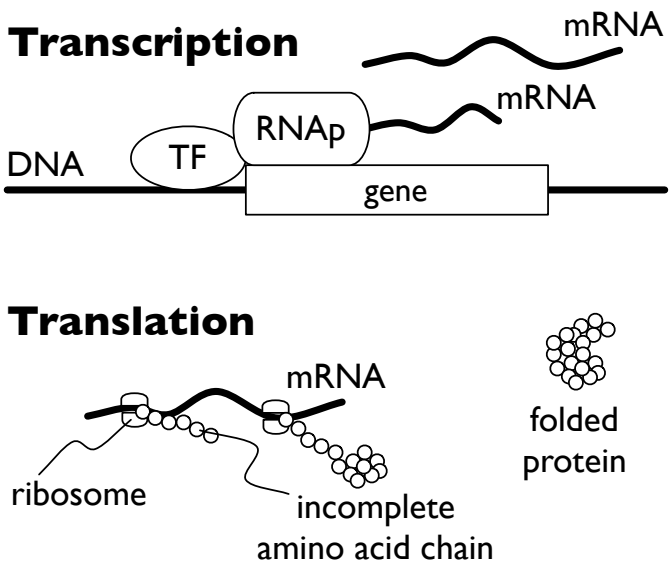
What functions do gene expression levels represent?

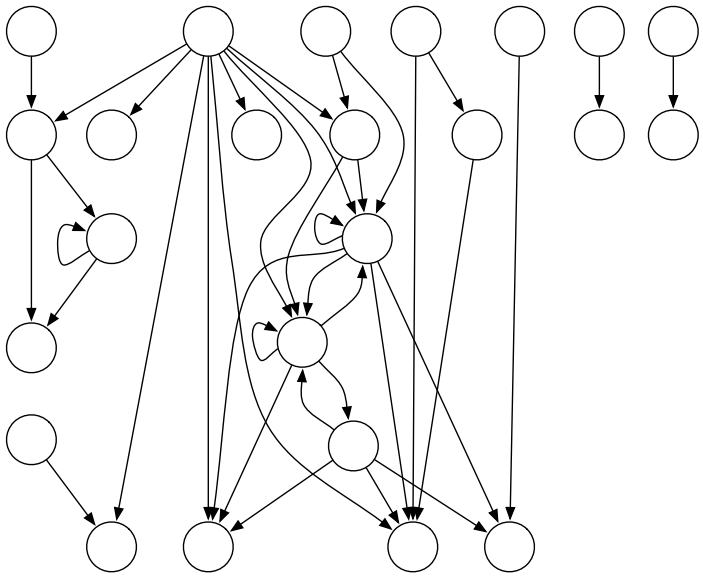
Elaine Angelino

March 17, 2014

Previous talk

- Evolvability of sparse linear functions
- Based on the observation that gene expression levels are regulated by transcription factors networks that have low depth and fan-in





Current goals

Identify a framework describing:

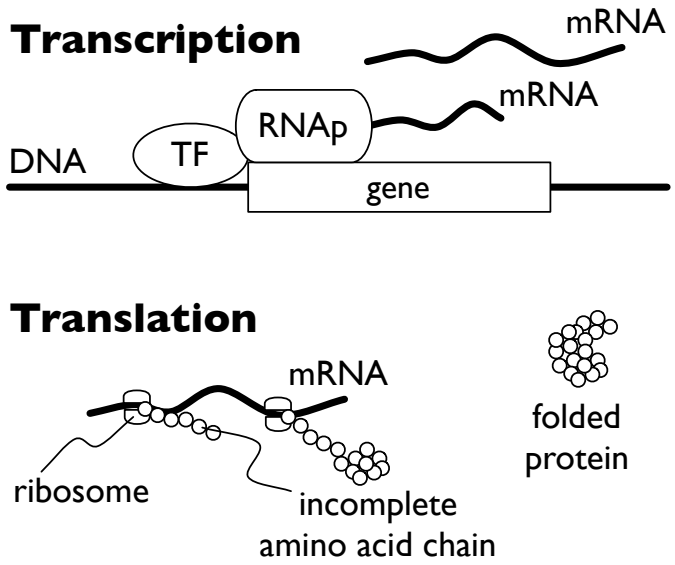
- More realistic models of gene expression
- Parameters corresponding to knobs tuned by evolution

Outline

- Review some gene regulation mechanisms
- Summarize a known framework (Markov chains, dynamical systems)
- Study the functional form of solutions (Kirchoff's Matrix Tree Theorem)
- Consider generalizations and implications for evolvability

Gene expression basics

- DNA is transcribed into mRNA, which is translated into protein
- Gene expression level is proportional to the fraction of promoters bound by RNA polymerase (RNAp)

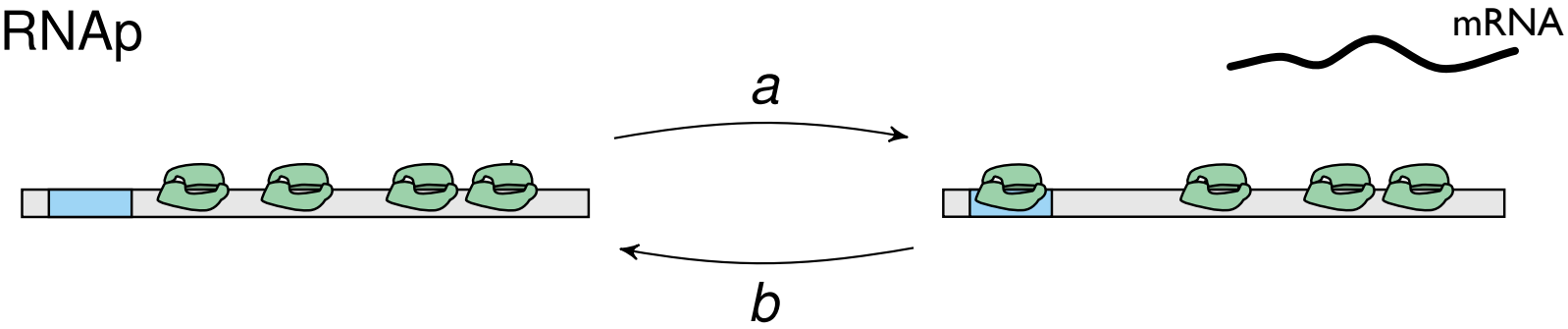


- Physical state of the promoter can make RNAp binding (un)favorable, equivalent to changing the “effective” amount of RNAp

