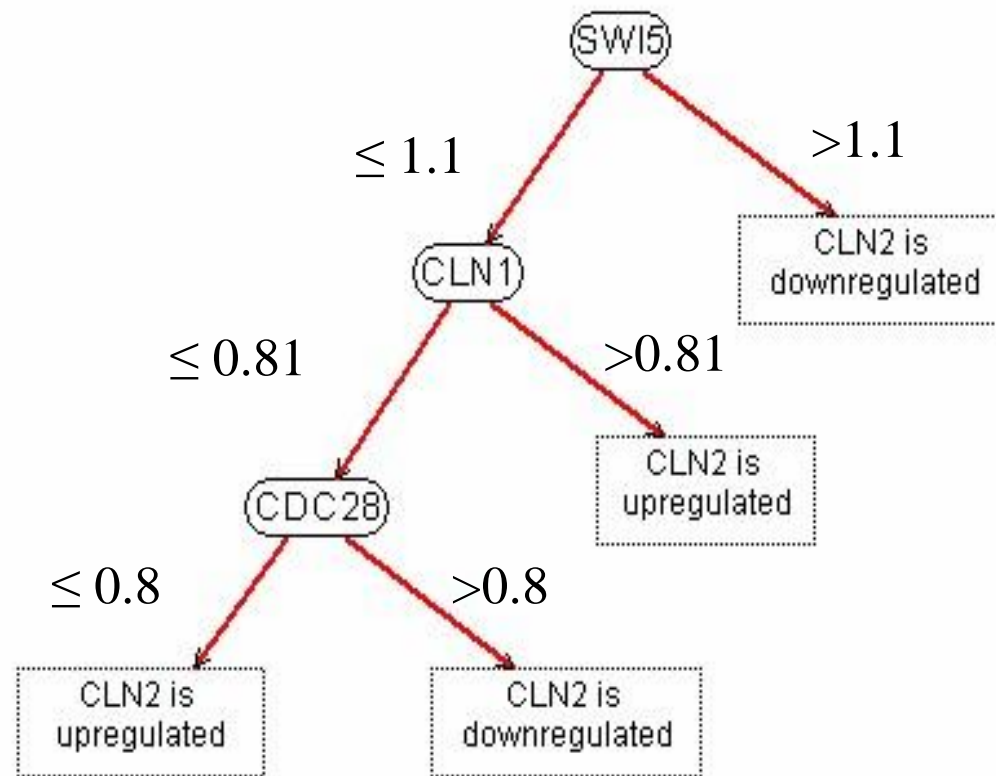
Logics – high throughput data is only now beginning to have impact

- Predicting gene expression from combination of expression levels of other genes (Soinov et al, 2003)

  - Limited to about 20 genes

  - For instance, by choosing genes well known to be involved in yeast cell cycle regulation it is possible to derive decision trees describing the combinatorial regulatory effects for these genes

  - At least some of the conclusions are supported by a priori knowledge

What is known about the regulatory logics from classical low throughput approaches?
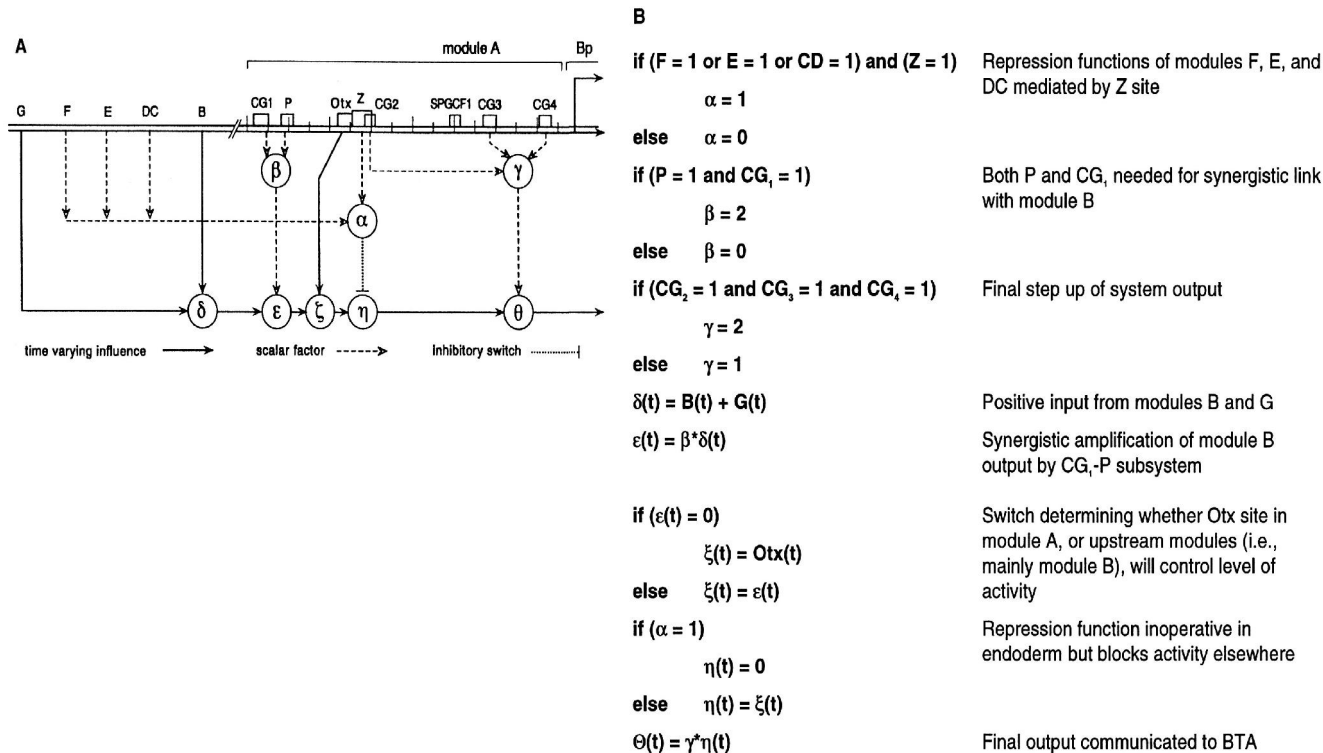
Yuh, C.H., Bolouri, H. and Davidson, E.H. (1998) Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene. Science 279, 1896-902

