What functions do gene expression levels represent?

Elaine Angelino

March 17, 2014
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
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 \( a, b \)?

We calculate \(p_{\text{bound}}\) by considering the distribution of RNAp on the non-specific sites (NNS), which make up the genome itself, and a single promoter. Then we distinguish two classes of outcomes (shown in Figure 1): all promoters unoccupied and the remaining promoters occupied.

The overall statistical weight of these states is based not just on how many of them there are but also on their statistical weight. The statistical weights associated with these different states of promoter occupancy are also shown.

The statistical question of how many ways are distributed among the non-specific sites. The statistical weights associated with these different states of promoter occupancy are also shown.

As noted above, this simple model includes two broad outcomes (shown in (a)) all promoter unoccupied \(\Delta E\) RNAp molecules bound non-specifically. For the T7 promoter, this results in

\[ \Delta E = \frac{E_N - E_P}{k_B T} \]

where \(E_N\) and \(E_P\) are the energies of RNAP to the genomic background. The probability of state \(S\) is given by considering the distribution of RNAp molecules distributed among the non-specific sites. The statistical weights associated with these different states of promoter occupancy are also shown.

The overall statistical weight of these states is based not just on how many of them there are but also on their statistical weight. The statistical weights associated with these different states of promoter occupancy are also shown.

\[ P = \text{number of RNAp molecules} \]

\[ N = \text{number of non-specific (NS) sites where RNAp can bind to DNA} \]

\[ \text{probability of state } S \propto \exp(-E_S/k_B T) \]

Simple 2-state system

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

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

Simple 2-state system: Equilibrium probabilities

\[ Z = \frac{N!}{P!(N-P)!} e^{-PE_{NS}/k_B T} + \frac{N!}{(P-1)!(N-P-1)!} e^{-(P-1)E_{NS}/k_B T} e^{-E_P/k_B T} \]

Probability RNAp bound to promoter

\[ = \frac{Y}{X + Y} = \frac{1}{1 + \frac{X}{Y}} \]

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

(Part of) a real promoter system . . .

Promoters can have many states

\[
\begin{array}{ccc}
1 & 2 & 3 \\
\hline \\
\text{transcription factor} & \text{nucleosome} & \text{TATA box} \\
\end{array}
\]

\[
2 \times 3 \times 2 = 12 \text{ states}
\]

- Example: promoter composed of 3 physical regions
- The regions can be empty or occupied
- RNApol can bind only when the 3rd region is \(\text{transcription factor}\) (6/12 states)
Transitions change occupancy at 1 promoter region

- This transition “adds” or “removes” \( \bigcirc \) at the 3rd promoter region
- These rates \( \leftrightarrow \) are physical constants
Some transitions are due to binding / unbinding

- Rate of binding $\propto [\text{free } \text{transcription factor}]

- Assumption: $[\text{free } \text{transcription factor}] \approx \text{constant w.r.t. promoter system dynamics}$
(Part of) a real promoter system . . .

This graph is sparse and strongly connected
Some transitions are irreversible
Laplacian Dynamics on General Graphs

Fig. 2 Spanning trees and the Matrix-Tree theorem. (A) On the left, as strongly-connected graph. On the right, the spanning trees rooted at a vertex, with edges outlined in black. (B) The Laplacian matrix of the graph in (A) with two minor signs, $L(G)$ and $L(G)$, calculated, illustrating the Matrix-Tree formula in Eq. (16) of Theorem 1. The determinants are placed at the spanning trees in (A) to give the corresponding values for any graph $G$, as a subgraph $T$ is a spanning tree of $G$ if $T$ reaches each vertex of $G$ and is connected and acyclic as an undirected graph. $T$ is rooted at $i \in G$ if $i$ is the only vertex of $T$ with no edges leaving it, $i \not\rightarrow j$. This implies that any non-root vertex has exactly one edge leaving it, for otherwise there would be an additional root or an undirected cycle (Fig. 2 (A)). Let $\Theta^i(G)$ denote the set of spanning trees of $G$ rooted at $i$. Note that in a general graph there may be no spanning trees rooted at a particular vertex, so that $\Theta^i(G) = \emptyset$.