Simple 2-state system

= promoter

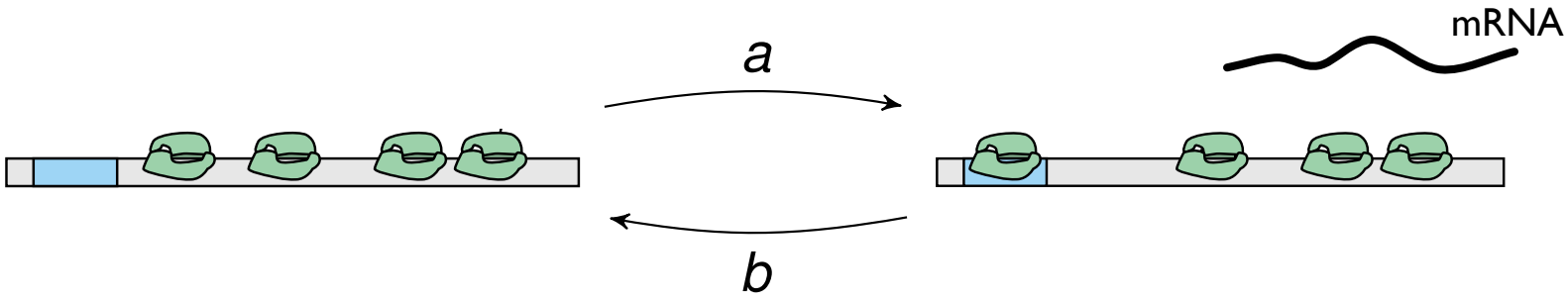
= RNAp



- Markov chain with 2 states: promoter is empty or bound by RNAp
- a and b are rate constants
- Let x = probability RNAp is not bound
  y = probability RNAp is bound $\Rightarrow$ Gene expression level $\propto y$
- What do the rate constants a, b depend on?
- What kind of functions are x, y in terms of a, b?

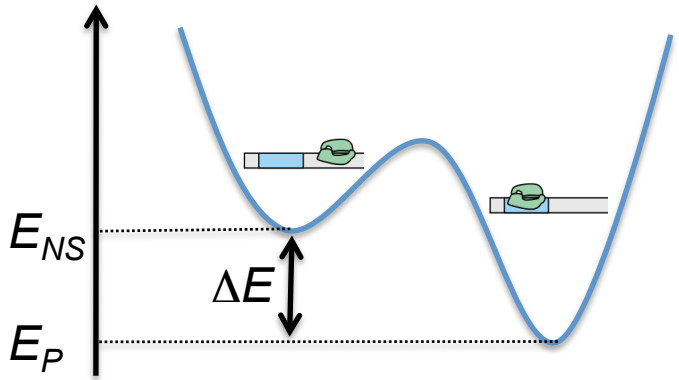
L. Bintu, *et al.* Transcriptional Regulation by the Numbers 1: Models, Current Opinion in Genetics & Development 15(2):116-124 (2005)

Simple 2-state system



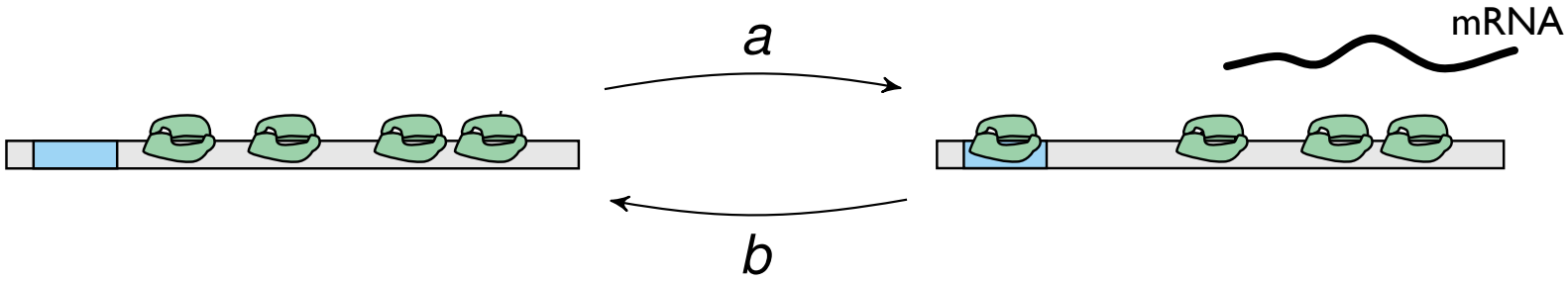
$P =$ number of RNAp molecules
$N =$ number of non-specific (NS) sites where RNAp can bind to DNA

$$\text{probability of state S} \propto \exp(-E_S/k_BT)$$



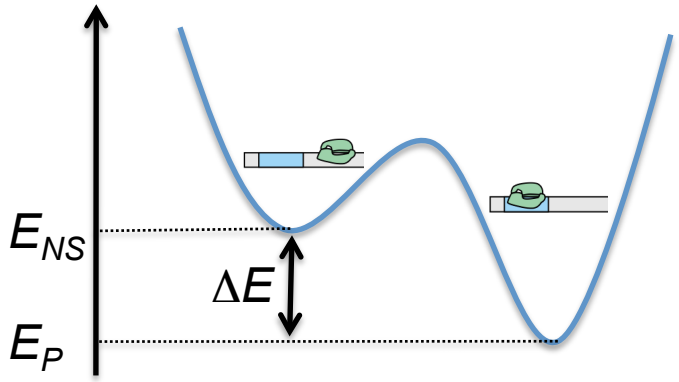
L. Bintu, *et al.* Transcriptional Regulation by the Numbers 1: Models, Current Opinion in Genetics & Development 15(2):116-124 (2005)

Simple 2-state system



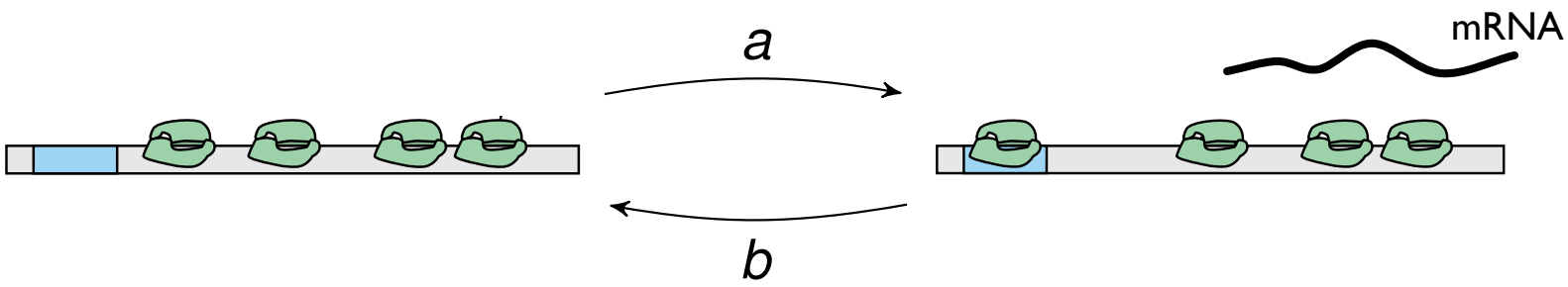
P = number of RNAp molecules
N = number of non-specific (NS) sites where RNAp can bind to DNA

$$Z = \frac{N!}{P!(N-P)!} e^{-PE_{NS}/k_BT} + \frac{N!}{(P-1)!(N-P-1)!} e^{-(P-1)E_{NS}/k_BT} e^{-E_P/k_BT}$$



L. Bintu, *et al.* Transcriptional Regulation by the Numbers 1: Models, Current Opinion in Genetics & Development 15(2):116-124 (2005)

Simple 2-state system: Equilibrium probabilities



N = number of non-specific (NS) sites where RNAp can bind to DNA
P = number of RNAp molecules

