Transcription and Gene Regulation

- Molecular stochasticity in single cells.
- Transcription factors and their regulatory motifs.
- Biophysics of recognition: facilitated diffusion and the search for regulatory motifs.
- Evolution of the regulatory vocabulary.
- Evolutionary rewiring of transcription networks.
Variation in Gene Expression in Single Cells

- Typically, gene expression is measured in populations of cells, obscuring the variation that exists at the individual level.

- Because the average numbers of proteins of a particular type is often well below 100 / cell, stochastic cell-to-cell variation can be large.

- Total among-cell variation in gene expression can be subdivided into three components:

  1) genetic variation among cells, \( \text{essential for evolutionary change} \)
  2) extrinsic environmental variance, \( \text{reduce the efficiency of selection} \)
  3) intrinsic noise due to the vagaries of random molecular motion and production. \( \text{minimum possible variation} \)

- Heritability \( (h^2) = \frac{V_G}{(V_G + V_{Ee} + V_{Ei})} \).

- Response to Selection = Heritability x Selection Differential
Central Kinetic Components of Protein Production

- Probability gene is on: \( P_{on} = \frac{k_{on}}{k_{on} + k_{off}} \)
Temporal Variation in Gene Expression Within a Cell

- Gene is stochastically on / off depending on the binding of cognate transcription factors.

- Messenger RNAs are produced during on periods, but decay away at exponential rates during off periods.

- Protein numbers rise during periods of mRNA abundance, and decline slowly via during periods of mRNA rarity. Fluctuations in protein numbers are damped, owing to their greater longevities than mRNA molecules.
• There is no correlation between the number of mRNAs and the number of protein molecules within individual cells.

• Proteins numbers are much higher than mRNA numbers.

• Raises issue about single-cell transcriptomics studies.

Median half lives of mRNAs:

<table>
<thead>
<tr>
<th>Organism</th>
<th>Half Life</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>E. coli</td>
<td>5 minutes</td>
<td>Taniguchi et al. 2010</td>
</tr>
<tr>
<td>S. cerevisiae</td>
<td>22 minutes</td>
<td>Wang et al. 2002</td>
</tr>
<tr>
<td>Mouse fibroblast</td>
<td>9 hours</td>
<td>Schwanhausser et al. 2011</td>
</tr>
</tbody>
</table>

From Li and Xie (Nature, 2011)
Stochastic Gene Expression in Single Cells

- Two identical genes in the same bacterial cells, with the proteins labeled by different fluorescent markers, will have the same expression level only if there is no intrinsic noise within the cell.

From Elowitz et al. (Science, 2002)
The Number of mRNAs / Cell is Expected to be Poisson Distributed for a Constitutively Expressed Gene

\[ p(n_m) = 0 \xleftarrow{k_m \gamma_m} 1 \xrightarrow{k_m \gamma_m} 2 \xleftarrow{k_m \gamma_m} 3 \xrightarrow{k_m \gamma_m} 4 \]

Mean number of mRNAs \((n_m) / \text{cell} = k_m / \gamma_m\)

Mean = Variance
If the Gene is Regulated and Only Active for a Fraction of Time $P_{on}$, Number of mRNA Molecules / Cell Follows a Mixture of Distributions

Mean number of mRNAs ($n_m$) / cell = $P_{on} k_m / \gamma_m$
Distributions of Transcript Numbers per Cell: regulated genes exhibit elevated levels of among-cell variance in expression, and can even be bimodal.

Constitutive expression (Poisson)

Regulated expression (mixture)

Munsky et al. (Science, 2012)
Extension to Protein Numbers / Cell

Changes in transcript numbers:

Changes in protein numbers:

transcript
protein
Average Numbers of Protein Molecules in Individual *E. coli* Cells Are Small

Individual distributions are roughly gamma in form, with $a$ being a measure of burst frequency, and $b$ a measure of burst size. Poisson: variance, $\sigma^2 = \text{mean}, \mu$

From: Taniguchi et al. (2010, Science)
Biological Features of Transcription Factors

• Transcription is generally nonautonomous, as one to several accessory TFs must be present simultaneously for transcriptional activation.