Boolean, linear and decision tree concepts are all used – 12 input variables

Probabilistic approaches

Canalizing Boolean functions

- There is one input and one value for that input that determines the output regardless of the values of other inputs
  - $F = x \lor y$ – canalizing – $x=1$ -> $F=1$
  - $F = x \ \& \ y$ – canalizing – $x= 0$ -> $F=0$
  - $F = x \oplus y$ – not canalizing – none of the values of none of the inputs can determine the value of F
- For Boolean functions of many inputs only a small number of the possible functions are canalizing

Gene Networks - four levels of hierarchical description

- **Parts list** – genes, transcription factors, promoters, binding sites, …

- **Topology** – a graph depicting the connections of the parts

- **Control logics** – how combinations of regulatory signals interact (e.g., promoter logics)

- **Dynamics** – how does it all work in real time

Classification of dynamic network models

- Continuous versus discrete state (e.g, boolean)
- Deterministic versus probabilistic state transitions (e.g. differential equations versus Bayesian models);
- Ignoring spatial effects vs modelling spatial effects

Differential equation based models

The basic assumption – the rate of changes in gene product abundance at a particular time are determined by the abundance of gene products at the time

$g_i(t)$ – the abundance of the product of gene i at time t
w_{ij} – the weight of the contribution of gene j to the expression of gene i

Differential equation based models

Difference equation model:

$$g_1(t+\Delta t) - g_1(t) = (w_{11}\ g_1(t) + \ldots + w_{1n}\ g_n(t))\ \Delta t$$
$$\ldots$$
$$g_n(t+\Delta t) - g_n(t) = (w_{n1}\ g_1(t) + \ldots + w_{nn}\ g_n(t))\ \Delta t$$

where:

$g_i(t)$ – the abundance of the product of gene i at time t

w_{ij} – the weight of the contribution of gene j to the expression of gene I

The main problem – we don't know the constants w_{ij}

Differential equation model for drosophila embryo development
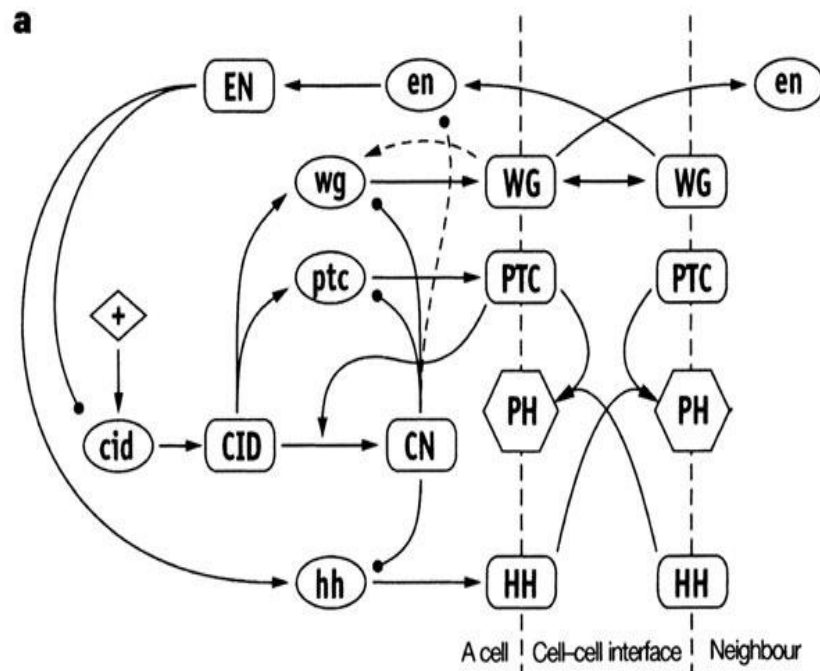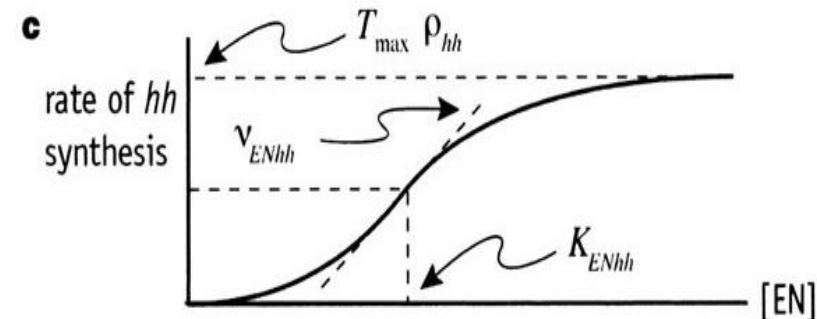
von Dassow, G., Meir, E., Munro, E.M. and Odell, G.M. (2000)
The segment polarity network is a robust developmental
module. Nature 406, 188-92.

Synchronous Boolean networks – the assumptions

- Each gene the system (cell) can be in one of two states –
    - 'expressed' – 1,
    - 'not expressed' – 0
- The genes can switch from state to state all simultaneously in synchronous manner
- The next state of each gene is determined by previous states of all genes by Boolean functions describing the network

$$Y=X\&Z, \quad X=Y, \quad Z= \neg X$$
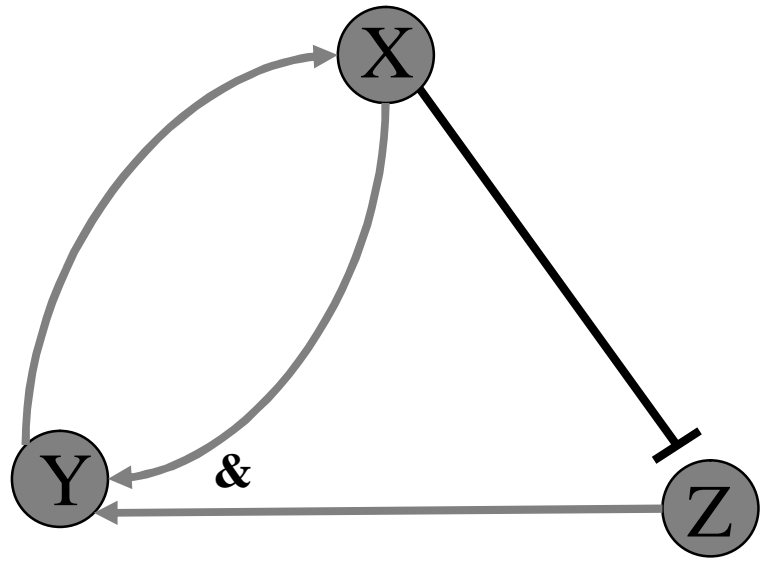
t			t+1		
X	Y	Z	X	Y	Z
0	0	0	0	0	1
0	0	1	0	0	1
0	1	0	1	0	1
0	1	1	1	0	1
1	0	0	0	0	0
1	0	1	0	1	0
1	1	0	1	0	0
1	1	1	1	1	0