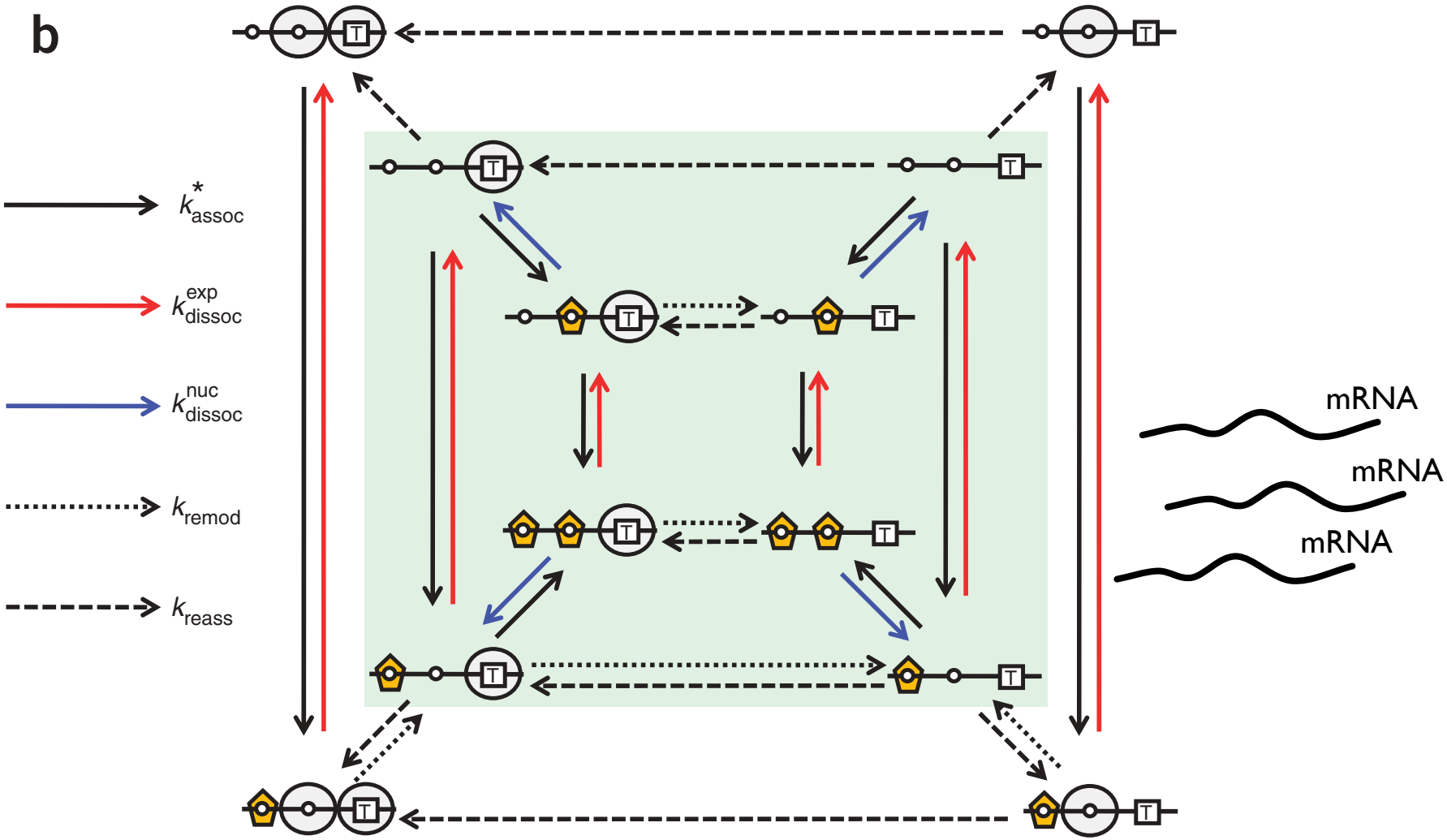
$$Z = \underbrace{\frac{N!}{P!(N-P)!} e^{-PE_{NS}/k_BT}}_{X} + \underbrace{\frac{N!}{(P-1)!(N-P-1)!} e^{-(P-1)E_{NS}/k_BT} e^{-E_P/k_BT}}_{Y}$$

$$\text{Probability RNAp bound to promoter} = \frac{Y}{X+Y} = \frac{1}{1+\frac{X}{Y}}$$

$$\approx \frac{1}{1+\frac{N}{P}e^{\Delta E/k_BT}} = \frac{1}{1+\frac{1}{[\text{RNAp}]}\frac{K_S}{K_{NS}}}$$

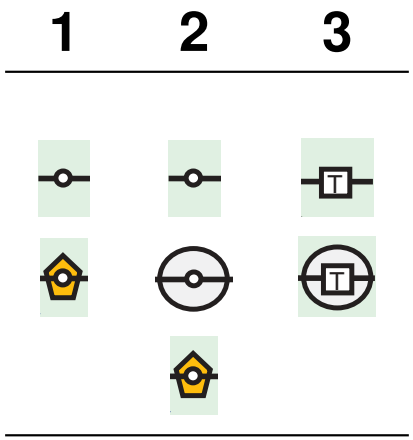
L. Bintu, *et al.* Transcriptional Regulation by the Numbers 1: Models, Current Opinion in Genetics & Development 15(2):116-124 (2005)

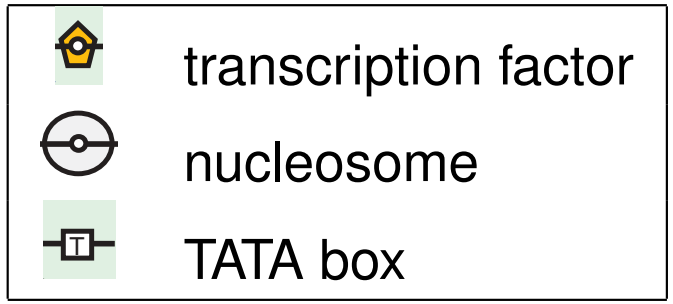
(Part of) a real promoter system . . .



H.D. Kim and E.K. O'Shea, A quantitative model of transcription factor activated gene expression, Nature Structural Molecular Biology, 15(11), 1192-1198 (2008).

Promoters can have many states

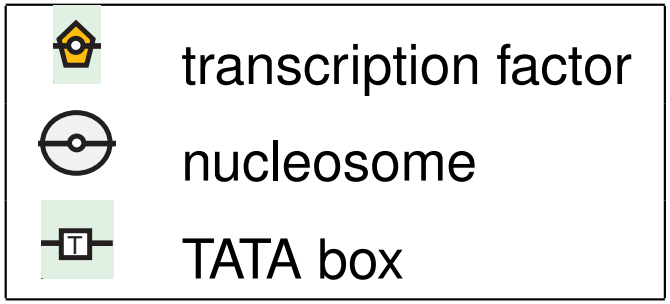


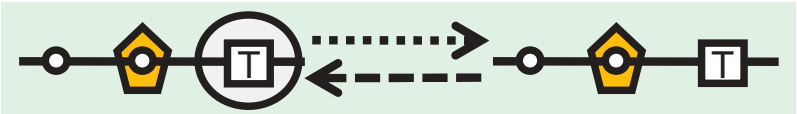


$2 \times 3 \times 2 = 12$ states

- Example: promoter composed of 3 physical regions
- The regions can be empty or occupied
- RNAp can bind only when the 3rd region is (6/12 states)

