The shaping of network features by biological constraints

Olivier Martin
UMR de Génétique Végétale, INRA
and
Laboratoire de Physique Theorique et Modeles Statistiques
University of Paris-Sud at Orsay, France

With:
Z. Burda, A. Krzywicki and M. Zagorski

OUTLINE

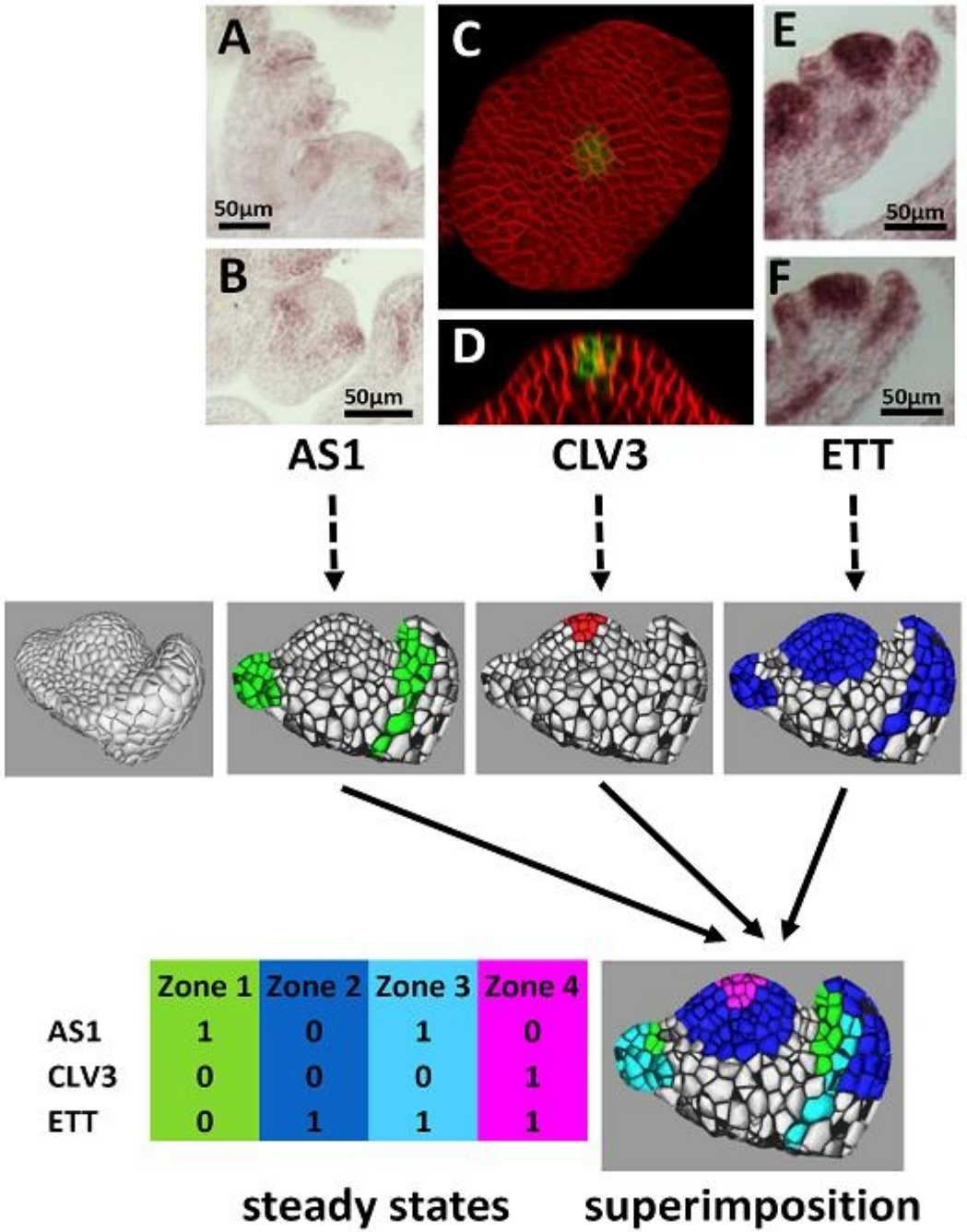
(Static) gene networks and influence graphs

Dynamical models of gene networks

How network structure and motifs emerge from functional constraints and mechanistic aspects

[I]: (Static) gene networks

Typical steps for “gene network reconstruction”