State transitions

$$111 \rightarrow 110 \rightarrow 100 \rightarrow 000 \rightarrow 001 \circlearrowright$$

$$011 \rightarrow 101 \leftrightarrows 010$$
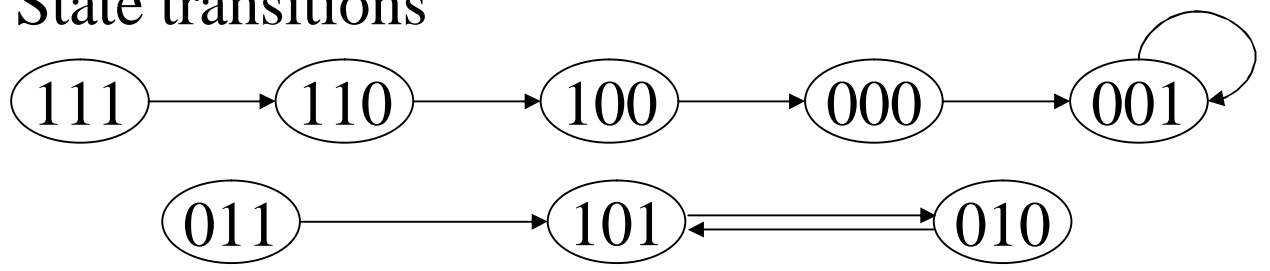
State space

Reverse engineering:

- Given the state space transitions:
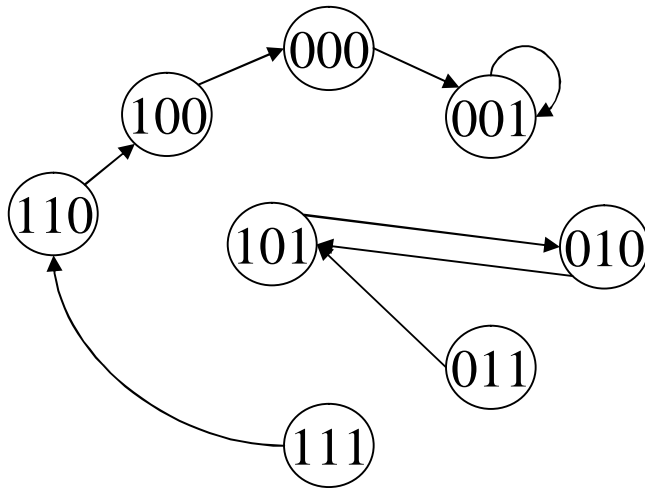
- Reconstruct the network:
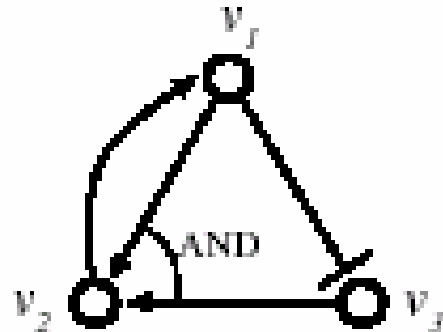
Reverse engineering problem

- On one hand the problem is trivial – the stage space immediately gives one a transition table, which is an equivalent representation to the wiring diagram

- However the problem of building the smallest wiring diagram from the table is NP-hard, i.e., it takes exponential time to do this

- For small networks (3 genes as above) this is not a problem, but for thousands of genes this is not a solution

Exponential algorithm

- Assume that all genes depend on all, i.e., in the wiring diagram connect each to all
- The Boolean function is the disjunctions of all vectors as given in the table
- This gives a hugely long Boolean functions for each gene (i.e, $n2^n$ for a network of each gene)
- The minimisation of this Boolean function to the smallest equivalent one is a classic NP hard problem

Solution

- Instead of the minimal possible network look for simply 'small' network
- Somogyi et al – algorithm using mutual information – not clear how good is this heuristics

Attractors in the state space
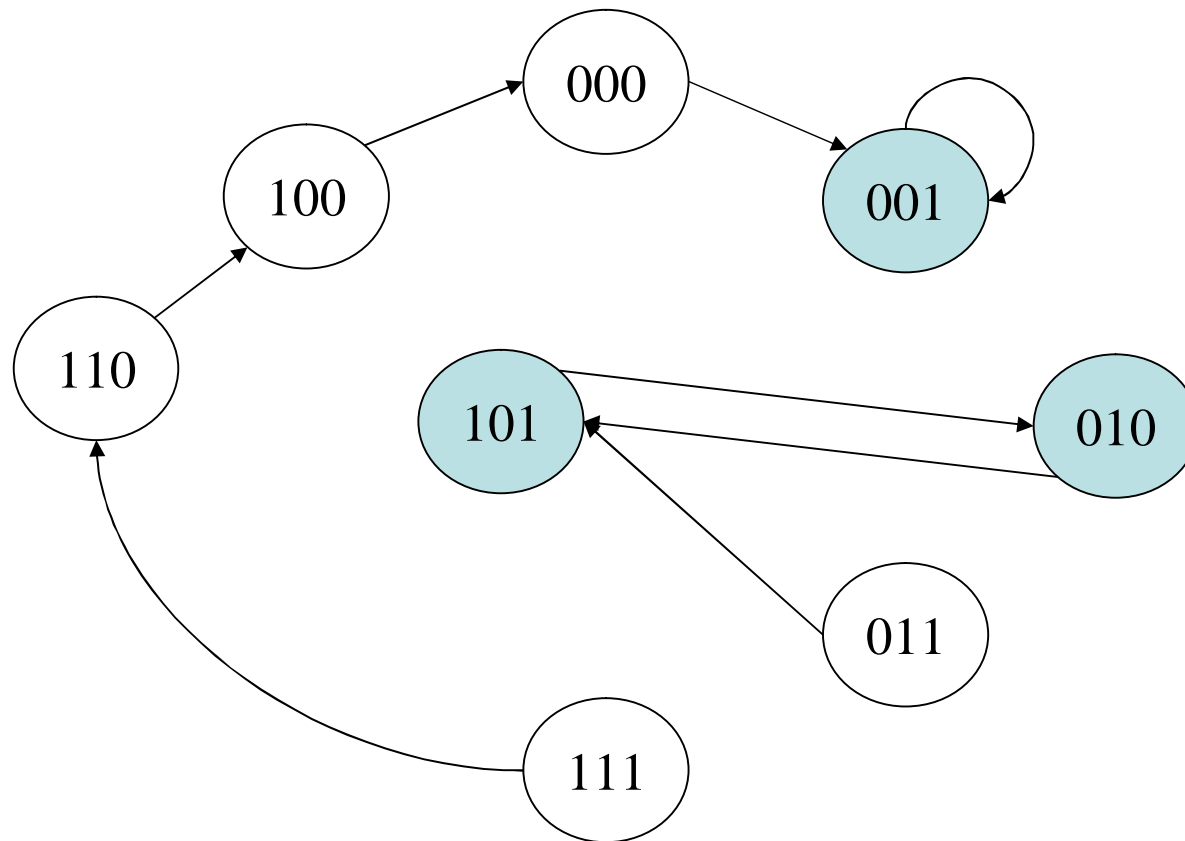
Canalizing Boolean functions