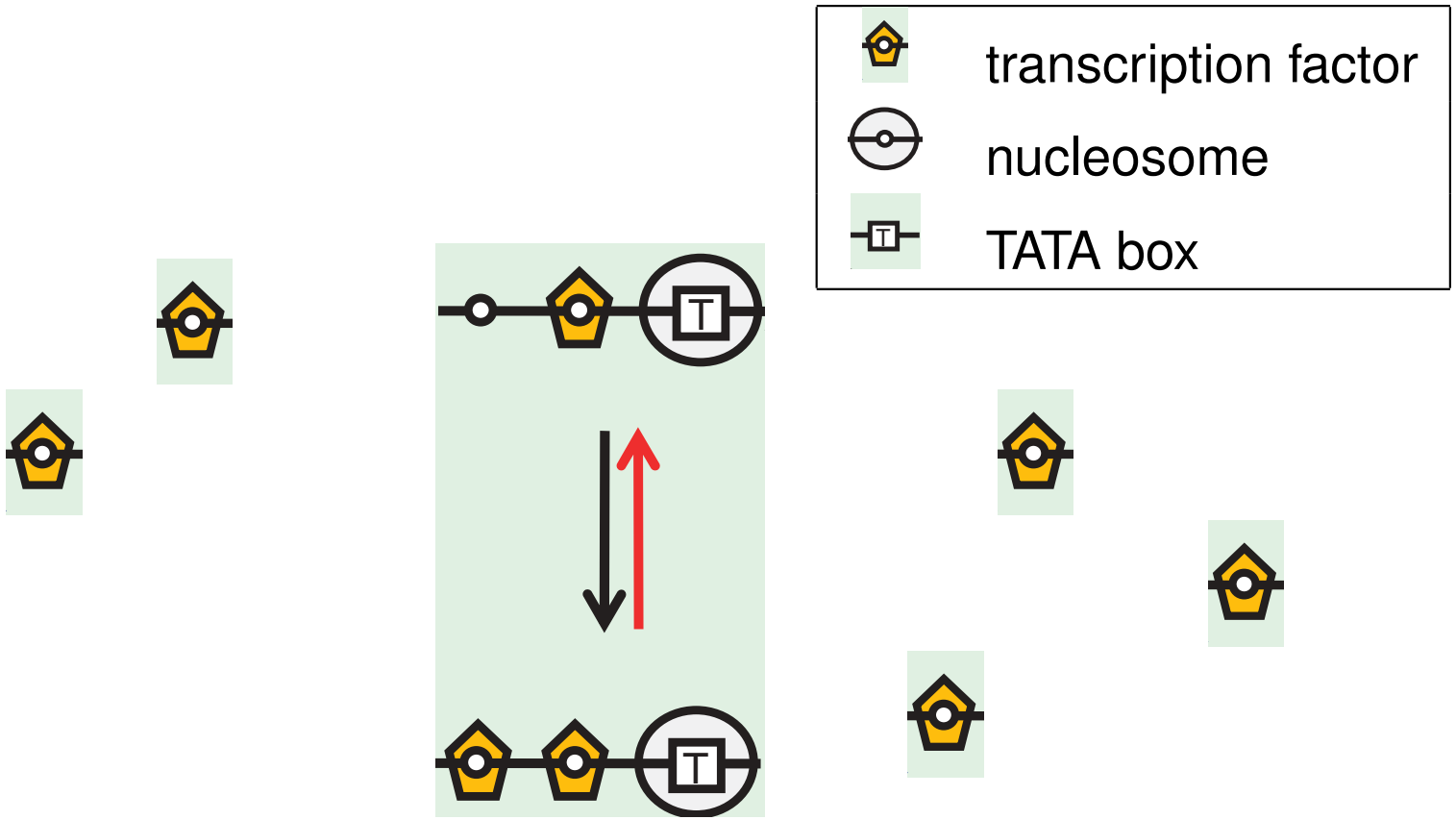
Transitions change occupancy at 1 promoter region





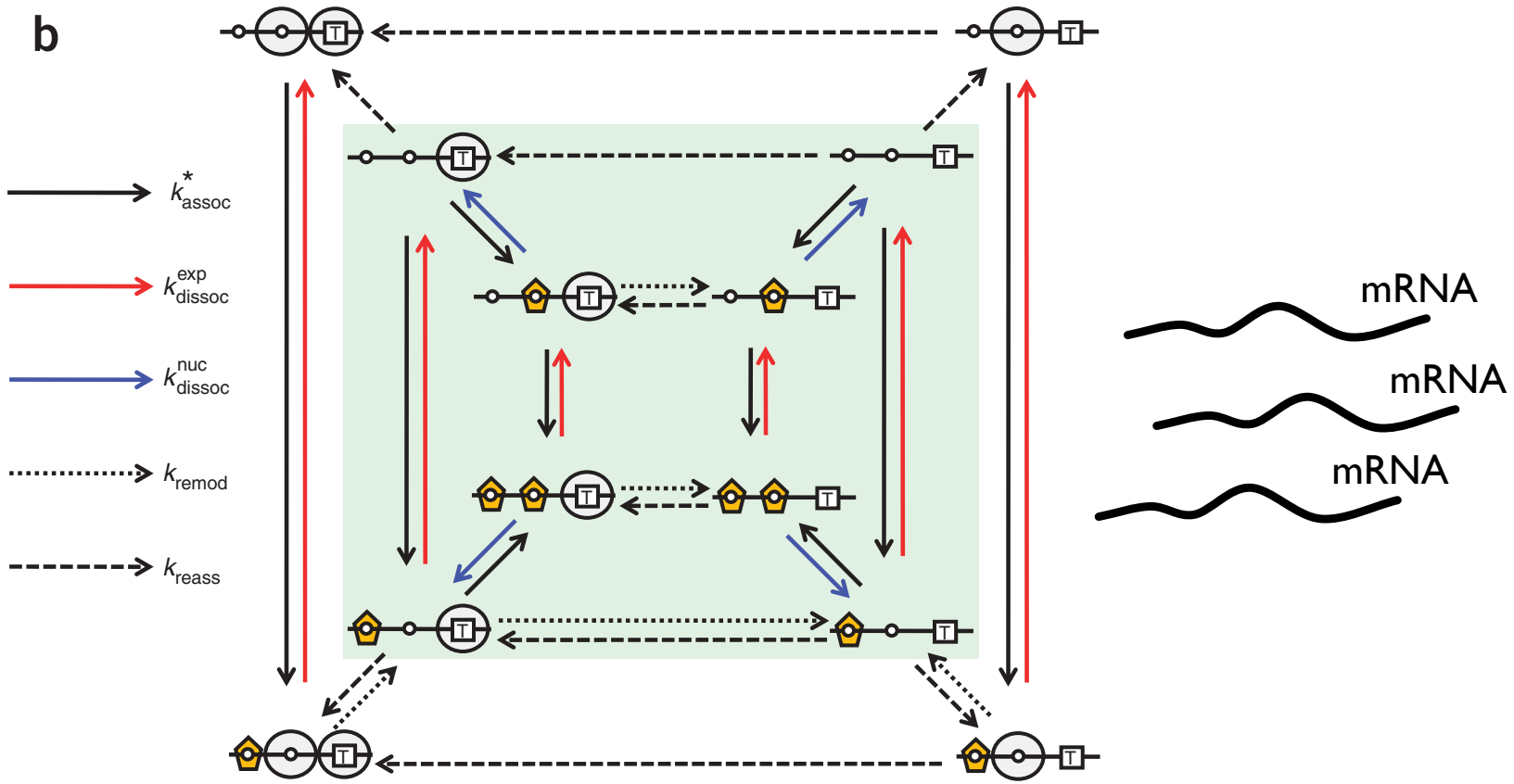
- This transition “adds” or “removes” at the 3rd promoter region
- These rates are physical constants