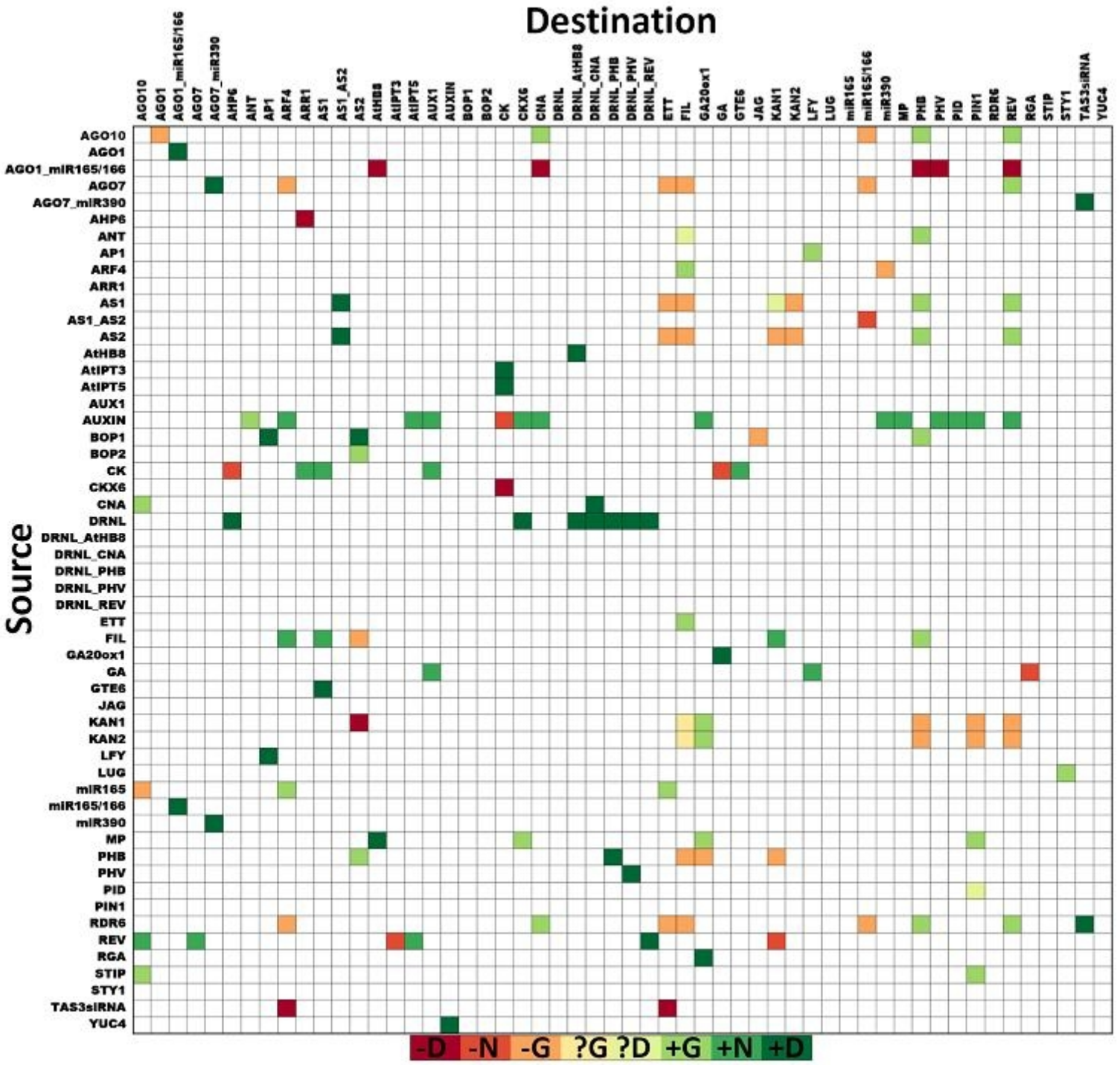


Reporter genes
2 or 3d imaging

Image segmentation
and analysis

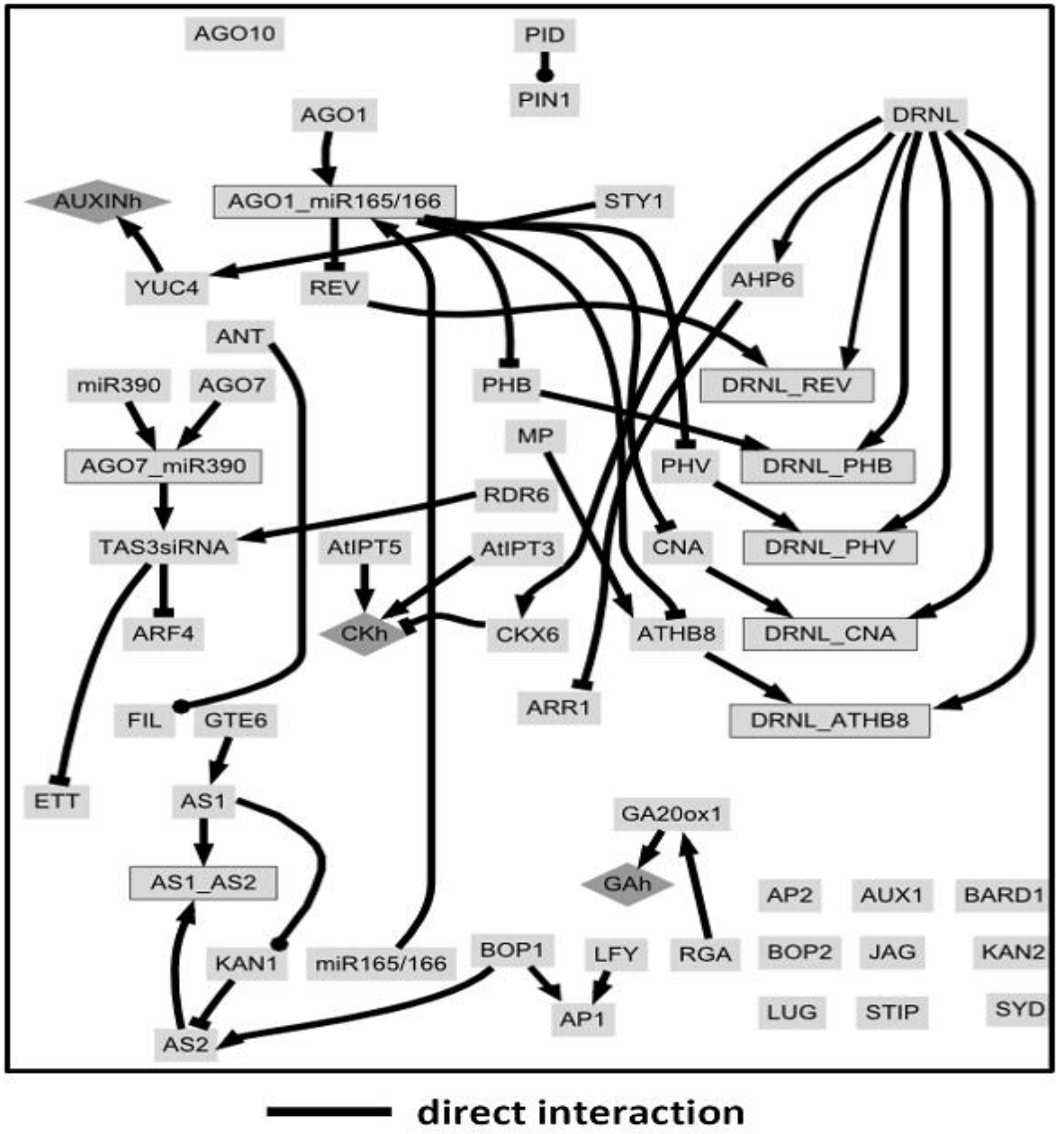
Virtual 3d reconstruction

Courtesy of F. Moneger



Genetic Analysis: effects of changing one gene on other genes

Resulting influence graph



[II]: Dynamical models of gene networks

To understand and act on gene networks, it is necessary to go beyond a description of interactions

Focusing on transcriptional dynamics:

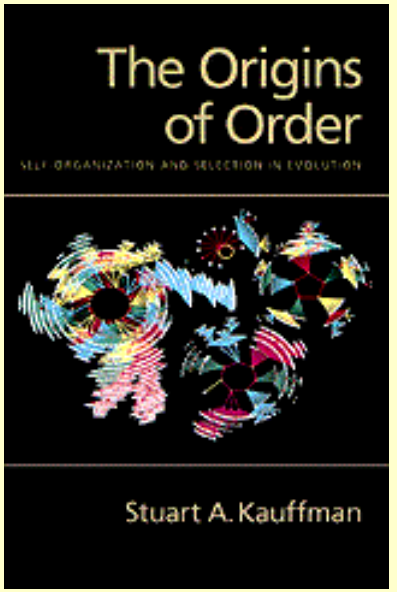
Boolean networks (genes on-off) (e.g., Stuart Kauffman)

Threshold networks (e.g., inspired by neural dynamics, Andreas Wagner)

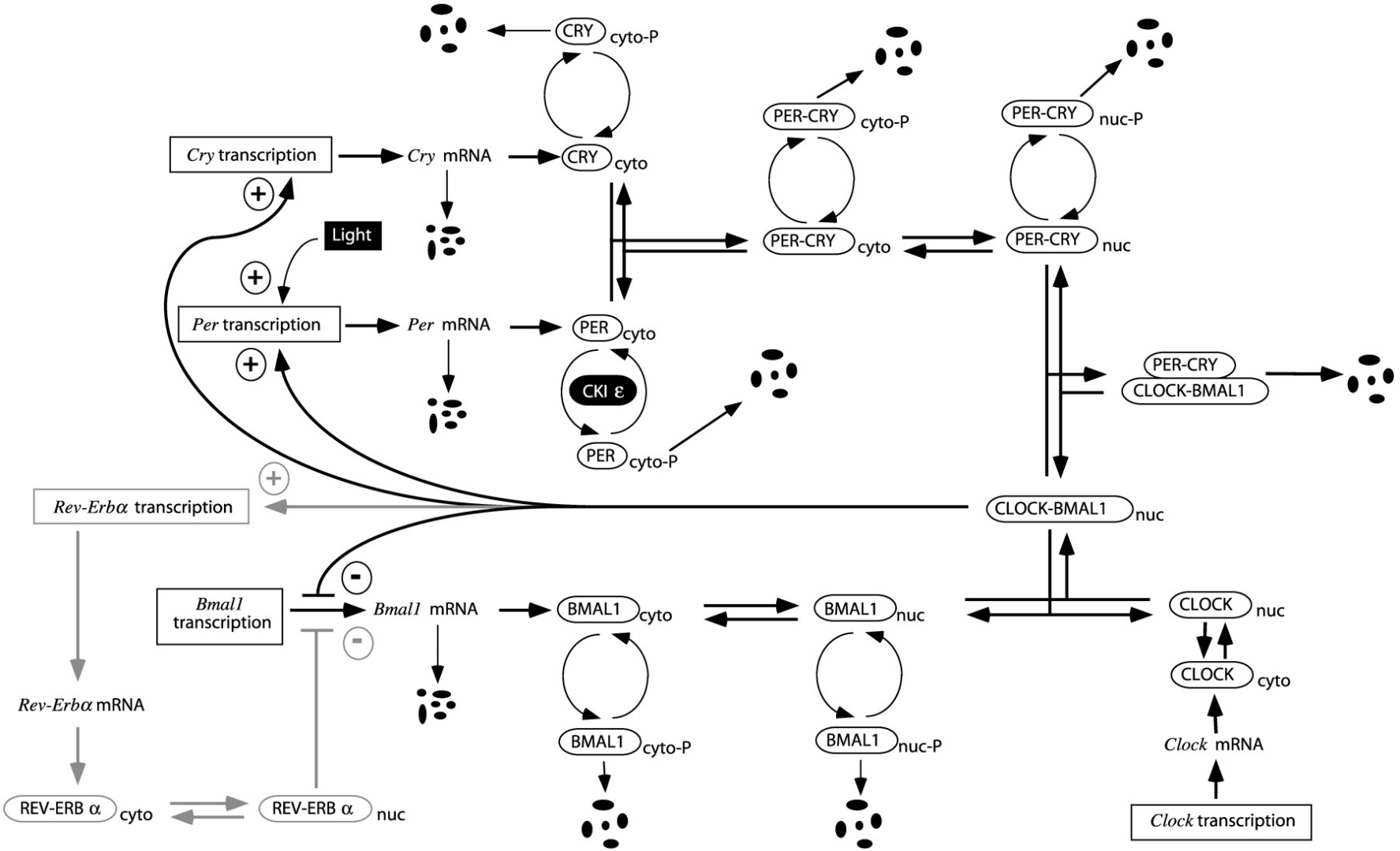
Piece-wise linear input-output relations (e.g., Hidde de Jong)

Differential equations with rates (e.g., Albert Goldbeter)

Individual event tracking (e.g., Daniel Gillepsie)



Example: model for the circadian clock in mammals



16-variable model including *per*, *cry*, *bmal1*, *rev-erbα*

Leloup J-C & Goldbeter A (2003) Toward a detailed computational model for the mammalian circadian clock. *Proc Natl Acad Sci USA.* 100: 7051-7056.

Wagner's gene network model

Model proposed by Andreas Wagner (Evolution, 1996)

Used since to fit quantitatively Drosophila gene expression levels during early development.

Used also for *in silico* studies by Bergman-Siegel, Li et al., Azevedo, … to get insights into genetic and evolutionary questions such as canalisation, epistasis, robustness, …

Regulatory network of ***N*** transcriptional regulators

We focus on their expression levels

$$\mathbf{S}(t)=(S_1(t), S_2(t), \ldots, S_N(t))$$

at some time *t* during a developmental or cell-biological process and in one cell or domain of an embryo.