- There is one input and one value, which determines the output regardless of the values of other imputs
- $F = x \lor y$ – canalizing – $x=1$ -> $F=1$
- $F = x \& y$ – canalizing – $x= 0$ -> $F=0$
- $F = x \oplus y$ – not canalizing – none of the values of none of the inputs can determine the value of F
- For Boolean functions of many inputs only a small number of the possible functions are canalizing

Kaufman's simulations

- Randomly constructed Boolean networks such that
  - the number of inputs of each 'gene' is small
  - the control functions are canalizing

have a property that
  - their state space consists of a relatively small number of attractors
  - most of the time the spend in attractor states

Attractors in the state space
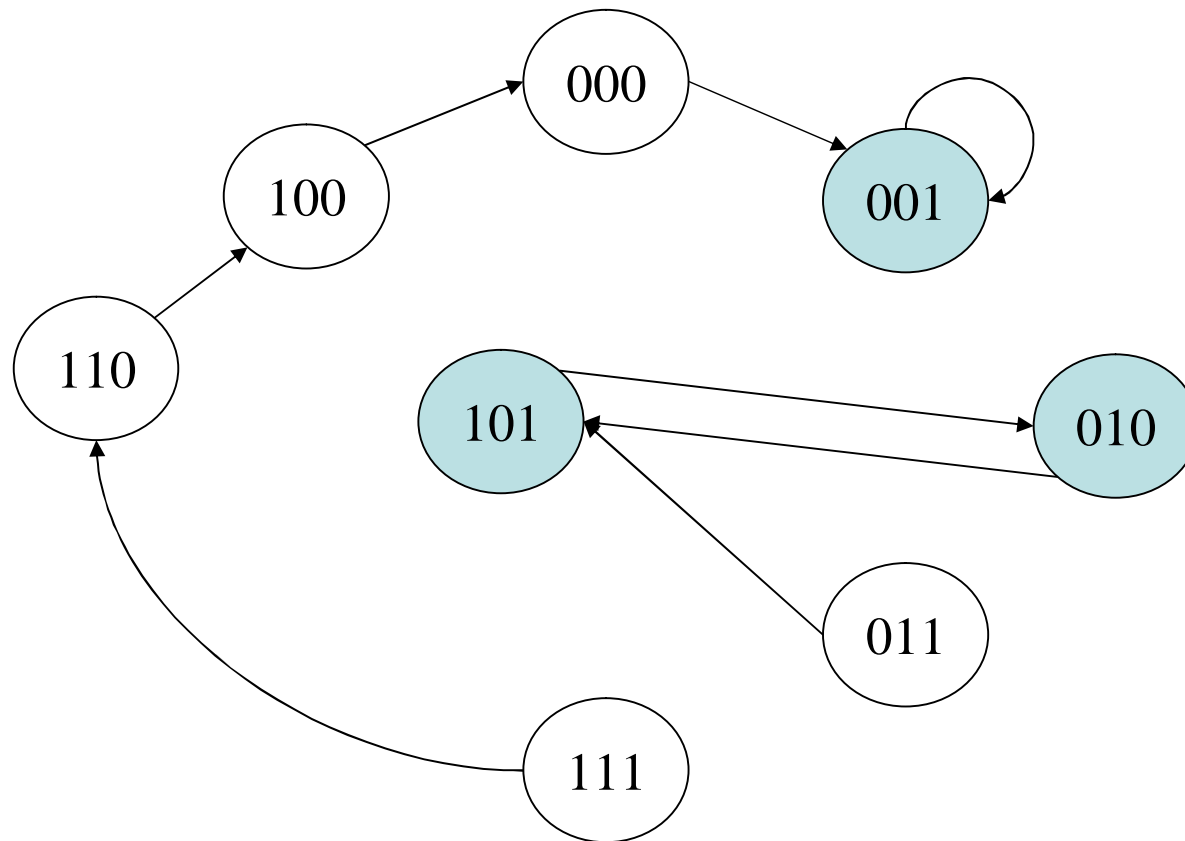
Kaufman's hypothesis

- Gene networks are predominantly controlled by canalizing functions

- Attractors are cell types

- He estimated that under certain conditions on network connectivity and assuming 100000 genes, there should be a few hundred different cell types