Some transitions are due to binding / unbinding



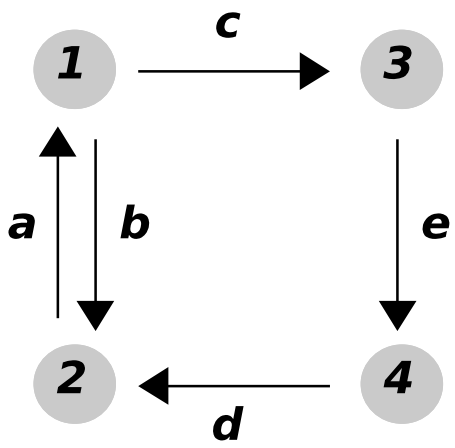
- Rate of binding $\propto$ [free]
- Assumption: [free] $\approx$ constant w.r.t. promoter system dynamics

(Part of) a real promoter system . . .



- This graph is sparse and strongly connected
- Some transitions are irreversible

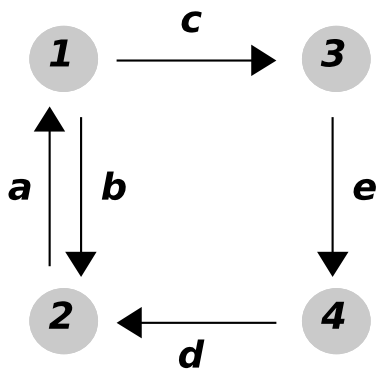
Representation: Graph $G(V, E)$



- Vertices correspond to states
- Edges are directed and labeled by non-negative transition rates
- No self-edges

I. Mirzaev, J. Gunawardena, “Laplacian dynamics on general graphs”, Bull Math Biol, 75:2118-49 2013.

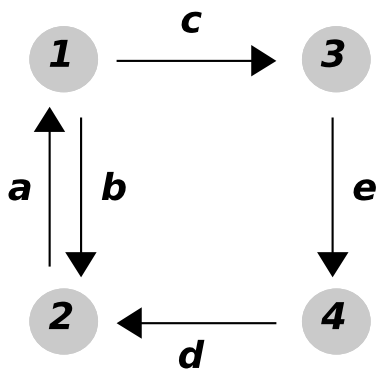
Continuous-time Markov chain



- $x_i =$ probability of state i
- $\frac{d\vec{x}}{dt}$ describes the rates at which the state probabilities change

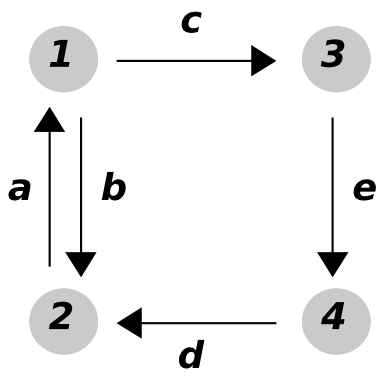
$$\begin{aligned}
\frac{dx_1}{dt} &= ax_2 - (b+c)x_1 \\
\frac{dx_2}{dt} &= bx_1 + dx_4 - ax_2 \\
\frac{dx_3}{dt} &= cx_1 - ex_3 \\
\frac{dx_2}{dt} &= ex_3 - dx_4
\end{aligned}$$

Continuous-time Markov chain



$$\frac{d\vec{x}}{dt} = \begin{pmatrix} -(b+c) & a & 0 & 0 \\ b & -a & 0 & d \\ c & 0 & -e & 0 \\ 0 & 0 & e & -d \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} \equiv \mathcal{L}\vec{x}$$

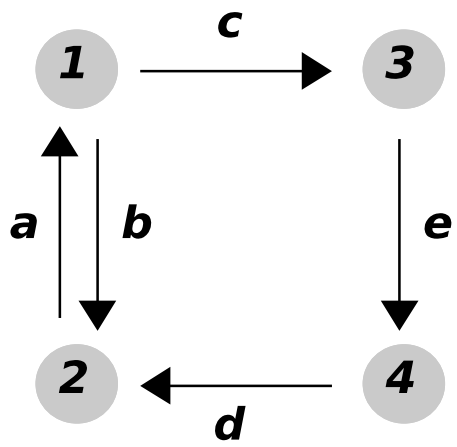
Continuous-time Markov chain



$$\frac{d\vec{x}}{dt} = \begin{pmatrix} -(b+c) & a & 0 & 0 \\ b & -a & 0 & d \\ c & 0 & -e & 0 \\ 0 & 0 & e & -d \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} \equiv \mathcal{L}\vec{x}$$

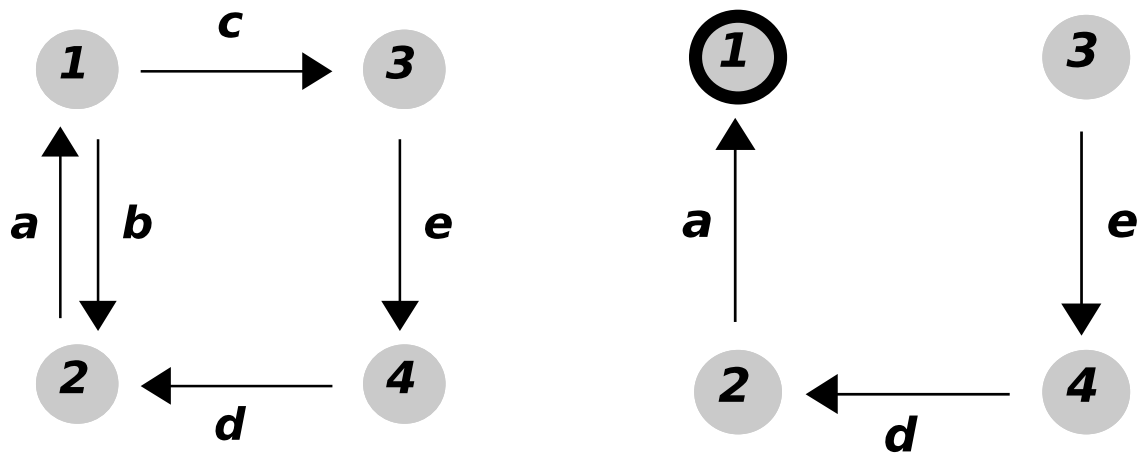
- $\mathcal{L}(G) =$ Laplacian of graph G
- Columns of $\mathcal{L}$ sum to zero
- Conservation of mass: $x_1(t) + \cdots + x_n(t) = c$