• In eukaryotes, individual TFs often service multiple genes, which facilitates coregulation of gene expression.

• *E. coli* has ~300 transcription factors: 7 control the expression of ~50% of regulated genes, whereas ~60 service single genes.
Figure 2 Drafting the yeast transcriptional regulatory map. a, Portions of chromosomes illustrating locations of genes (grey rectangles) and conserved DNA sequences (coloured boxes) bound in vivo by transcriptional regulators. b, Combining binding data and sequence conservation data. The diagram depicts all sequences matching a motif from our compendium (top), all such conserved sequences (middle) and all such conserved sequences bound by a regulator (bottom). c, Regulator binding site distribution. The red line shows the distribution of distances from the start codon of open reading frames to binding sites in the adjacent upstream region. The green line represents a randomized distribution.

From Harbison et al. (Nature, 2004)
Scaling of the Number of TFs with the Number of Protein-coding Genes – typically, 1 to 5% of the protein-coding genes within a genome are TFs

- Prokaryotes: number of TFs scales with ~2 power of gene number.
- Eukaryotes: number of TFs scales with ~1 power of gene number.

From Charoensawan et al. (Nuc. Acids Res., 2010)
Transcription-factor Binding Motifs Are Small, and Longer in Prokaryotes Than Eukaryotes

- Typically, 10 to 50 amino-acid residues in the TF are involved in contacts with the DNA.

**Figure 1** The lengths of binding sites range from 5 nt to \(\sim 30\) nt, in both eukaryotes (left, 454 curated transcription factor motifs) and prokaryotes (right, 79 motifs). The information content per nucleotide ranges from \(\sim 0.25\) bits to 2 bits (see Figure S1).

- Although each TF has maximum affinity for a specific DNA motif, there is no general regulatory code in TFs, i.e., no specific language involving one-to-one matching between the amino-acid sequence of a TF and the nucleotide sequence of its binding site.
Transcription-Factor Molecules Are Especially Rare in Cells

Human lymphoblastoid cell line (Marinov et al. 2014, Genome Research)
Transcription-factor Molecules Locate Their Binding Sites by Facilitated Diffusion

• All TFs engage in promiscuous interactions with off-target sites as a consequence of the negatively charged phosphate backbones of the DNA and positively charged residues on the protein.

In bacteria, transcription and translation are colocalized, and genes often appear in operons, increasing the chance of rapid localization.

A combination of one-dimensional sliding and three-dimensional jumping dramatically reduces the search time, relative to random diffusion.

From Kolesov et al. (PNAS, 2007)
How Rapidly Do TFs Find Their Cognate TFBs by Facilitated Diffusion?

- With one-dimensional diffusion in *E. coli*, once on the DNA, it would take ~29 days for a TF to find a specific binding site.

- With three-dimensional diffusion, the encounter rate between jumps = \[ 4\pi(D_3 + D_{3p})(r_n + r_p) \]

- Accounting for the size (r) of nucleotides and TFs, and the diffusion rate of a TF (D_{3p}), the time to jump from one location to another is \(~2.5 \times 10^{-7}\) seconds.

- Once on the DNA, a TF spends ~0.0026 sec diffusing over ~100 bp, before falling off, so essentially all of the search time is spent directly interrogating the DNA, rather than jumping from spot to spot.
  - Because \(~10^5\) 100-bp scans are required to cover the entire genome, the estimated time to locate a site is \(10^5 \times 0.0026 = 260\) seconds.
  - With \(N_{tf}\) molecules in the cell, the search time would be reduced to \(260/N_{tf}\) seconds.
Gene regulation in LUCA must have relied on transcription factors, but only a small fraction of known DNA-binding domains are shared across the three super-kingdoms.
Transcription Factors Bind to Specific Motifs With Different Binding Affinities

- The motifs generally vary among individual client genes, and seldom match the consensus sequence.
Towards a Physical Model for Understanding Gene-Expression Evolution:
Characterizing Binding Sites by the Strength of Their Motif Sequences

- Affinity between TFs and their cognate TFBSs on the DNA are governed by hydrogen bonds.

\[
E(a) = \sum_{i=1}^{\ell} c_i(a_i)
\]

Table 1. Features of the motifs of well-studied transcription factors (TFs). Motif size is based on consensus sequences. The estimated costs of mismatches are obtained from binding-strength experiments in which single-base changes were made in motifs. Costs of single-base mismatches are in units of kcal/mol; these average to 1.4 across the full set of studies, or in terms of Boltzmann units \((k_B T \approx 0.6 \text{ kcal/mol})\), to 2.3.