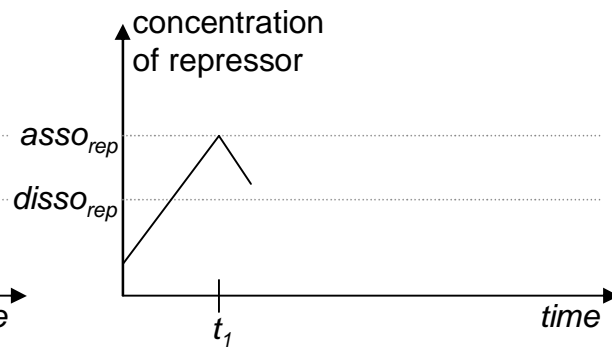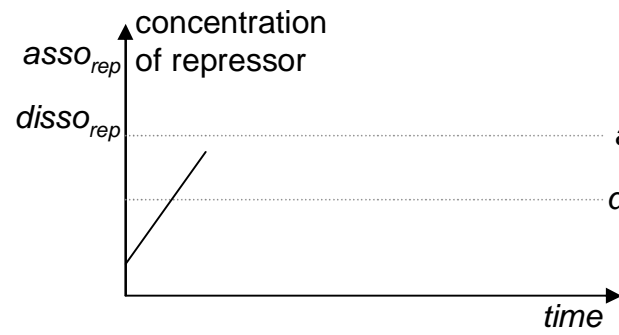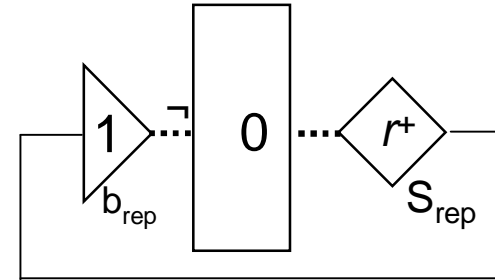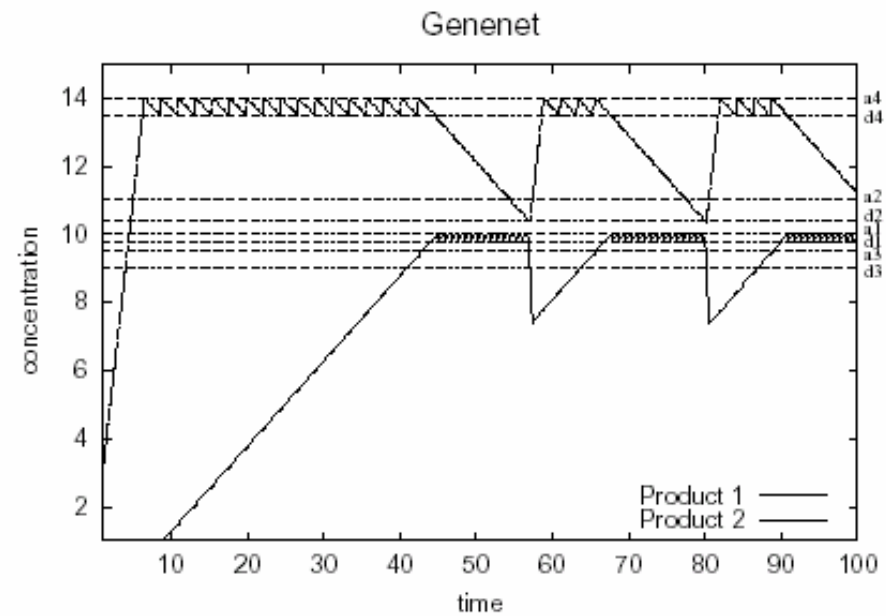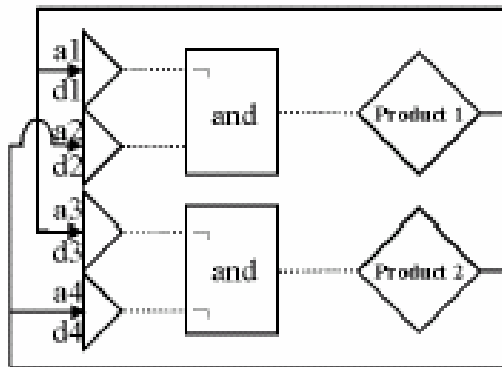# A hybrid models –