Focus on strongly connected graphs



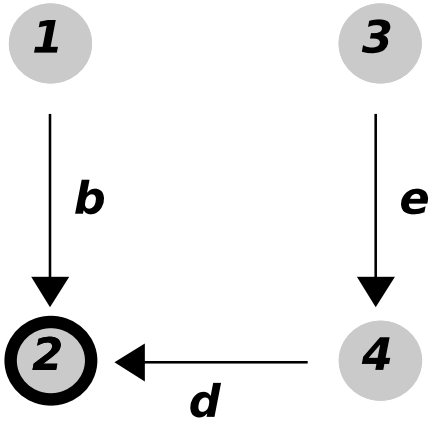
- Physically meaningful setting
- Results can be generalized to graphs that are not strongly connected
- Strongly connected: $\geq$ 1 directed path between any 2 vertices
  $\geq$ 1 spanning tree rooted at each vertex

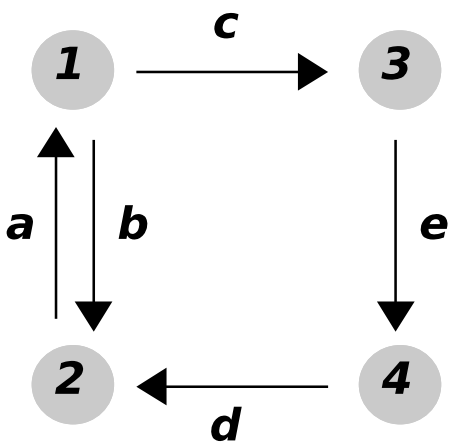
Spanning trees rooted at vertex 1

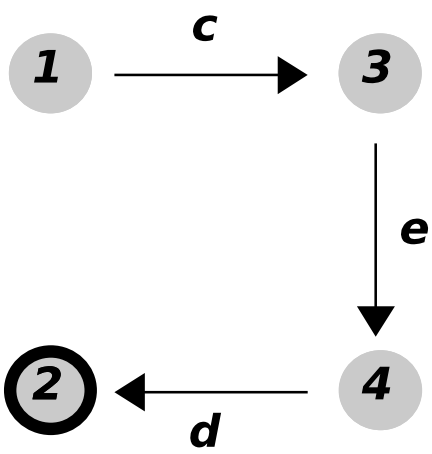


- A tree is a graph with no cycles
- A spanning tree has all n vertices $\Rightarrow n - 1$ edges
- Rooted at 1: All paths lead to 1

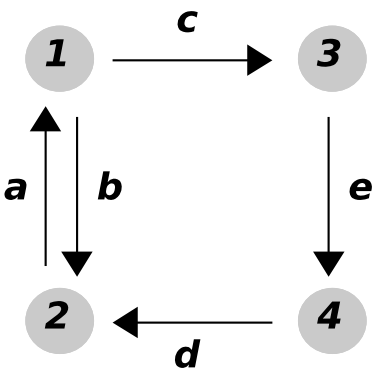
Spanning trees rooted at vertex 2

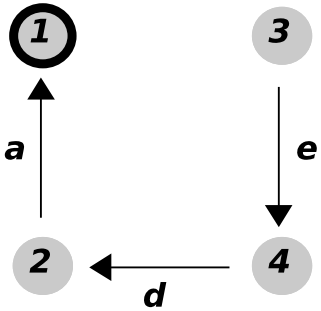


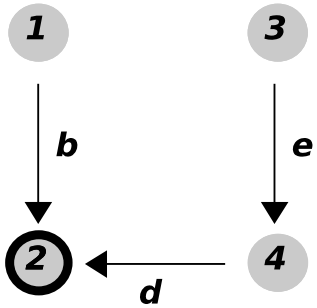


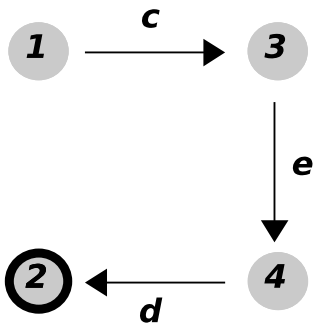


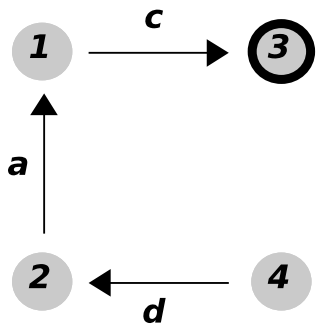
All spanning trees

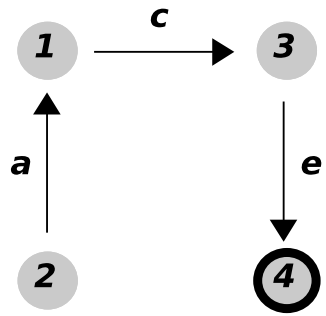












Solve for steady state when G is strongly connected

- Steady state: $\vec{x}(t)$ converges to a stationary distribution $\vec{x}^*$
- Want to solve for $\vec{x}^*$ s.t. $\frac{d\vec{x}}{dt} = \mathcal{L}\vec{x}^* = 0$
- Strongly connected $\Rightarrow$ rank of $\mathcal{L}$ is exactly $n-1$
- Recall: columns of $\mathcal{L}$ sum to zero
  $\Rightarrow 1^T\mathcal{L} = 0 \Rightarrow 1$ is a left eigenvector with eigenvalue 0
  $\Rightarrow \exists$ corresponding right eigenvector $\vec{\rho}$ s.t. $\mathcal{L}\vec{\rho} = 0$
  $\Rightarrow$ Solution $\vec{x}^* \propto \vec{\rho}$

What is the functional form of $\vec{x}^*$?

- $\mathcal{L}$ is not full-rank $\Rightarrow \det(\mathcal{L}) = 0$
- Determinant expansion $\Rightarrow \forall j,\ \det(\mathcal{L}) = \sum_k (-1)^{j+k} \ell_{jk} \det(\mathcal{L}_{j,k}) = 0$

  $\det(\mathcal{L}_{j,k}) = (j,k)$-minor of $\mathcal{L}$ = determinant of submatrix made by deleting the j-th row and k-th col

What is the functional form of $\vec{x}^*$?

- $\mathcal{L}$ is not full-rank $\Rightarrow \det(\mathcal{L}) = 0$
- Determinant expansion $\Rightarrow \forall j,\ \det(\mathcal{L}) = \sum_k (-1)^{j+k} \ell_{jk} \det(\mathcal{L}_{j,k}) = 0$

  $\det(\mathcal{L}_{j,k}) = (j,k)$-minor of $\mathcal{L}$ = determinant of submatrix made by deleting the j-th row and k-th col
- Fact for Laplacians: $\det(\mathcal{L}_{j,k}) = (-1)^{j+k} \det(\mathcal{L}_{k,k})$
- Then: $\forall j,\ \sum_k \ell_{jk} \det(\mathcal{L}_{k,k}) = 0 \quad \Rightarrow \quad \boxed{\mathcal{L}\vec{\rho} = 0 \text{ for } \rho_k = \det(\mathcal{L}_{k,k})}$

What is the functional form of $\vec{x}^*$?