Given a $n \times n$ matrix $A$, recall that $A(ij)$ is the minor given by the determinant of the $(n-1) \times (n-1)$ matrix obtained from $A$ by removing the $i$th row and $j$th column.

Theorem 1 (Matrix-Tree Theorem (Tutte 1948)) If $G$ is any graph with $n$ vertices then the minors of its Laplacian are given by

$$L(G)(ij) = (-1)^{n+i+j-1} \sum_{T \in \Theta^j(G)} \prod_{k \rightarrow l \in T} a_{kl}.$$  \hspace{1cm} (16)


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

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

\[
\begin{align*}
\frac{dx_1}{dt} &= ax_2 - (b + c)x_1 \\
\frac{dx_2}{dt} &= bx_1 + dx_4 - ax_2 \\
\frac{dx_3}{dt} &= cx_1 - ex_3 \\
\frac{dx_2}{dt} &= ex_3 - dx_4
\end{align*}
\]
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: ≥ 1 directed path between any 2 vertices
               ≥ 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

For any graph $G$ with $n$ vertices, $s$ so that $i \rightarrow j$ is the set of spanning trees of $G$, and $A$ the matrix obtained from $G$ by removing the only vertex of $G$ with no edges leaving it, for otherwise there would be an additional root or an undirected cycle (Fig. 2). Let $L(G)$ denote the Laplacian matrix of $G$. Then, the minors of its Laplacian are given by

$$L(G)_{ij} = (-1)^{n+i+j-1} \sum_{T \in \Theta_j(G)} \prod_{a \rightarrow l \in T} a.$$  

(16)

What functions do gene expression levels represent?
All spanning trees

Fig. 2 Spanning trees and the Matrix-Tree theorem. (A) On the left, as strongly-connected graphs, the minors of its Laplacian are given by

\[ L(G)(ij) = (-1)^{n+i+j-1} \sum_{T \in \Theta_j(G)} \left( \prod_{a \rightarrow l \in T} a \right) \]

(16)

What functions do gene expression levels represent?
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: $\forall j, \sum_k \ell_{jk} \det(\mathcal{L}_{k,k}) = 0 \Rightarrow \mathcal{L}\vec{\rho} = 0$ 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 column.
- Fact for Laplacians: $\det(\mathcal{L}_{j,k}) = (-1)^{j+k} \det(\mathcal{L}_{k,k})$
- Then: $\forall j$, $\sum_k \ell_{jk} \det(\mathcal{L}_{k,k}) = 0$ $\Rightarrow$ $\mathcal{L}\vec{\rho} = 0$ 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 $e$
What is the functional form of $\vec{x}^*$?

- $\mathcal{L}$ is not full-rank $\Rightarrow$ $\det(\mathcal{L}) = 0$

- Determinant expansion $\Rightarrow$ $\forall j$, $\det(\mathcal{L}) = \sum_k (-1)^{j+k} \ell_{jk} \det(\mathcal{L}_{j,k}) = 0$
  \[ \det(\mathcal{L}_{j,k}) = (j, k)\text{-minor of } \mathcal{L} = \text{determinant of submatrix made by deleting the } j\text{-th row and } k\text{-th col} \]

- Fact for Laplacians: $\det(\mathcal{L}_{j,k}) = (-1)^{j+k} \det(\mathcal{L}_{k,k})$

- Then: $\forall j$, $\sum_k \ell_{jk} \det(\mathcal{L}_{k,k}) = 0$ $\Rightarrow$ $\mathcal{L}\vec{\rho} = 0$ 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 $n = \text{number of vertices}$
  \[ \Theta_k = \{ \text{spanning trees rooted at vertex } k \} \]
  \[ a_e = \text{weight of edge } e \]

- 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

\[ \rho_j > 0 \text{ since } \geq 1 \text{ spanning tree at each vertex} \]

Normalize for steady state: \( x_j^* = \frac{\rho_j}{(\rho_1 + \rho_2 + \rho_3 + \rho_4)} \)
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

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

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

microstates

1
2
3
4
5
6
7
8
9
10
11
12

graph

edge labels  \( a = k_{assoc} \);  \( b = k_{exp}^{dissoc} \);  \( c = k_{nuc}^{dissoc} \);  \( d = k_{remod} \);  \( e = k_{reass} \)

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

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)
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