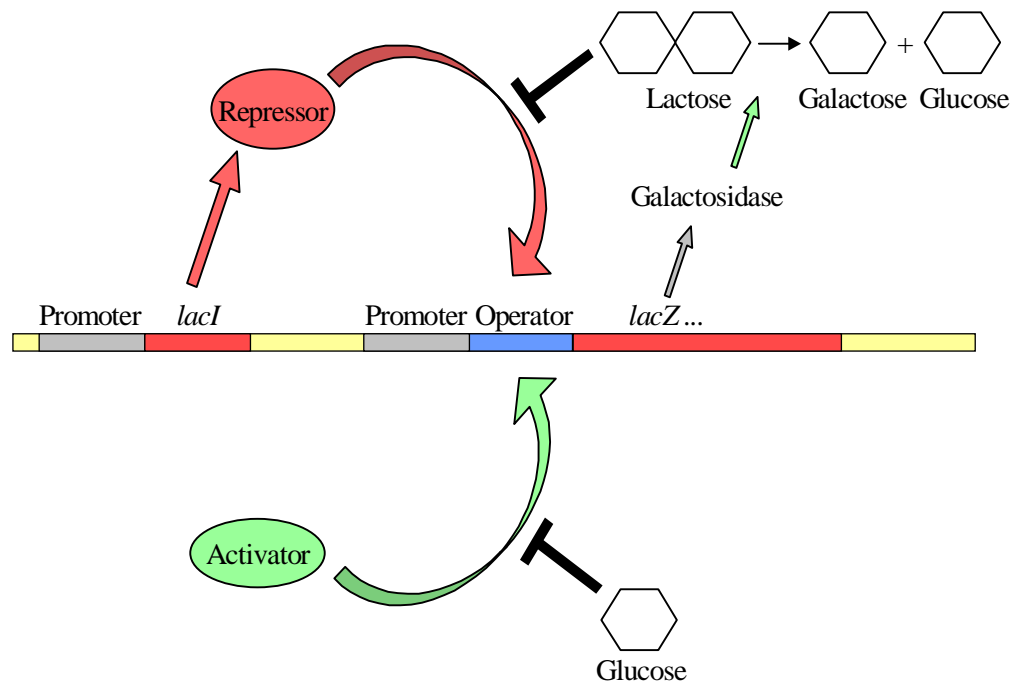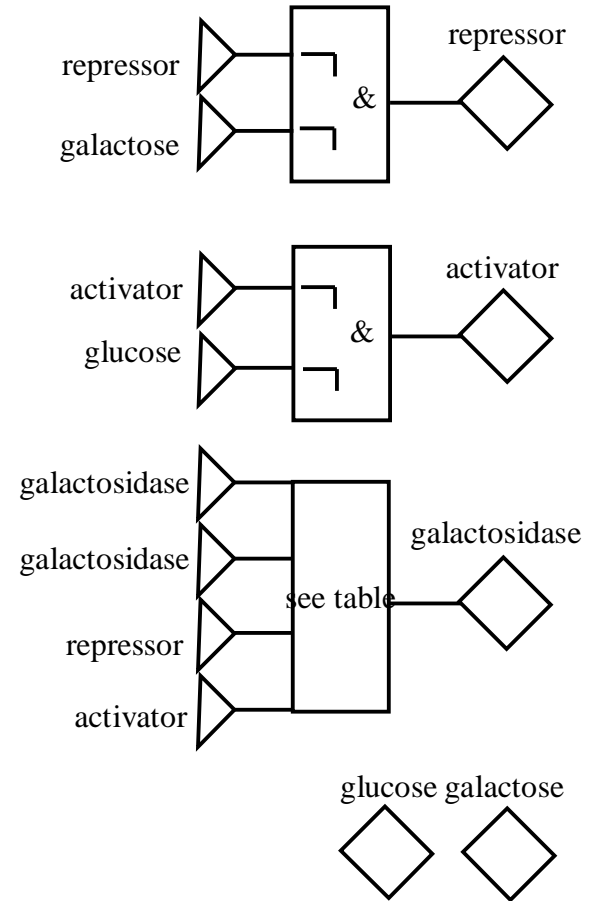# the finite state linear model
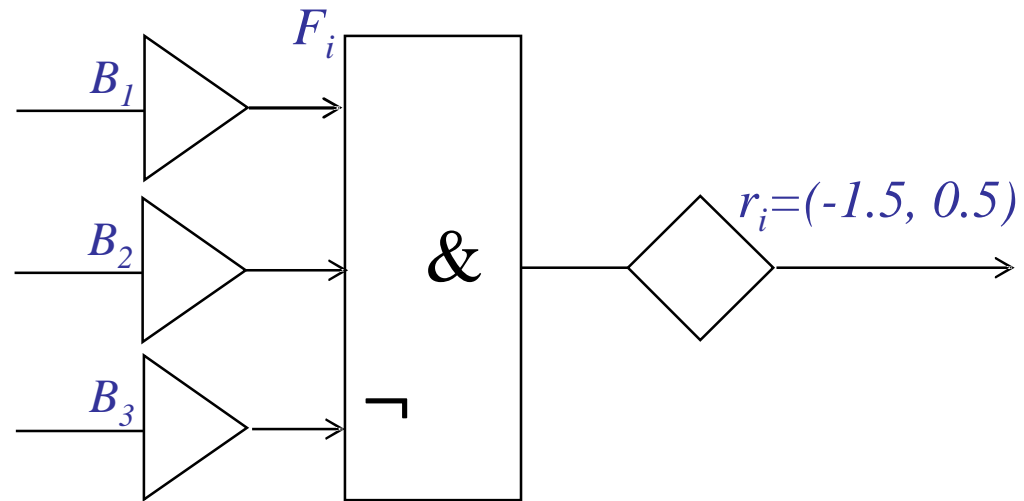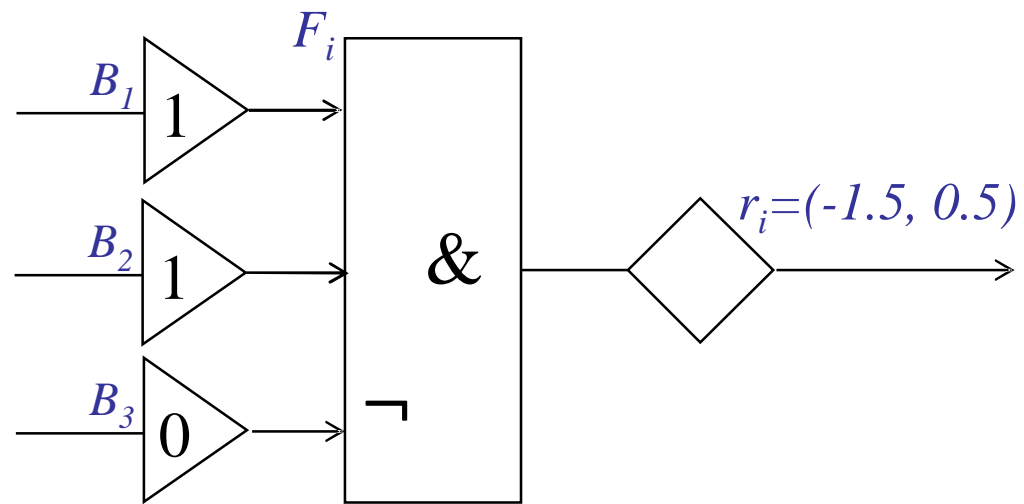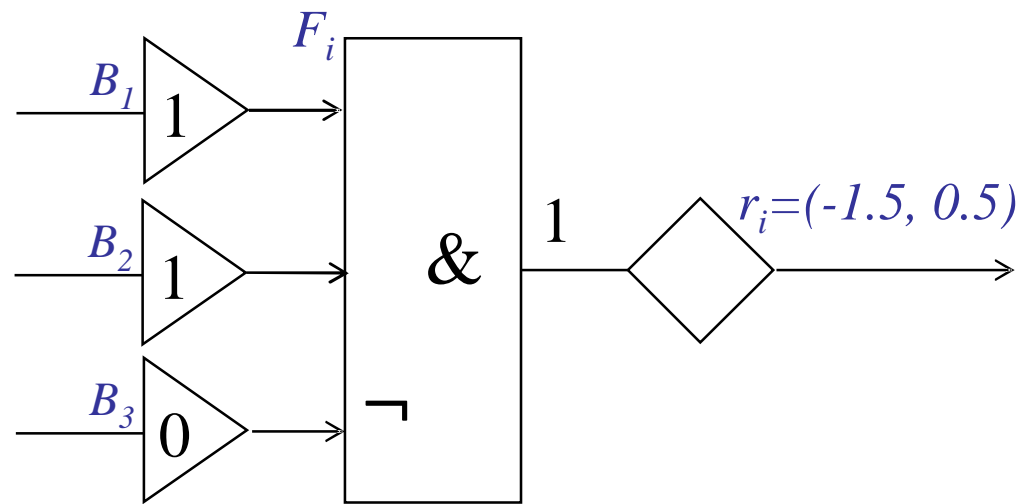