- $\mathcal{L}$ is not full-rank $\Rightarrow \det(\mathcal{L}) = 0$
- Determinant expansion $\Rightarrow \forall j,\ \det(\mathcal{L}) = \sum_k (-1)^{j+k} \ell_{jk} \det(\mathcal{L}_{j,k}) = 0$

  $\det(\mathcal{L}_{j,k}) = (j,k)$-minor of $\mathcal{L}$ = determinant of submatrix made by deleting the j-th row and k-th col
- Fact for Laplacians: $\det(\mathcal{L}_{j,k}) = (-1)^{j+k} \det(\mathcal{L}_{k,k})$
- Then: $\forall j,\ \sum_k \ell_{jk} \det(\mathcal{L}_{k,k}) = 0 \quad \Rightarrow \quad \boxed{\mathcal{L}\vec{\rho} = 0 \text{ for } \rho_k = \det(\mathcal{L}_{k,k})}$
- Kirchhoff's Matrix Tree Theorem $\Rightarrow$

$$\det(\mathcal{L}_{k,k}) = (-1)^{n-1} \sum_{T \in \Theta_k} \prod_{e \in T} a_e$$

where $n =$ number of vertices

$\Theta_k = \{$spanning trees rooted at vertex $k\}$

$a_e =$ weight of edge e

What is the functional form of $\vec{x}^*$?

- $\mathcal{L}$ is not full-rank $\Rightarrow \det(\mathcal{L}) = 0$
- Determinant expansion $\Rightarrow \forall j,\ \det(\mathcal{L}) = \sum_k (-1)^{j+k} \ell_{jk} \det(\mathcal{L}_{j,k}) = 0$

  $\det(\mathcal{L}_{j,k}) = (j,k)$-minor of $\mathcal{L}$ = determinant of submatrix made by deleting the j-th row and k-th col
- Fact for Laplacians: $\det(\mathcal{L}_{j,k}) = (-1)^{j+k} \det(\mathcal{L}_{k,k})$
- Then: $\forall j,\ \sum_k \ell_{jk} \det(\mathcal{L}_{k,k}) = 0 \quad \Rightarrow \quad \boxed{\mathcal{L}\vec{\rho} = 0 \text{ for } \rho_k = \det(\mathcal{L}_{k,k})}$
- Kirchhoff's Matrix Tree Theorem $\Rightarrow$

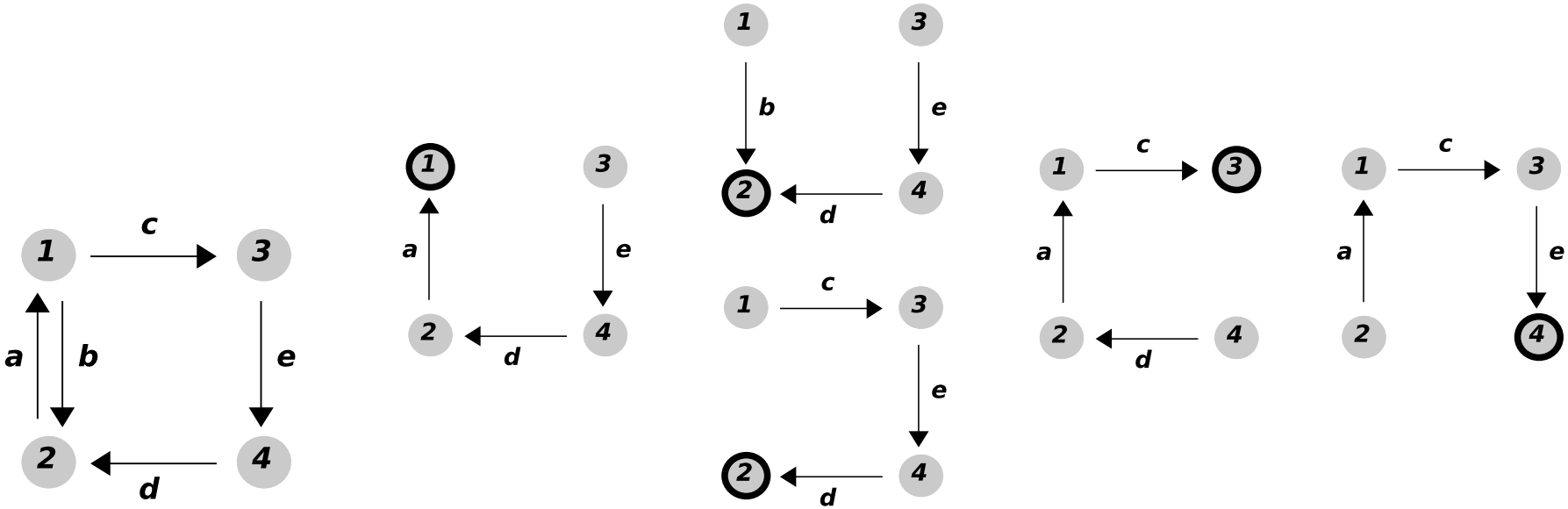
$$\det(\mathcal{L}_{k,k}) = (-1)^{n-1} \sum_{T \in \Theta_k} \prod_{e \in T} a_e$$

  where $n =$ number of vertices

  $\Theta_k = \{\text{spanning trees rooted at vertex } k\}$

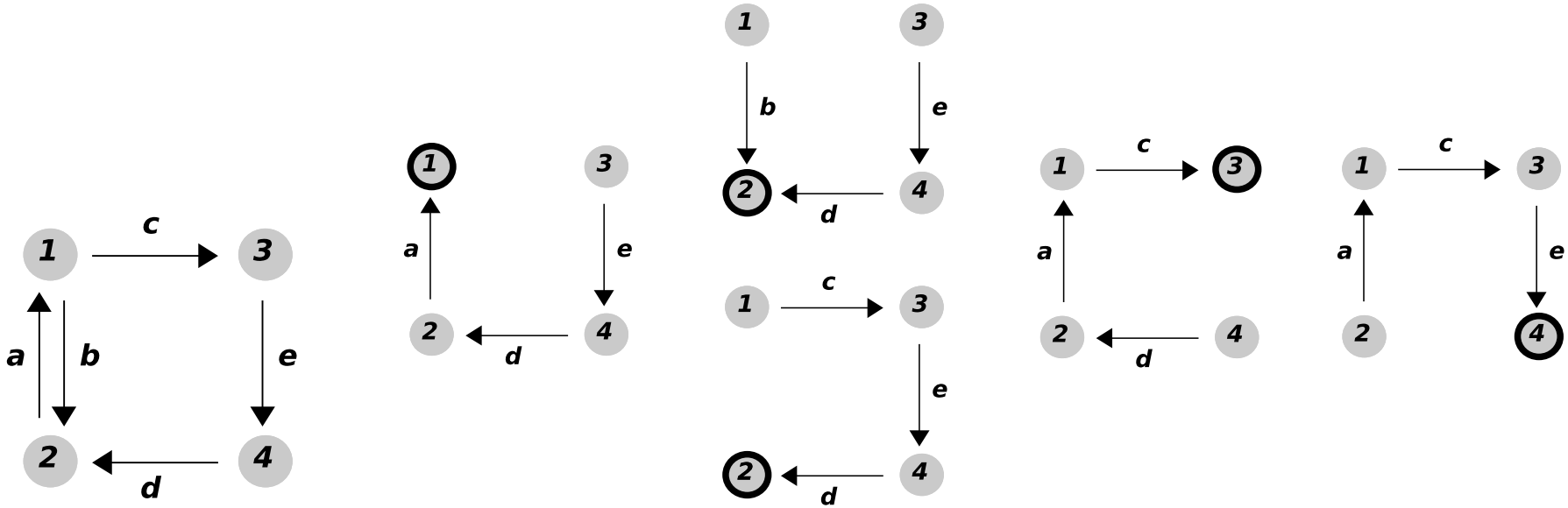
  $a_e =$ weight of edge e
- Thus, the entries of $\vec{\rho}$ are homogeneous polynomials in the rates (degree $= |T| = n - 1$)