Regulatory dynamics of the Wagner model

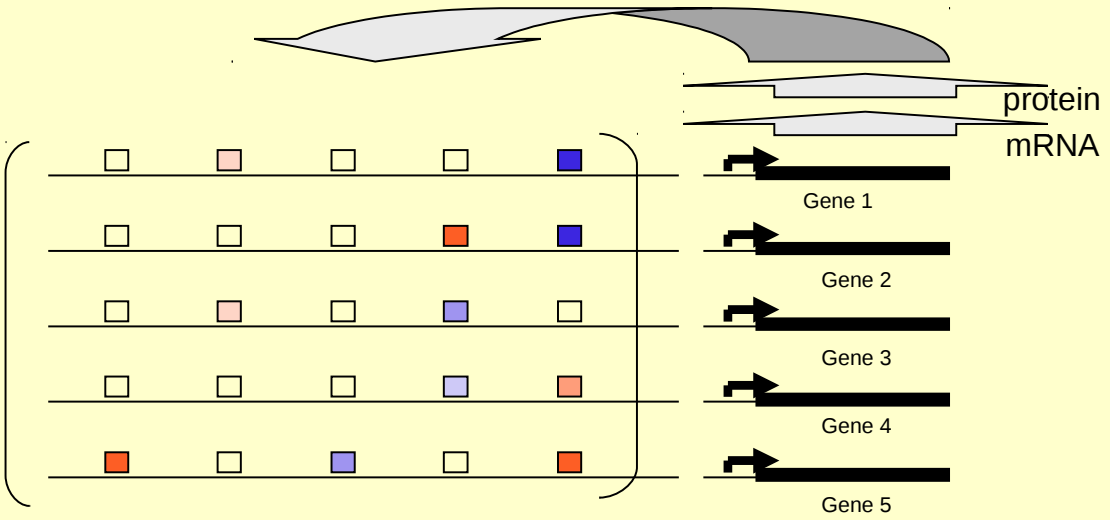
Discrete time **dynamics** for simplicity:

$$S_i(t+1) = F(W_{i1} S_1(t) + W_{i2} S_2(t) + \ldots)$$

where *F* is a sigmoidal function. Analog of a Hopfield neural network using a *threshold* weighted sum for input → output.

The elements W_{ij} of this matrix indicate the strength of the regulatory influence that gene ***j*** has on gene ***i***. This influence can be either activating ($W_{ij}>0$), repressing ($W_{ij}<0$) or absent ($W_{ij}=0$).

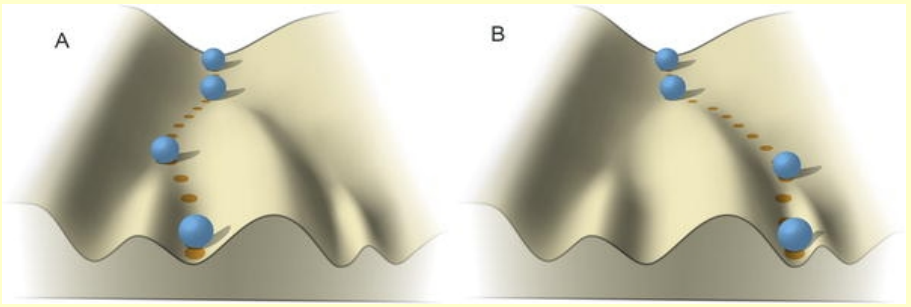
W =



Genotypes and phenotypes

The matrix ***W*** represents the **genotype** of the network.

The **phenotype** is the behavior of the regulatory circuit, given for instance by its expression pattern(s). In the case of developmental processes, we can take the initial conditions as given (e.g., by maternal mRNA or upstream genes); "proper" development requires that the network reach at large times the right *target pattern* for the S_i. One can also simply impose multi-stability: force several patterns to be steady states of the transcriptional dynamics, thereby modeling different cell types. In the case of circadian oscillators or the cell cycle, the target pattern will be a periodic orbit.



Strong selection limit:

A genotype is **viable** if and only if it goes to the steady state given by the target pattern. This corresponds to a *0-1 fitness landscape*.

Some drawbacks of this model

The work on the Wagner model has used binary expression levels: a gene is either on or off. Allow for continuous values but base the extension on a *motivated* choice of the sigmoidal function for input-to-output.

The number of non-zero entries in ***W*** is *a priori* arbitrary. How can the actual number of interactions emerge without putting in constraints by hand?

In an evolutionary context, since the entries in ***W*** have no molecular origin, their evolutionary dynamics is arbitrary. Thus generalize the model so that the interaction strengths have a biophysical meaning.

[III]: How functional constraints and mechanistic aspects shape network structure

Q1: Why are gene networks sparse ?

Paradox : there are far fewer sparse than non sparse networks. So why do real gene networks have low in degree?

Biologically, having interactions requires TF binding, and mutations can break these interactions. The more interactions there are, the easier it may be to disrupt function.

Can this argument be put on sound footing? It could be that having multiple weak interactions leads to more robust behavior than having a few strong interactions.

We will consider this using a model taking into account the molecular nature of the interactions.

Molecular nature of genetic interactions: modeling transcription factor binding to DNA

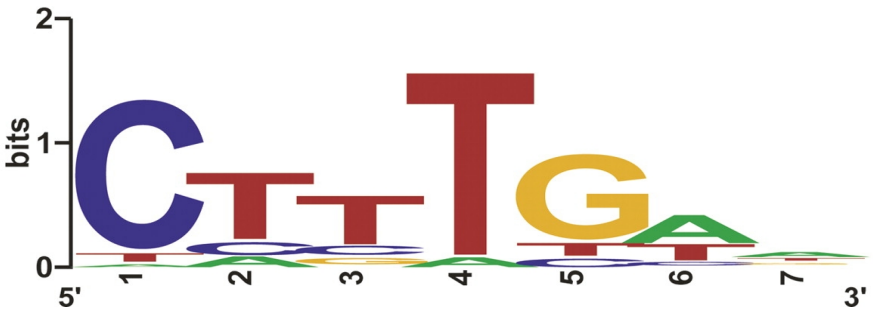
(1) Consensus sequence, e.g., TATAAT
Each mismatch lowers the binding affinity.

(2) Position Weight Matrix
Provides a score for each sequence. Corresponds to a model for binding via additive affinities from each site (Berg + von Hippel).

A

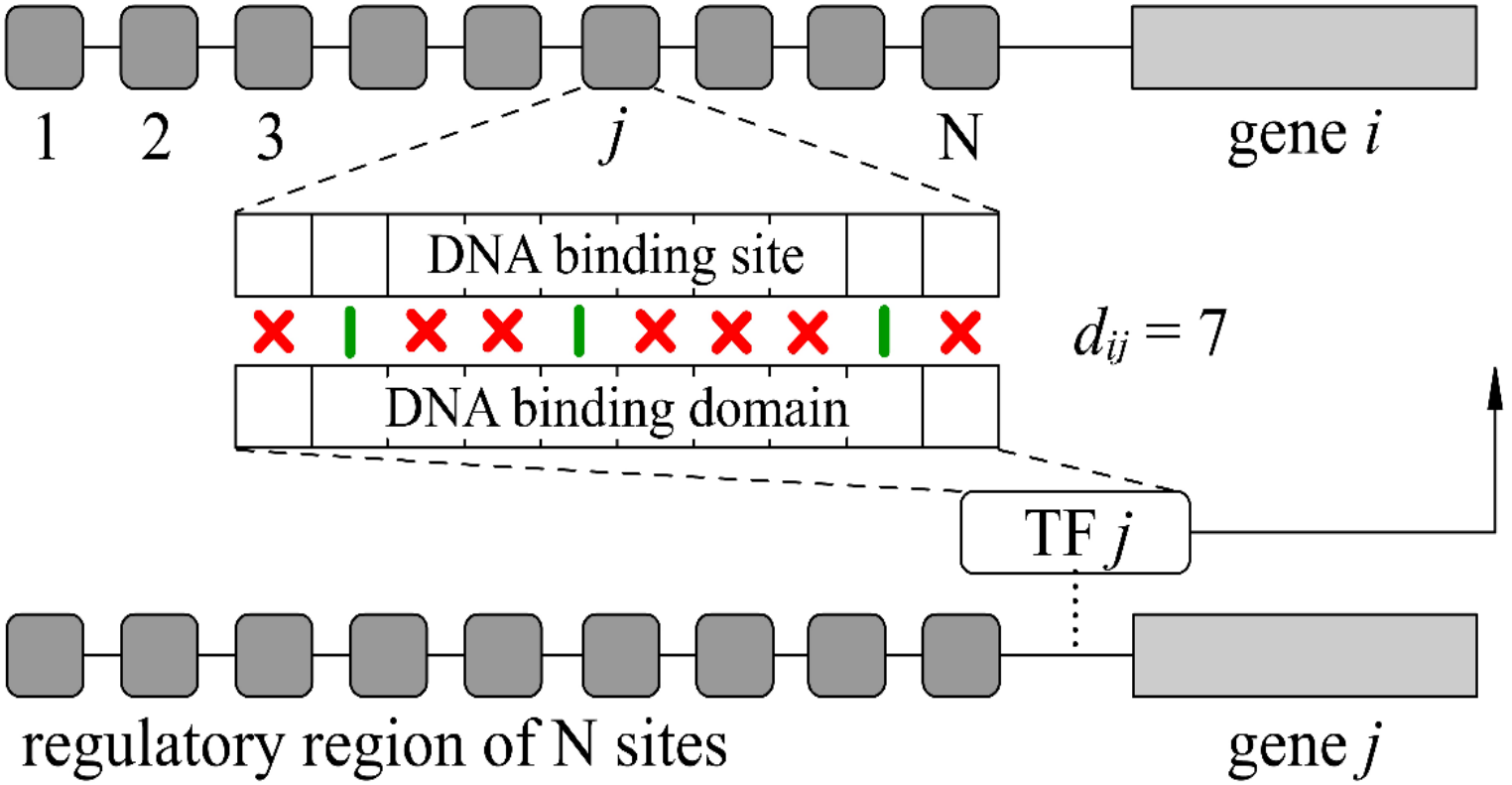
	1	2	3	4	5	6	7
A	1	4	1	2	0	17	13
C	28	5	5	0	3	3	2
G	0	0	4	0	25	1	7
T	2	22	21	29	4	10	9

B



Use of mismatches in a “molecular” model

Consider N regulatory genes, each coding for a transcription factor. The expression of gene *i* is potentially affected by these transcription factors (TF) if they bind to its regulatory region. We simplify by considering that each TF *j* can bind to a *dedicated* site in the regulatory region of each gene *i*.