<table>
<thead>
<tr>
<th>TF</th>
<th>Species</th>
<th>Motif (bp)</th>
<th>Cost of Mismatch</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>CI</td>
<td>Lambda phage</td>
<td>17</td>
<td>1.4</td>
<td>Sarai and Takeda (1989)</td>
</tr>
<tr>
<td>Cro</td>
<td>Lambda phage</td>
<td>9</td>
<td>1.4</td>
<td>Takeda et al. 1989</td>
</tr>
<tr>
<td>Mnt</td>
<td>Salmonella phage P22</td>
<td>21</td>
<td>1.0</td>
<td>Fields et al. (1997); Berggrun and Sauer (2001)</td>
</tr>
<tr>
<td>CRP</td>
<td>Escherichia coli</td>
<td>22</td>
<td>1.7</td>
<td>Gunasekera et al. (1992); Kinney et al. (2010)</td>
</tr>
<tr>
<td>CRP</td>
<td>Synechocystis sp.</td>
<td>22</td>
<td>1.8</td>
<td>Omagari et al. (2004)</td>
</tr>
<tr>
<td>ArcA</td>
<td>Shewanella oneidensis</td>
<td>15</td>
<td>1.3</td>
<td>Schildbach et al. (1999); Wang et al. (2008)</td>
</tr>
<tr>
<td>Gcn4</td>
<td>Saccharomyces cerevisiae</td>
<td>11</td>
<td>1.0</td>
<td>Nuti et al. (2011)</td>
</tr>
<tr>
<td>c-Myb</td>
<td>Homo sapiens</td>
<td>6</td>
<td>1.6</td>
<td>Oda et al. (1998)</td>
</tr>
</tbody>
</table>
Probability a Particular Target Site with \( m \) Matches is Bound

\[ P_{on} \simeq \frac{1}{1 + Be^{-2m}}, \]

where \( B = G/(N_{tf} - N_{ot}) \) is a measure of the concentration of background (non-specific) binding sites relative to the number of TF molecules available for the target site.
What is the Magnitude of Background Off-Target Promiscuity?

- The number of off-target sites, $G$, is generally in the range of $10^6$ to $10^{10}$ bp, with prokaryotes falling at the lower end and multicellular eukaryotes at the higher end of the range.

- In bacteria, the numbers of molecules per cell for particular TFs, $N_{tf}$, are often in the range of 100 to 1000, with just a few cases ranging as high as 50,000. The number of genes serviced by a particular TF, $N_{ot}$, is generally <100.

- Thus, $B$ is on the order of $10^3$ to $10^6$ for prokaryotes, and estimates for eukaryotes are in the same range.

- If other sources of interference exist (such as promiscuous binding to other proteins), $B$ would be higher.
The Probability of Binding-site Occupancy Typically Saturates at a Small Number of Matches

\[ P_{on} \approx \frac{1}{1 + Be^{-2m}} \]

- Unless the level of promiscuous binding is enormous, there is little advantage of a binding-site lengths > 10 bp.

- Two costs / limitations of using TFs to regulate genes:
  1) Changing the number of matches is a coarse-grained tuning mechanism.
  2) Owing to promiscuous binding to nonspecific sites, 100s of TF molecules need to be present in a cell to ensure that a host gene is turned on.
Relating Binding Strength to Fitness: function of the number of matching sites, \( m \)

\[
W(m) = 1 + \alpha P_{on} = 1 + \frac{\alpha}{1 + e^{-2(\ell-n) + \ln(B)}}
\]