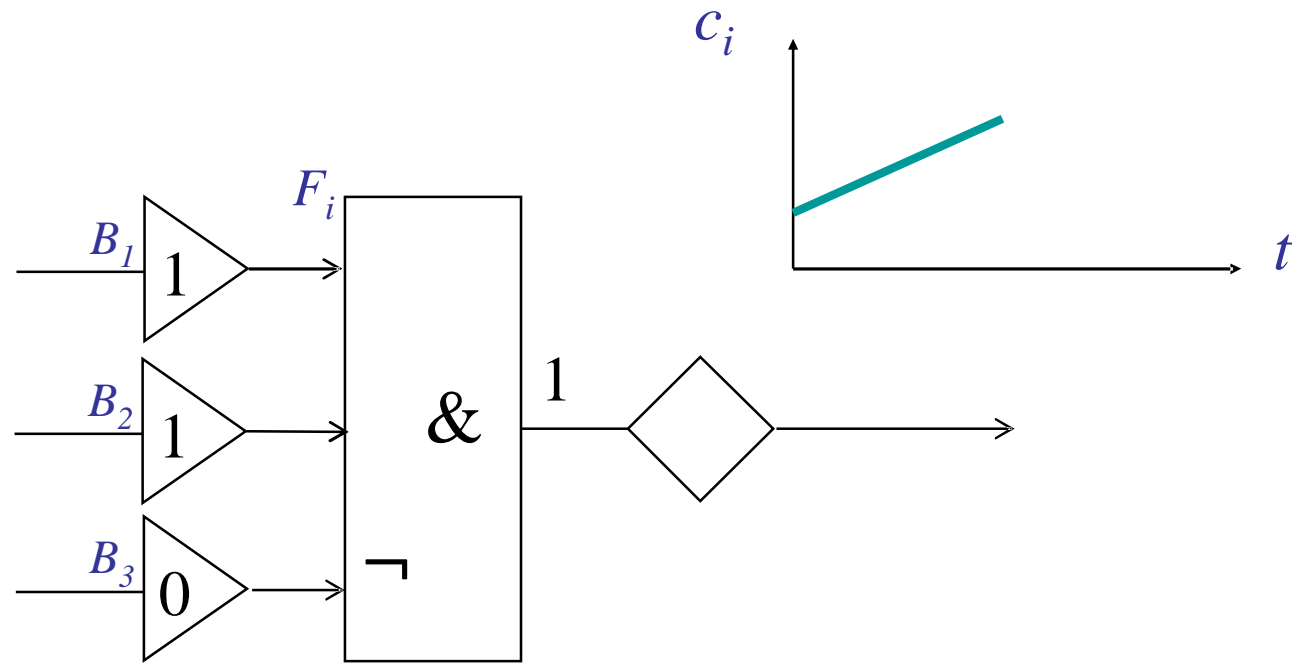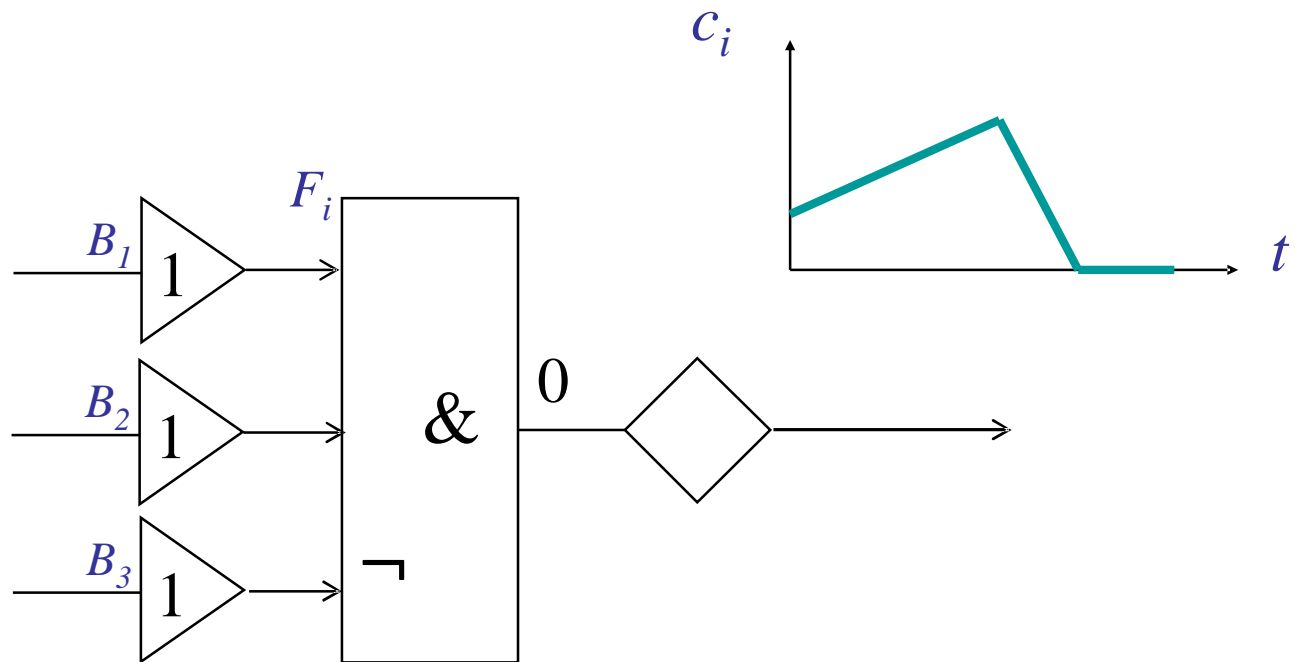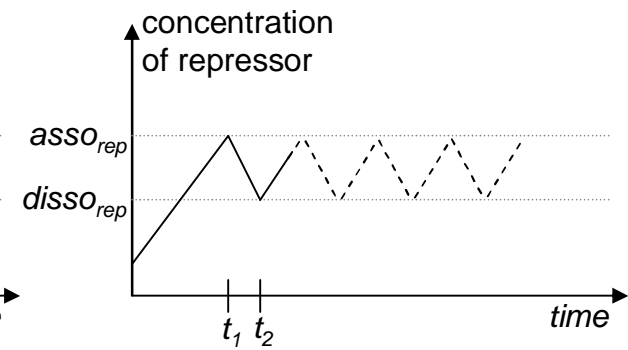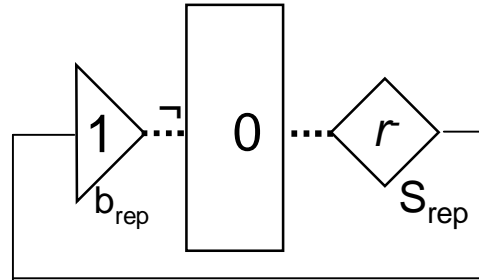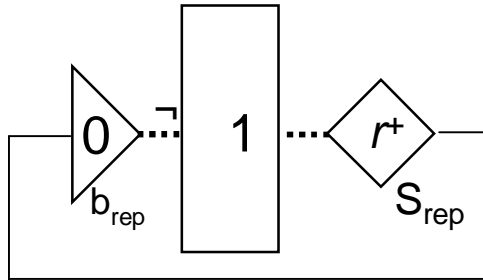
Simulations on a FSLM