The binding is thermodynamic (equilibrium), with a free energy linear in the mismatch.

Z. Burda, A. Krywicki, O.C. Martin and M. Zagorski, PNAS 2011

The binding is treated within thermodynamics: a transcription factor's probability of being bound depends on the binding energy, calculated from the mismatch d_{ij} between the associated DNA and TF strings.

If one has a *single TF molecule*, the probability of it being bound is :

$$W_{ij} = e^{-\varepsilon d_{ij}} / Z$$

This is a measure of the interaction strength between TF *j* and the target gene *i*. Following Gerland et al. (2002), the formula can be generalized to the case where there are multiple TF molecules competing for the same binding site: the probability P_{ij} that one of the n_j molecules of TF *j* is bound to site *i* is like the Fermi function:

$$P_{ij} = \frac{W_{ij}\, n_j}{1 + W_{ij}\, n_j}$$

Given the probabilities of finding a binding site in the regulatory region of gene *i* occupied, we still have to model how this will affect the expression level of this gene. Transcriptional dynamics remains incompletely understood, but there has been a lot of modeling.

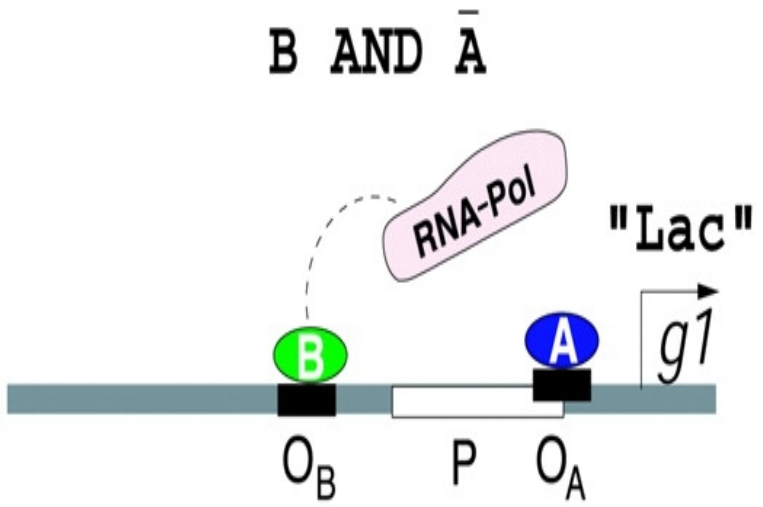
Transcriptional dynamics: combinatorial control logic

Some possible gene responses (ON or OFF) according to the specific activation patterns of two TFs, A and B, as denoted by their cellular concentrations (high or low).

a

A	B	"Lac" g1	AND g2	OR g3	NAND g4	XOR g5	EQ g6
low	low	OFF	OFF	OFF	ON	OFF	ON
high	low	OFF	OFF	ON	ON	ON	OFF
low	high	ON	OFF	ON	ON	ON	OFF
high	high	OFF	ON	ON	OFF	OFF	ON

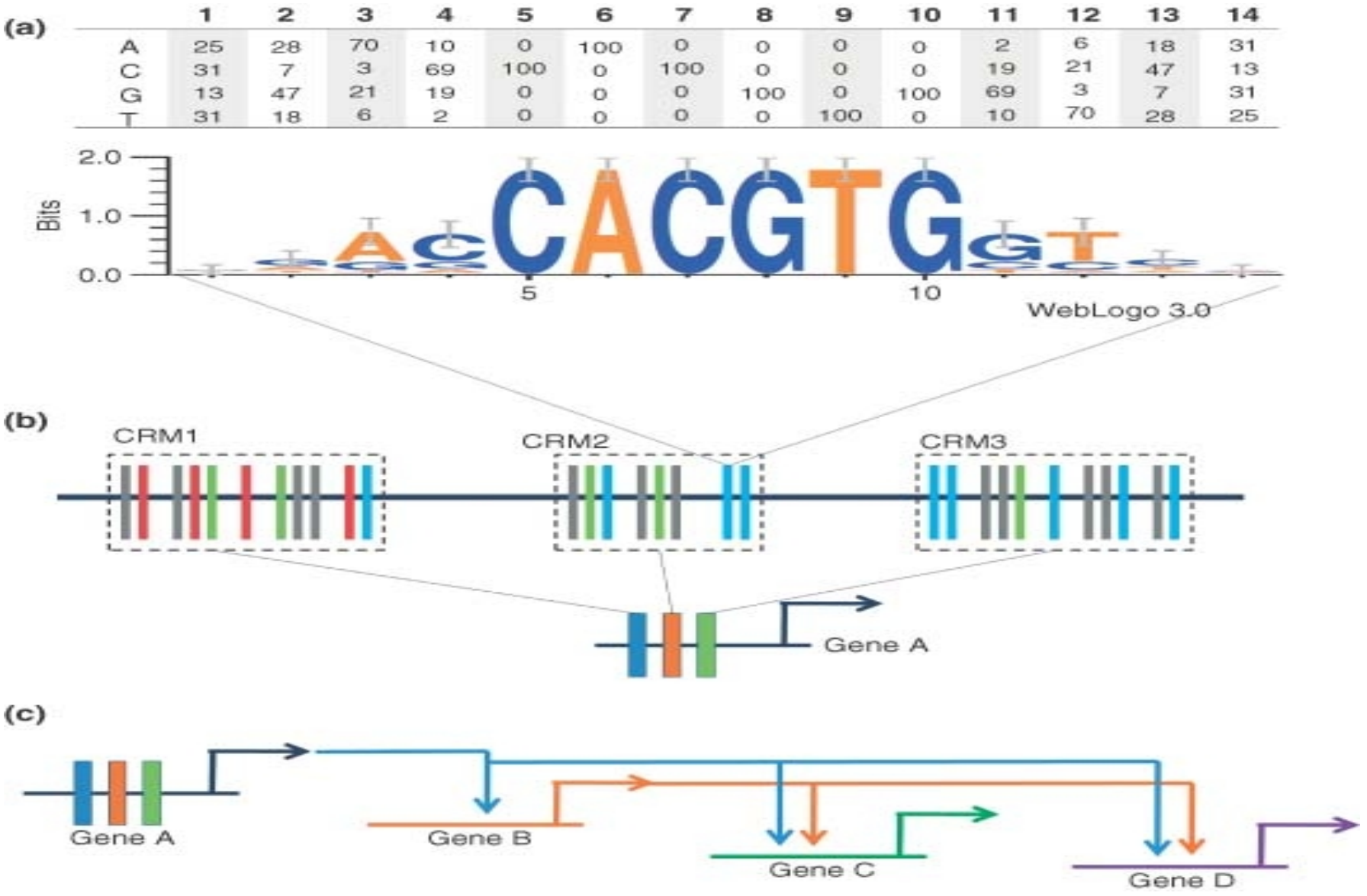
b



Buchler et al. PNAS 2003;100:5136-5141

Use of POCC = Probability of OCCupation within a region

Regulatory regions often have multiple CRM (cis-regulatory modules) each with multiple copies... Use biophysical binding model to estimate the POCC as a proxy for the genetic interaction strength.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	25	28	70	10	0	100	0	0	0	0	2	6	18	31
C	31	7	3	69	100	0	100	0	0	0	19	21	47	13
G	13	47	21	19	0	0	0	100	0	100	69	3	7	31
T	31	18	6	2	0	0	0	0	100	0	10	70	28	25

Regulatory logic in our molecular model

If no inhibitors are bound to the regulatory region:
use the "OR" logic → transcription is ON as soon as at least one activator is bound
The probability of being ON is then the POCC
→ mean expression level is given by POCC

If one or more inhibitors are bound to the regulatory region:
we consider that such transcription factors veto transcription
use the "OR" logic → transcription is OFF as soon as at least one inhibitor is bound
→ mean expression level is 0

We allow for no other cooperative effects.
If $P_{i,j}$ is the probability that some molecule of TF_j is bound to its site,
then our framework gives for the mean expression level of gene i,

$$S_i = [\, 1 - \Pi_j (1 - P_{i,j}) \,] \; \Pi_{j'} (1 - P_{i,j'})$$

j activators, j' inhibitors

Genotypes and phenotypes revisited

The matrix ***W*** represents the **genotype** of the network.

