BE 150: Design Principles of Genetic Circuits

Justin Bois

Caltech

Spring, 2018

Donna and Benjamin M.
Rosen Bioengineering Center

This document was prepared at Caltech with financial support from the Donna and Benjamin M. Rosen Bioengineering Center.

10 Noisy gene expression

Thus far, we have written deterministic systems of ordinary differential equations to describe the dynamics of mRNA and protein concentrations in cells. When we considered the statistical mechanical model for gene expression regulation, we considered the probability of a promoter being bound. This probability is considered over an *ensemble* of states. We assumed that there are enough of all the components, and enough time, to sample the states of this ensemble. So, we were looking at rates of gene expression considering large numbers of the molecular constituents.

In our study of noise this week, we will address two questions.

1. What are the sources of noise and how do we characterize them?
2. What are the dynamics of the entire probability distribution of protein levels? From this, can we identify design principles?

We will address the first question today and the second question in the next lecture. To do so, as suggested in the second question, we will use the mathematical machinery of probability, an important tool for every biologist and bioengineer to have in his or her toolbox.

10.1 Noise is present in many genetic circuits

The copy numbers of many proteins and mRNA molecules are small in a given cell. A recent study quantified counts of transcription factors. The result is shown in Fig. 14. As evident in the figure, many of the copy numbers of come transcription factors, relative to the number of binding sites, is between 1 and 10. These are small numbers! This means that they are susceptible to fluctuations.

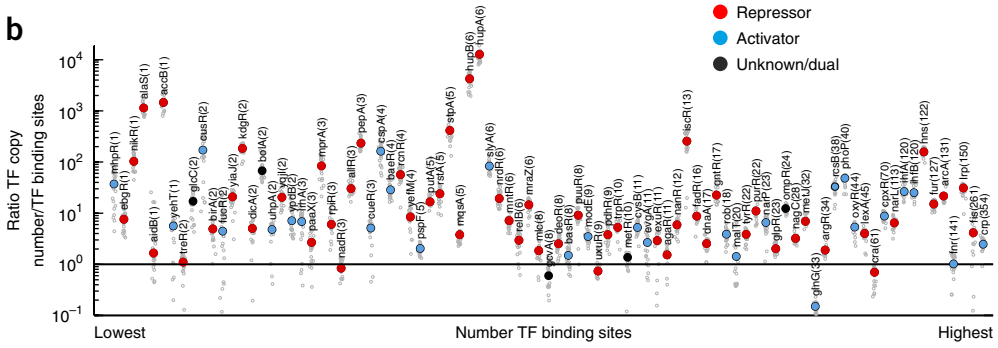

Figure 14: Quantification of copy number of transcription factors in *E. coli*. Taken from Schmidt, et al., *Nature Biotech.*, **34**, 104–113, 2016.

Generally speaking, we call deviations from what we might expect from our deterministic view of gene expression *stochasticity*, or *noise*. This is a key concept, because nearly all cellular processes are susceptible to noise, for a host of reasons, including low copy numbers of molecular regulators of gene expression.

10.2 Definition of total noise

To quantify the noise, we need to define a metric. To do so, we start with some other definitions. Let $n(t)$ be the copy number of a protein of interest.[2] Let $P(n)$ be the probability of observing n copies of the protein of interest. We define the noise, η_{tot}, as the coefficient of variation of gene expression. This is given by the standard deviation of n over its mean. Thus,

$$\eta_{\text{tot}}^2 = \frac{\mu_2 - \mu_1^2}{\mu_1^2}, \tag{10.1}$$

where μ_m is the mth moment of the probability distribution,

$$\mu_m = \sum_{n=0}^{\infty} n^m P(n). \tag{10.2}$$

If the standard deviation is comparable to the mean, as we would expect in the case of large copy numbers, we have low noise, but if it is large compared to the mean, we have high noise.

Because we do not know the probability distribution of copy number, $P(n)$, but rather can only measure copy number (or often a quantity approximately proportional to copy number, as is the case with fluorescence measurements), we can approximate the noise with its plug-in estimate from the experimental data. For example, if the fluorescence intensity c is related to the copy number by $ac = n$, then we have

$$\eta_{\text{tot}}^2 = \frac{\langle n^2 \rangle - \langle n \rangle^2}{\langle n \rangle^2} = \frac{\langle c^2 \rangle - \langle c \rangle^2}{\langle c \rangle^2}. \tag{10.3}$$

10.3 Extrinsic and intrinsic noise

We would like to know more about the noise, specifically where it comes from. We can separate the noise into intrinsic and extrinsic noise.

Intrinsic noise: Transcription and translation can occur at different times and rates in otherwise identical systems. This results in fluctuations in $n(t)$. The fluctuations in the copy number of the protein of interest are due to fluctuations that affect *only* the gene of interest. Operationally, intrinsic noise causes the failure of identical genes in identical environments to correlate. This fundamentally limits the precision of regulation.

Extrinsic noise: Other molecular species, such as RNA polymerase, ribosomes, chemical species in the cell's environment, vary over time and affect the gene of interest. The fluctuations in the copy number of the protein of interest are due to fluctuations that affect *all* genes in a cell.