- \( n \) = number of mismatches
- \( \ell \) = length of the optimal motif sequence
- \( \alpha \) = scaling factor for the strength of selection
- \( B \) = measure of background interference
Quasi-Equilibrium Evolutionary Distribution of Binding-Site Affinities: the Neutral Case

\[ P(n) = 0 \quad \frac{3\ell\mu}{\mu} \quad 1 \quad \frac{3(\ell-1)\mu}{2\mu} \quad 2 \quad \frac{3(\ell-2)\mu}{3\mu} \quad 3 \quad \frac{3(\ell-3)\mu}{4\mu} \]

Decreasing number of matching sites

\( \mu = \) the rate of mutation from nucleotide x to nucleotide y (reversion rate)

\( 3\mu = \) rate of loss of a correct site
The fixation probabilities are obtained from Kimura’s (1962) diffusion equation for newly arisen mutations,

$$\hat{\phi}_{x,y} = \frac{1 - e^{-2N_es_{x,y}/N}}{1 - e^{-2N_es_{x,y}}}$$

where $N_e$ is the effective population size, $1/N$ is the initial frequency of a mutation, and $s_{x,y}$ is the fractional selective advantage of allelic class $y$ over $x$. 
Equilibrium Distributions of the Number of Matches

- Distribution is independent of the mutation rate.

- Depends only on the multiplicity under neutrality.

- The distribution under selection is simply a weighting of the neutral expectation.

- Unless the power of selection relative to drift ($N_e \alpha$) is extremely strong, the majority of motifs will contain mismatches.

\[ \tilde{P}(m) = C \left[ \frac{3^n}{n!} \right] \]

the neutral expectation
Using the Theory to Estimate the Strength of Selection on Binding Sites

- CRP motifs in *E. coli*.

From Mustonen and Lassig (2005, PNAS)
Transcription-factor Binding Motifs Appear to Wander Over Evolutionary Time

Figure 2. Comparison of isoform specificities. DNA-binding specificities of 17 Lola isoforms generated through alternate splicing. MatAlign clustergram emphasizing the diversity within the recognition motifs of the various Lola isoforms. All of the characterized ZFPs utilize a pair of zinc fingers to recognize DNA. Identical fingers are present in the lola-PN and -PY isoforms and the lola-PT and -PU isoforms, and both pairs have identical specificity.

Fig. 1. DNA binding-site motifs bound by forkhead domain proteins. A representative member of each class of binding site discussed in the text is shown. Solid symbols are used to represent binding specificities in subsequent figures.

From Enuameh et al. (Gen. Res., 2013) From Nakagawa et al. (PNAS, 2013)
Coevolution of the Regulatory Vocabulary: TFs and Their Binding Sites

- TFs with larger numbers of target genes are more evolutionarily conserved at the amino-acid sequence level, including at the level of the recognition sequence.

- Decline in binding-site specificity with increasing numbers of genes serviced by a TF in both E. coli and yeast (Sengupta et al. 2002).

- Does such a condition evolve by selection so as to minimize the mutational burden on an organism?

- Or are TFs with low specificity recruited more frequently into various regulatory pathways over evolutionary time?
Dramatic Rewiring of Regulatory Mechanisms Is Commonly Observed in Yeasts

- Massive differences in the regulatory machinery associated with the ribosomal protein genes in the two yeasts *Saccharomyces cerevisiae* and *Candida albicans*. Nearly every TF used in Sc is utilized in a different way in Ca, and shifts in the consensus motifs for orthologous TFs occur as well.

From Weirauch and Hughes (Trends Genetics, 2009)
Potential Paths of Rewiring of Regulatory Modules Involving Intermediate Stages of Redundancy

• Intermediate states of shared (redundant) regulatory motifs; subsequently experiencing reciprocal loss.

From Weirauch and Hughes (Trends Genetics, 2009)
Mechanisms of Regulatory Rewiring:
How can a large set of co-regulated genes experience a coordinated switch of their regulatory pathways?

Fig. 2. A plausible pathway to the concurrent rewiring of a large set of genes. In this scenario an interaction is acquired between TRs A and B, after which interactions between B and DNA are optimized gene-by-gene. Rewiring in this manner could avoid fitness barriers imposed by initially changing regulation one gene at a time.