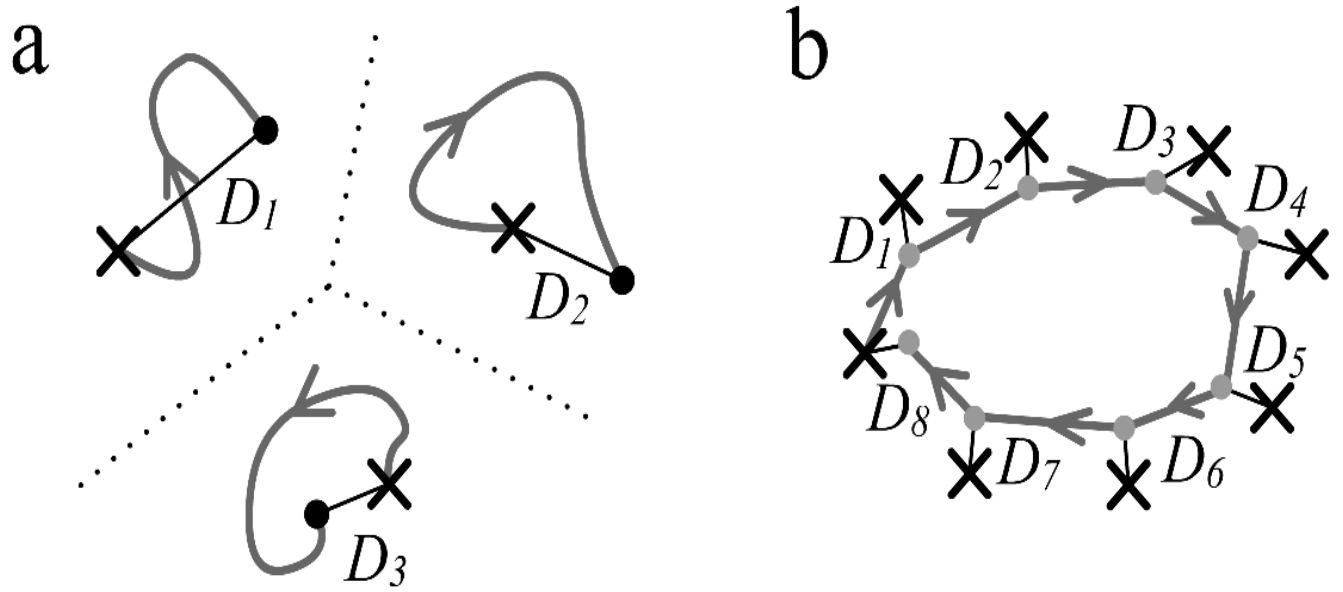
The **phenotype** is associated with the expression patterns produced by the transcriptional dynamics of this network. You can think of the phenotype as the “function” of the network.

We consider alternatively two kinds of “target” (desired) expression patterns:

(1) A set of stationary (fixed point) expression vectors. This is motivated by the different expression patterns in different tissues for instance in development.

(2) A cycle of expression vectors. This is motivated by the behavior of oscillators (circadian clocks and the cell cycle).

Genotypes and phenotypes revisited



Since our expression levels S_j are continuous, we have to allow for some tolerance: the actual expression vectors will not be exactly the target ones. We do this by simply having a fitness function which depends on the distance between the actual behavior and the target behavior. In practice, we use the fitness function

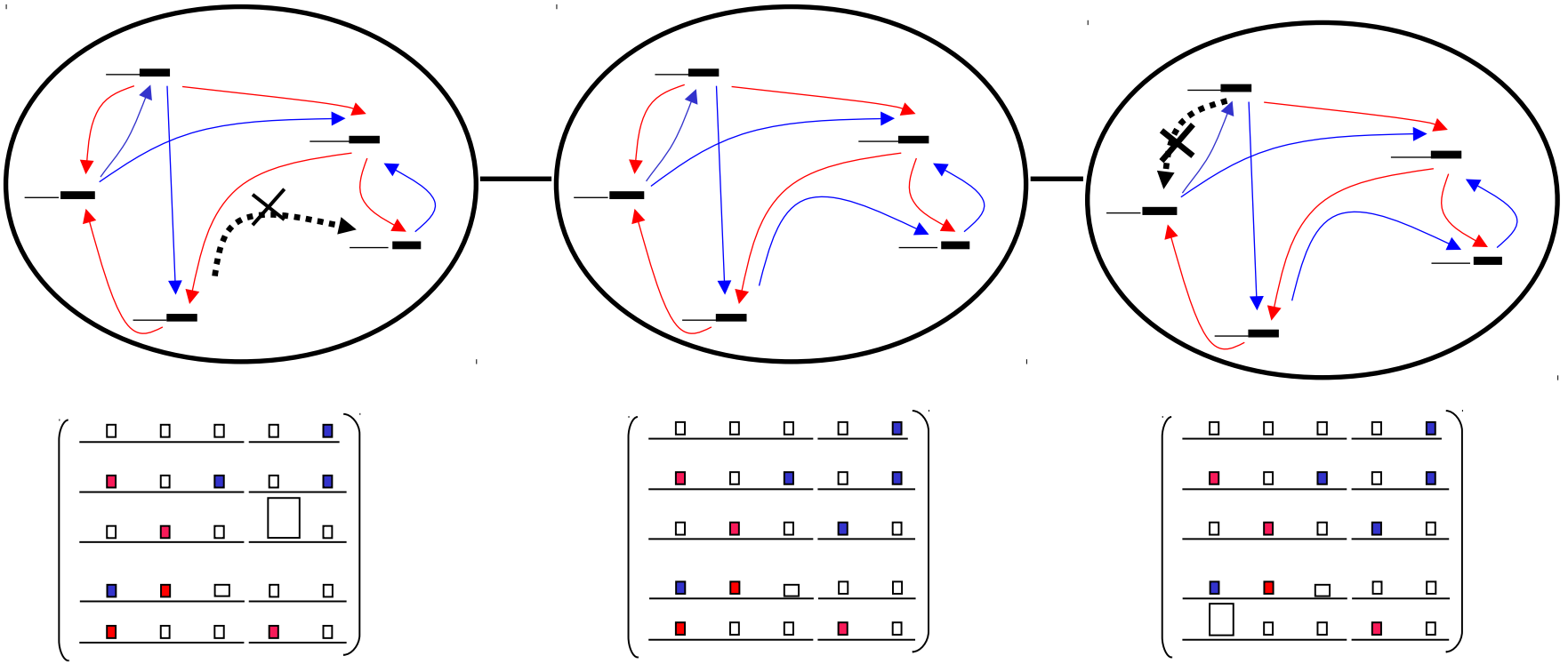
Fitness(phenotype) = exp[- f Dist(phenotype, target_phenotype)]

Computational aspects: using MCMC

The set of all genotypes is easily sampled. But we are only interested in "functional" networks, that is genotypes that have phenotypes close to the desired target. If one takes a random matrix $\boldsymbol{W}$ it almost surely will lead to a bad phenotype (low fitness). To sample genotypes having *good* phenotypes, we use Markov Chain Monte Carlo (MCMC).

Principle behind the MCMC: perform a biased random walk in the genotype space in such a way that the equilibrium measure of the walk is the desired one. Here we sample genotypes with a probability proportional to their fitness: we have a fitness landscape. A genotype producing expression patterns far from the target will simply not arise in the sampling.

In practice, our MCMC sampling procedure necessitates the use of steps to produce the walk, to go from one genotype to another. For efficiency reasons, the steps have to be small and so generally are taken to be point mutations.

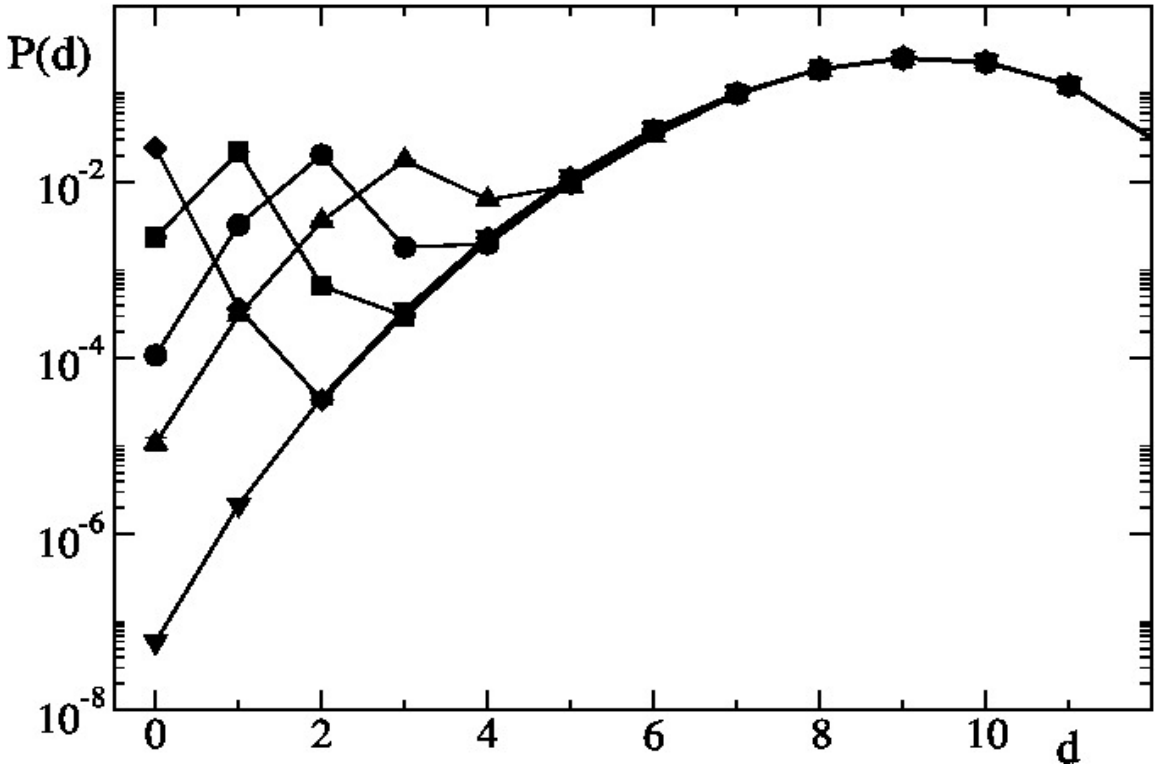


For our model, we perform mutations on the *character strings* of the binding sites in the regulatory regions.