Steady state solution: Rational polynomial in the rates



$$\vec{\rho} = \begin{pmatrix} \rho_1 \\ \rho_2 \\ \rho_3 \\ \rho_4 \end{pmatrix} = \begin{pmatrix} ade \\ bde + ced \\ cad \\ ace \end{pmatrix}$$

Steady state solution: Rational polynomial in the rates



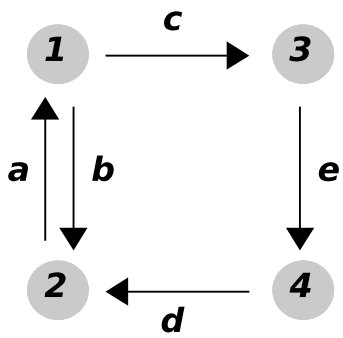
$$\vec{\rho} = \begin{pmatrix} \rho_1 \\ \rho_2 \\ \rho_3 \\ \rho_4 \end{pmatrix} = \begin{pmatrix} ade \\ bde + ced \\ cad \\ ace \end{pmatrix}$$

- $\rho_j > 0$ since ≥ 1 spanning tree at each vertex
- Normalize for steady state: $x_j^* = \rho_j/(\rho_1 + \rho_2 + \rho_3 + \rho_4)$

Gene expression level at steady state

- Gene expression level $\propto$ probability promoter bound by RNAp
- For example, suppose RNAp can bind to states 2 and 4

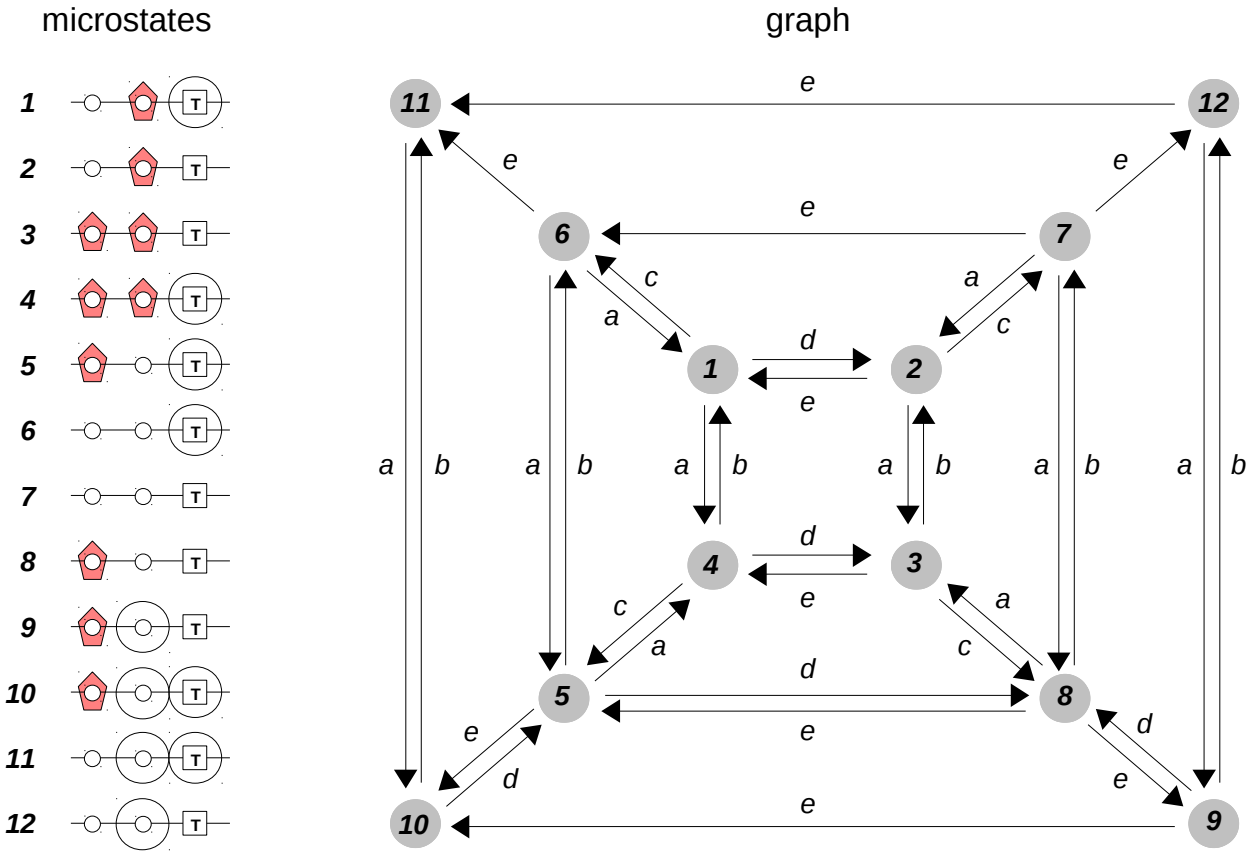
$$\text{Gene expression level} \propto x_2^* + x_4^* = \frac{\rho_2 + \rho_4}{\rho_1 + \rho_2 + \rho_3 + \rho_4} = \frac{1}{1 + \frac{\rho_1 + \rho_3}{\rho_2 + \rho_4}}$$



$$\vec{\rho} = \begin{pmatrix} \rho_1 \\ \rho_2 \\ \rho_3 \\ \rho_4 \end{pmatrix} = \begin{pmatrix} ade \\ bde + ced \\ cad \\ ace \end{pmatrix}$$

- Gene expression level = rational homogeneous polynomial
- Each term is degree $n - 1$ and positive (product of rates)

Gene expression level: Rational polynomial of rates



edge labels $a = k^{*}_{assoc}$; $b = k^{exp}_{dissoc}$; $c = k^{nuc}_{dissoc}$; $d = k_{remod}$; $e = k_{reass}$

T. Ahsendorf, F. Wong, R. Eils, J. Gunawardena. A framework for modelling gene regulation that accommodates non-equilibrium mechanisms, submitted (2013)

What is the functional form of the rates?

- Example real system: 4/5 rates are constants
- Rate of binding transcription factor $\propto$ concentration of free TF
- More complicated dependencies are possible, *e.g.*, dimerization

What could be under selection in this framework?

Sequence mutations to DNA can occur at:

- Promoter regions that bind transcription factors and other molecules
- Promoter regions that regulate expression of these upstream genes

What could be under selection in this framework?

Sequence mutations to DNA can occur at:

- Promoter regions that bind transcription factors and other molecules
- Promoter regions that regulate expression of these upstream genes

Mutations can perturb the model by:

- Adding new states (new site where TF can bind; new TF that can bind)
- Changing rates (modify binding properties of DNA or protein)
- Removing states (*e.g.*, send a rate to 0)

Acknowledgements

- Jeremy Gunawardena
- Zorana Zeravcic