Lac operon in E.coli bacteria

- There are two modes in E.coli – glucose or lactose utilisation mode that is regulated by the presence or absence of lactose
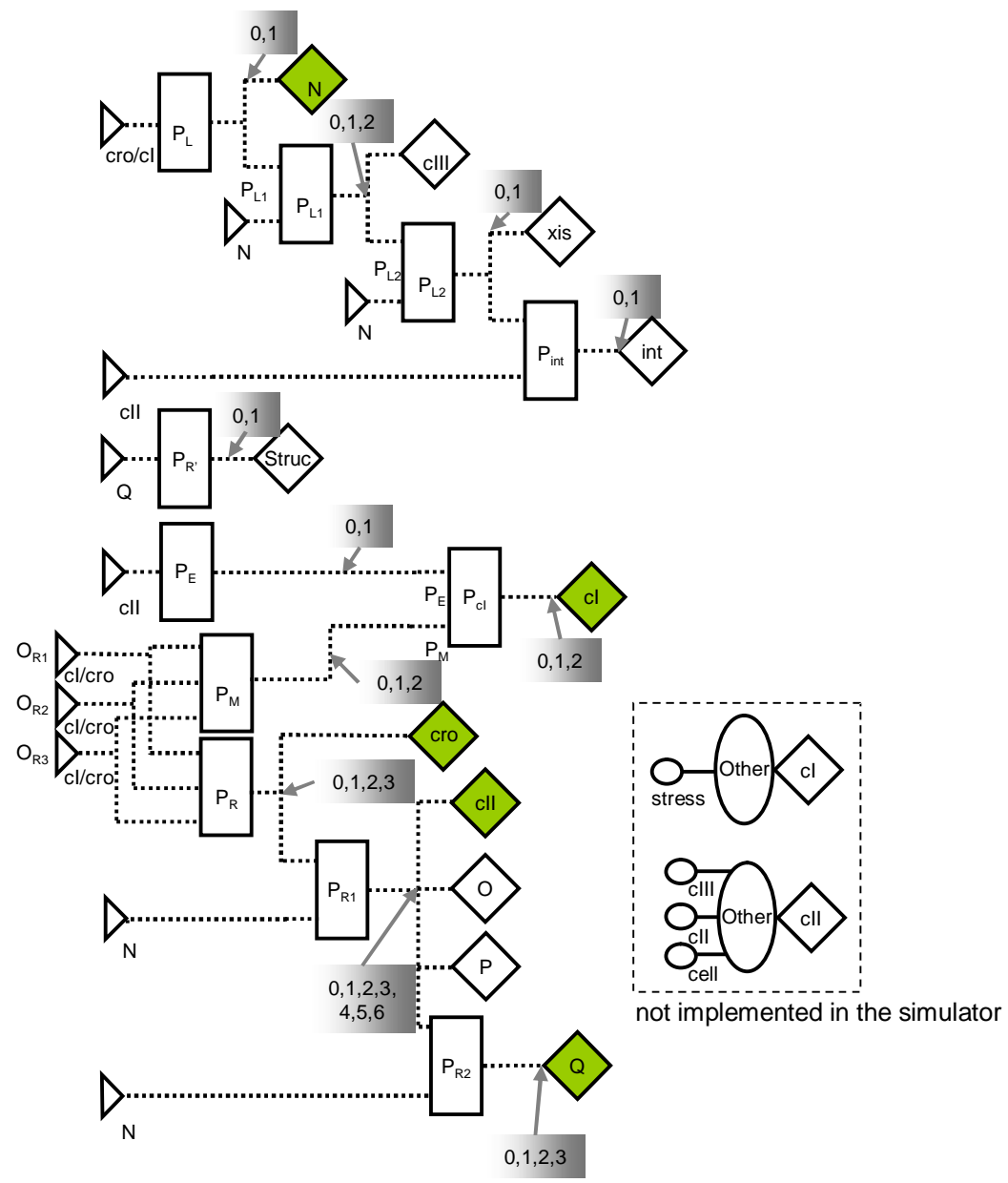
Finite state model for Lac-Operon network

FSLM representation

Finite state linear model for lambda phage lytic/lysogenic mode switch

not implemented in the simulator

Network dynamics – the state of art

- Most of the current dynamic models include less than 10 genes, and the knowledge used in the modelling mostly comes from traditional biology studies

- There are no convincing examples where high throughput technologies had a substantial impact on network modelling on the dynamics level yet

Conclusions – what have we learned on each level so far?

- **Parts list** – we are dealing with thousands to tens of thousands of elements in these networks, and hundreds to thousands regulating elements;
- **Topology** – may be tens of thousands of connections, it seems to be scale free, no obvious modules
- **Control logics** – a gene can be controlled by dozens of transcription factors in a rather complex way
- **Dynamics** – we are not yet able to model dynamics of genome scale transcription regulation networks

What do I hope you have learned in this course

- Some feel what real microarray data are like, some idea of the basic methods (if you didn't know this before)

- How to use ArrayExpress and Expression Profiler if you need this

- A flavour what is our current knowledge how genes are regulated and how little we know