In this improved model, are interactions sparse?

The weights (binding probabilities) W_{ij} are never zero so **one cannot talk of networks with sparse interactions**. Nevertheless, some interactions must be weak and others strong. In practice, the distribution of the W_{ij} is bimodal: most interactions are weak and are distributed as if there were no viability constraints, but some interactions are strong.

Effect of increasing the parameter *n* in the Fermi function



d = mismatch

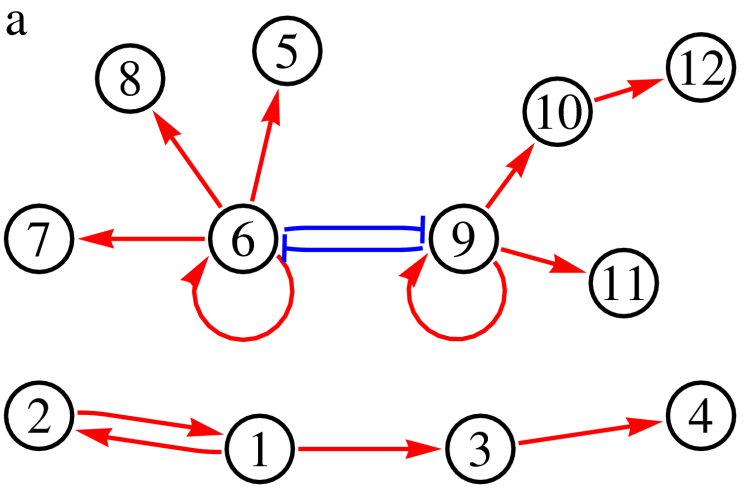
In this improved model, are interactions sparse?

To avoid an arbitrary distinction between weak and strong, we focus on the *functional* (biological) importance of an interaction. An interaction is called “essential” if its removal makes the network non functional. In practice, the distinction follows rather closely the the separation into weak and strong interactions.

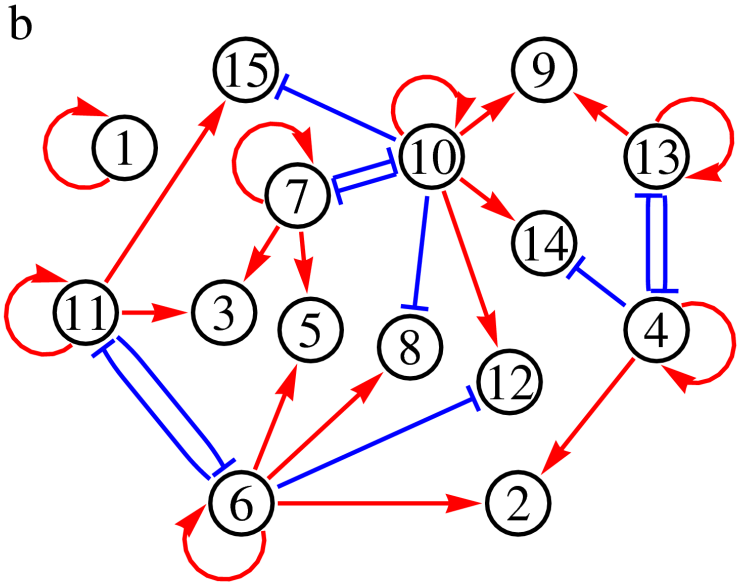
Rather than consider the full genotype, we can focus on the list of interactions that are essential, hereafter referred to as the “essential network”. This coarse-graining, which does not include continuous variables, provides a far more intuitive grasp of what the network is doing.

Essential networks are sparse !

Network having 2 given steady state expression patterns



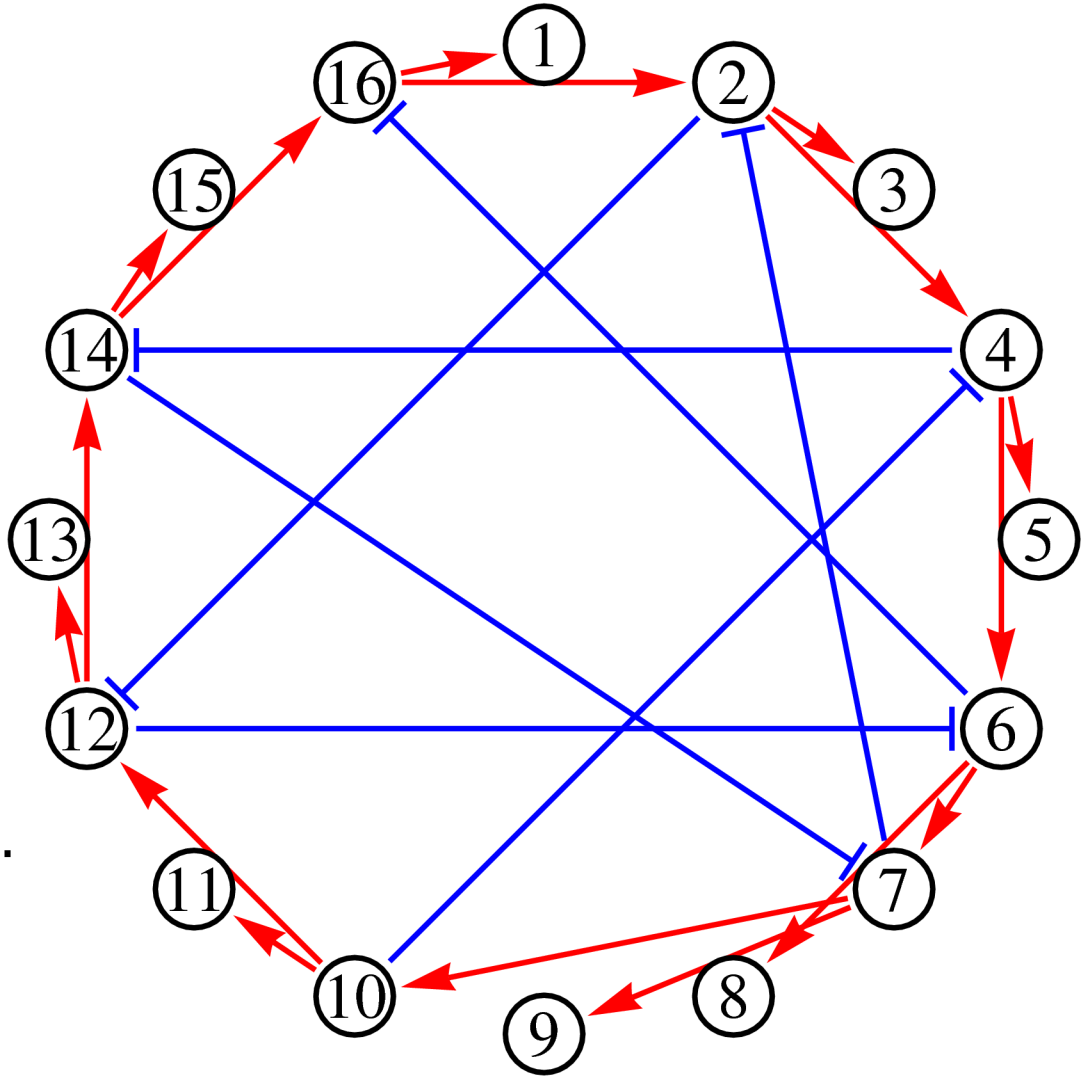
Network having 4 given steady state expression patterns



Z. Burda, A. Krywicki, O.C. Martin and M. Zagorski, PNAS 2011

Network having a given cyclic expression pattern:

0 1 1 1 1 1 1 0 0 0 0 …
0 0 0 1 1 1 1 1 1 0 0 0 …
0 0 0 0 0 1 1 1 1 1 1 0 0 …
0 0 0 0 0 0 0 1 1 1 1 1 1 0 0 …



Z. Burda, A. Krywicki, O.C. Martin and M. Zagorski, PNAS 2011

Justification for sparseness

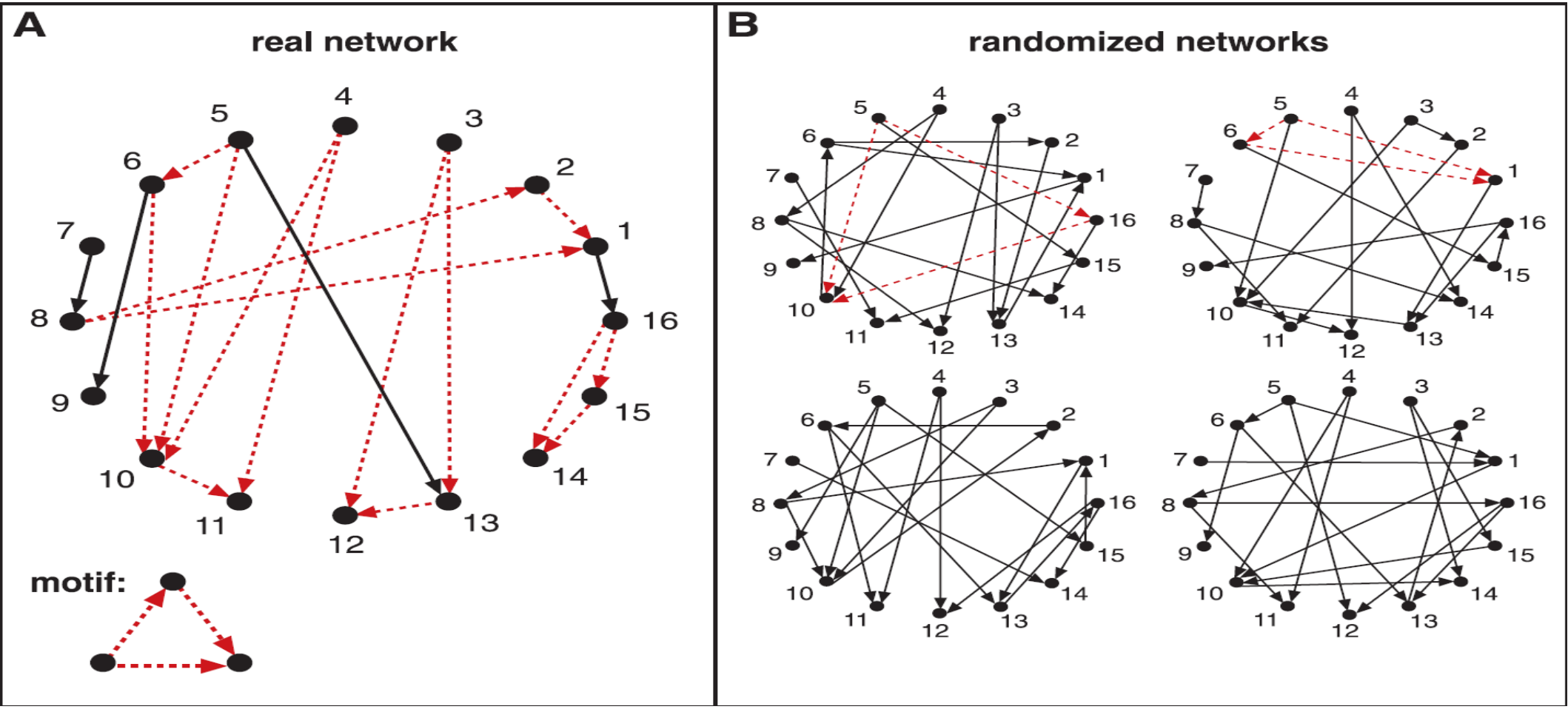
In the absence of the functional constraints, the weights $\boldsymbol{W_{ij}}$ have a known distribution, leading to low binding probabilities. For a TF to bind, its mismatch must be low but this is a **rare** event.

Entropy is favored when most mismatches are random, thus have large values. Networks thus tend to be *parsimonious* in low mismatches, which leads to few large weights $\boldsymbol{W_{ij}}$. (These are the interactions that are essential.) With our encoding of the weights via molecular genotypes, there are far more sparse networks than dense ones!

In a simplified version of the model, we showed that one has a condensation phase transition: the functional constraints are implemented for each gene by one incoming interaction.
In the more realistic cases, each gene is controlled by just a small number of (essential) interactions, so these networks have low in-degree. A broad distribution of the *out-degree* is obtained if the regulatory model is used in an evolving population.

Q2: Why do biological networks have motifs?

Comparing real gene regulatory networks to randomized networks reveals the existence of repeated (over-represented) motifs as shown by Uri Alon's group. Are these functional modules? Are they just the trace of evolutionary histories such as successive whole genome duplications?



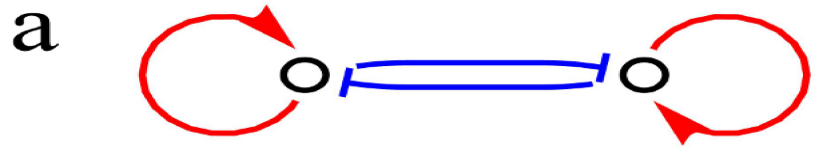
What happens in our improved model?

In our model framework, motifs arise, we find them in our MCMC sampled genotypes. Thus it is not necessary to appeal to evolutionary chance or whole genome duplications as the cause of excess occurrences of subgraph topologies. Instead, one has something similar to evolutionary convergence: because of the functional constraints, specific genetic architectures emerge.

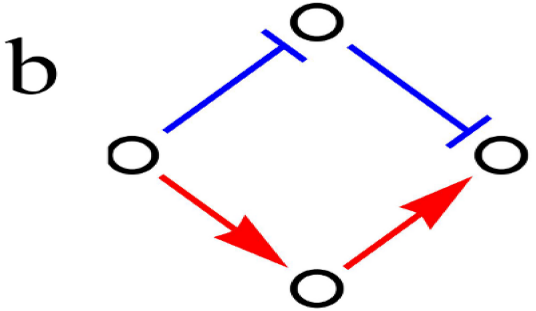
Some intuitive arguments from a design perspective:

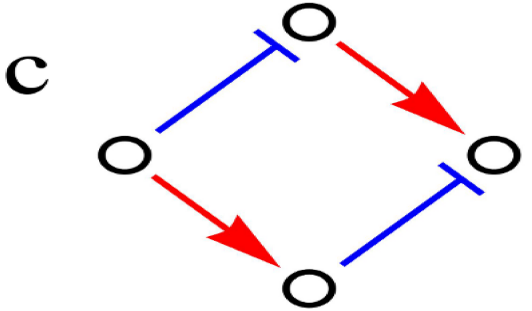
- In the case where one wants several steady state expression patterns: one can have two master genes that mutually inhibit one another. These form a switch that can drive other genes to the desired target levels.
- In the case of the target cyclic behavior, we took a shifting pattern on a ring. Genes then have to activate genes ahead of them and inhibit genes behind them.

Motifs found by analyzing the MCMC sampling

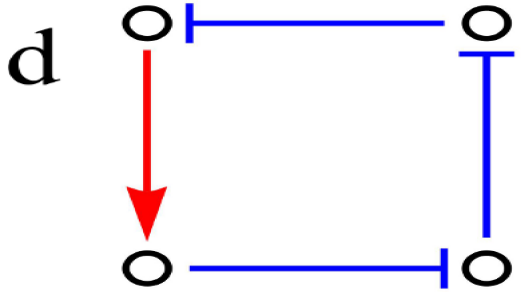


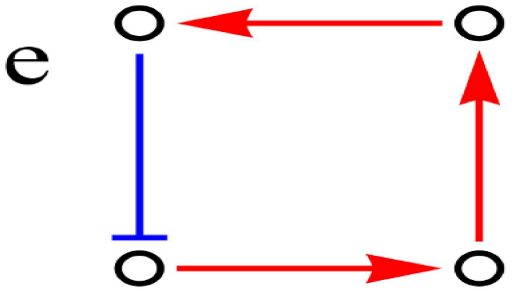
a: main motif for fixed point phenotypes

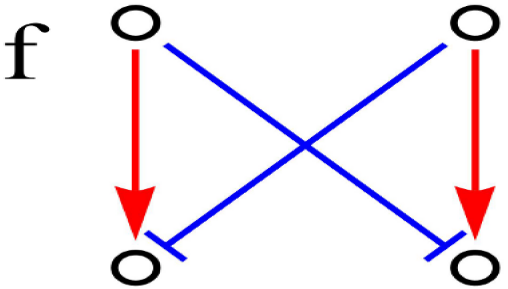




b,c,d,e,f: main motifs having loops for the cyclic phenotypes





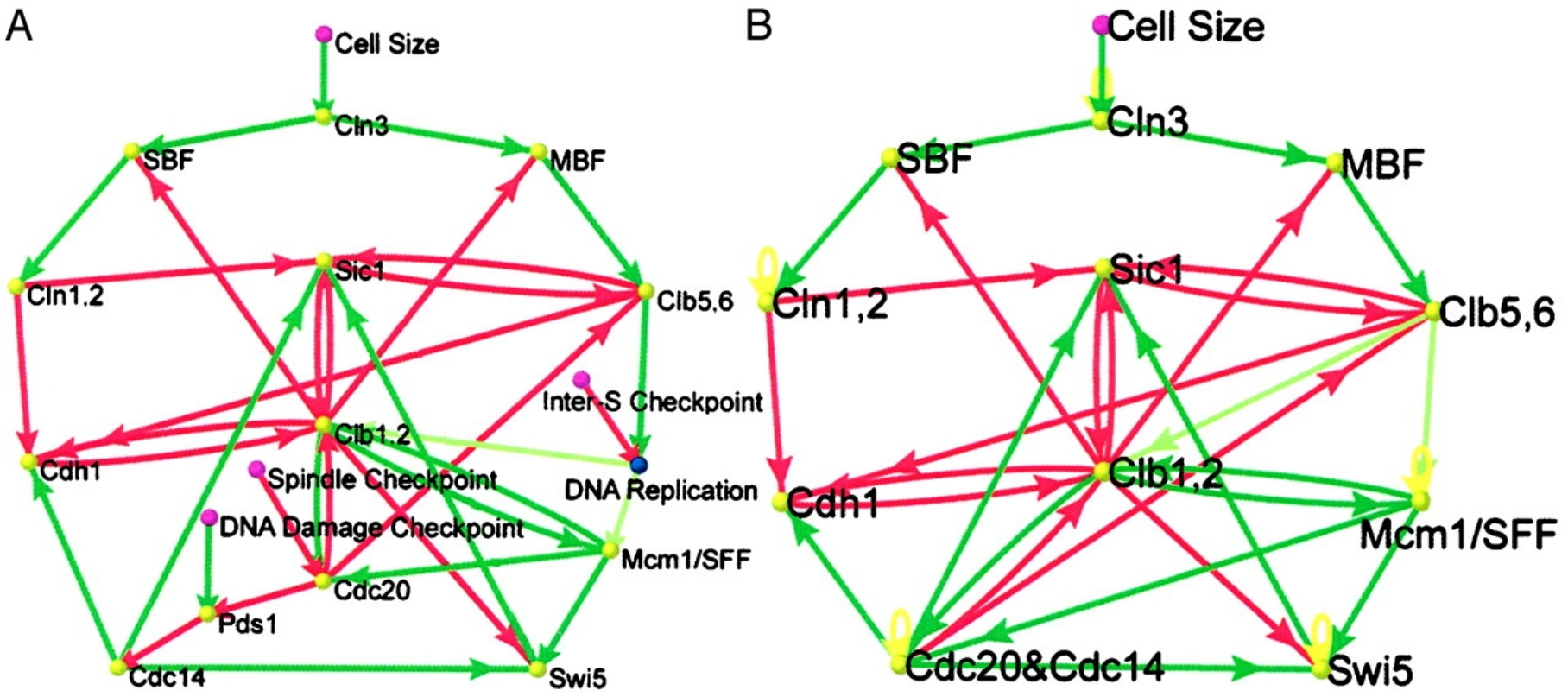


Z. Burda, A. Krywicki, O.C. Martin and M. Zagorski, PNAS 2011

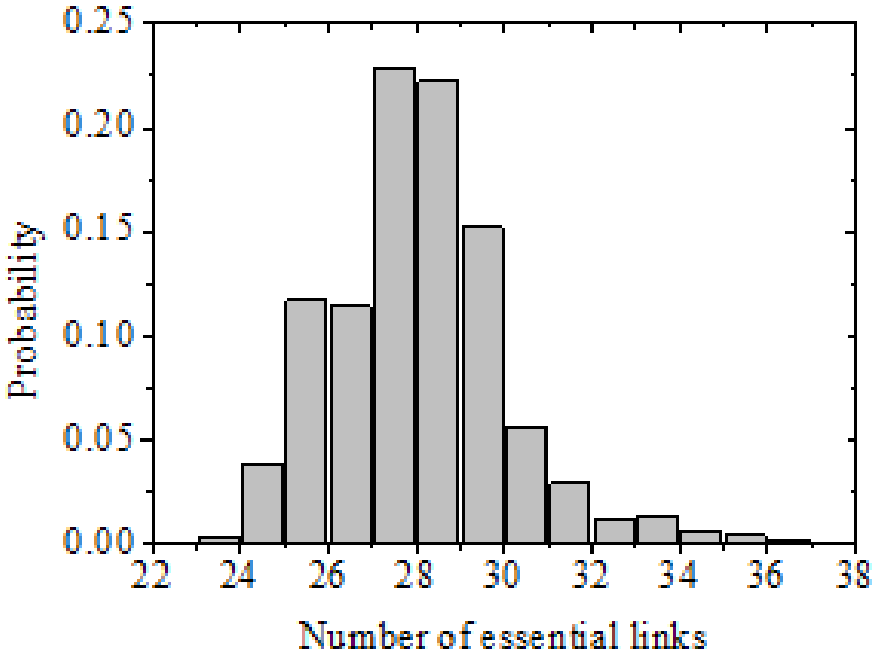
Extensions to a (Boolean) cell cycle model

Li, Long, Lu, Ouyang, and Tang
The yeast cell-cycle network is robustly designed,
PNAS 2004

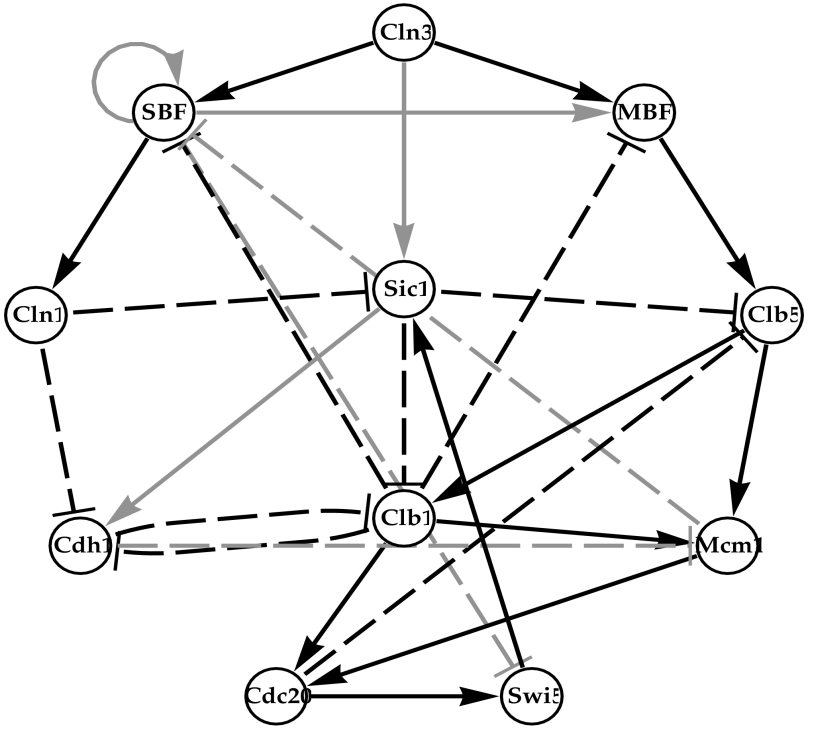
Lau, Ganguli, and Tang
Function constrains network architecture and dynamics: A case study on the yeast cell cycle Boolean network,
Phys Rev E 2007



Behavior obtained in our framework

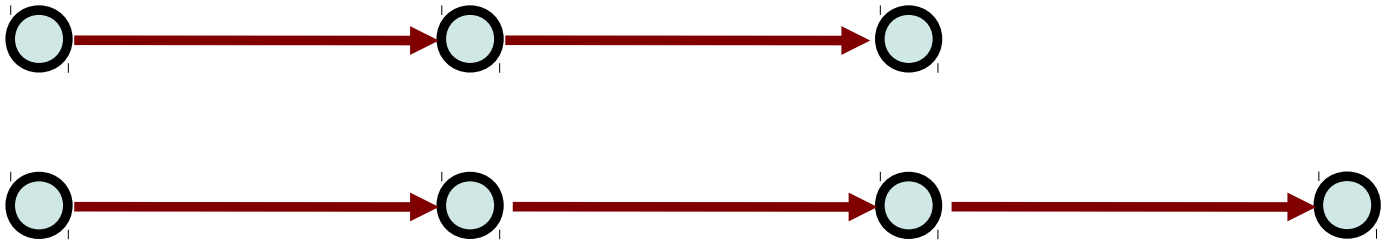


Distribution of number of interactions



Most frequent network

Motifs:
loopless feed-forward



Conclusions

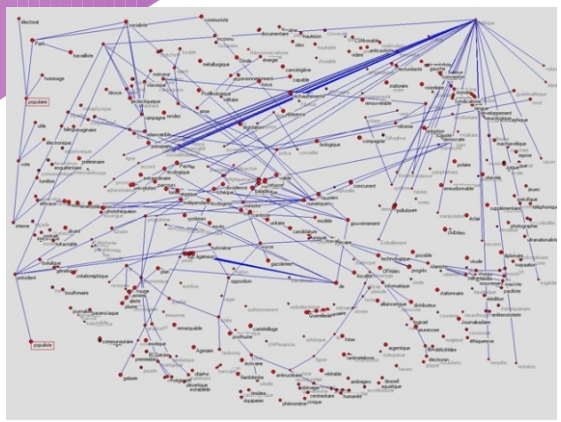
Imposing a function (phenotype) on a system is generally a severe constraint; associated genotypes are *rare*.

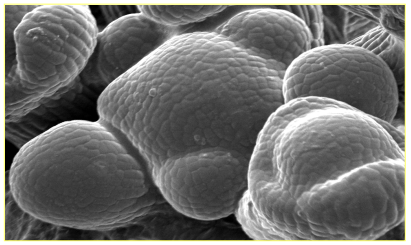
Although functional genotypes represent only a tiny fraction of all genotypes, they are typically astronomically many such genotypes forming a connected, and highly diverse set. This probably follows from the high degeneracy of the genotype to phenotype map and the existence of many near neutral mutations. Genotype networks associated with different phenotypes are interwoven, be-it for RNA molecules (Schuster et al., Sumedha et al.) or gene regulatory networks (Ciliberti et al.).

To improve the modeling of transcriptional regulation, we included the mechanistic nature of *genetic* interactions $\boldsymbol{W_{ij}}$, which builds-in the difficulty of maintaining large interaction strengths; this naturally leads to *sparse* (essential) interactions. The functional constraints then make certain genetic architectures more likely, revealed here by *emergent motifs*.

Analogous studies of *metabolic* networks (A. Samal et al.) show the emergence of modularity, large scale structural properties, etc. when imposing constraints on the metabolic phenotype.

Take home message…





Hopefully not an infinite loop...