The total noise is written as the sum of intrinsic and extrinsic contributions.[3]

$$\eta_{\text{tot}}^2 = \eta_{\text{int}}^2 + \eta_{\text{ext}}^2. \tag{10.4}$$

[2] I choose n here instead of p like we have been using so as to avoid confusion with the probability distribution P.

[3] It is not clear that we can write the noise as a sum of these two contributions in this way, and we will gloss over that here. For a more complete treatment of intrinsic versus extrinsic noise, see Hilfinger and Paulsson, *PNAS*, **108**, 12167–12172 [link].

With this in mind, we can define a probability distribution that also contains the extrinsic and intrinsic variables that contribute to gene expression level. We denote these respectively by $\mathbf{E}$ and $\mathbf{I}$. Let $P(n, \mathbf{E}, \mathbf{I})$ be the joint probability distribution for n, $\mathbf{E}$, and $\mathbf{I}$. This probability mass function $P(n)$ is found by marginalizing the extrinsic and intrinsic variables.

$$P(n) = \int d\mathbf{E} \int d\mathbf{I}\, P(n, \mathbf{E}, \mathbf{I}). \tag{10.5}$$

We have been a little cavalier here because some of the intrinsic and extrinsic variables may be discrete, and we have written continuous distribution. The general argument does not change if we explicitly consider discrete and continuous variables, so we will write them all as continuous for ease of notation.

The mth moment is then

$$\mu_m = \sum_{n=0}^{\infty} n^m P(n) = \sum_{n=0}^{\infty} n^m \int d\mathbf{E} \int d\mathbf{I}\, P(n, \mathbf{E}, \mathbf{I}) = \int d\mathbf{E} \int d\mathbf{I} \sum_{n=0}^{\infty} n^m P(n, \mathbf{E}, \mathbf{I})$$

$$= \int d\mathbf{E} \int d\mathbf{I}\, \mu_m^{\mathbf{E},\mathbf{I}}, \tag{10.6}$$

where we have defined $\mu_m^{\mathbf{E},\mathbf{I}}$ as the mth moment for a *particular* $\mathbf{E}$ and $\mathbf{I}$. We define

$$\bar{x} = \int d\mathbf{E}\, x \tag{10.7}$$

$$\text{and } \langle x \rangle = \int d\mathbf{I}\, x, \tag{10.8}$$

such that

$$\mu_m = \overline{\left\langle \mu_m^{\mathbf{E},\mathbf{I}} \right\rangle} = \overline{\left\langle \mu_m^{\mathbf{E},\mathbf{I}} \right\rangle}. \tag{10.9}$$

The total noise is then given by

$$\eta_{\text{tot}}^2 = \frac{\mu_2 - \mu_1^2}{\mu_1^2} = \frac{\overline{\left\langle \mu_2^{\mathbf{E},\mathbf{I}} \right\rangle} - \overline{\left\langle \mu_1^{\mathbf{E},\mathbf{I}} \right\rangle}^2}{\overline{\left\langle \mu_1^{\mathbf{E},\mathbf{I}} \right\rangle}^2}. \tag{10.10}$$

Now, say we are interested in the intrinsic noise. We first compute the coefficient of variation for a *fixed* $\mathbf{E}$. We define

$$\mu_m^{\mathbf{I}}(\mathbf{E}) = \left\langle \mu_m^{\mathbf{E},\mathbf{I}} \right\rangle = \int d\mathbf{I} \sum_{n=0}^{\infty} n^m\, P(n, \mathbf{E}, \mathbf{I}). \tag{10.11}$$

To find the contributions to the variance by the intrinsic part, we average the first and second moment over the extrinsic variables.

$$\text{var}_{\text{int}} = \int d\mathbf{E} \left[\mu_2^{\mathbf{I}} - \left(\mu_1^{\mathbf{I}}\right)^2 \right] = \overline{\mu_2^{\mathbf{I}} - \left(\mu_1^{\mathbf{I}}\right)^2} = \overline{\mu_2^{\mathbf{I}}} - \overline{\left(\mu_1^{\mathbf{I}}\right)^2} = \mu_2 - \overline{\left(\mu_1^{\mathbf{I}}\right)^2}. \tag{10.12}$$

So, we can define the intrinsic contribution to the noise as the intrinsic standard deviation over the mean, or

$$\eta_{\text{int}}^2 = \frac{\text{var}_{\text{int}}}{\mu_1^2} = \frac{\mu_2 - \overline{\left(\mu_1^{\mathbf{I}}\right)^2}}{\mu_1^2}. \tag{10.13}$$

Because the remaining noise must be extrinsic, we can use the expressions for the total noise, (10.10), and the intrinsic noise, (10.13), to write

$$\eta^2_{\text{tot}} = \eta^2_{\text{int}} + \eta^2_{\text{ext}} = \frac{\mu_2 - \overline{\left(\mu_1^{\mathbf{I}}\right)^2}}{\mu_1^2} + \frac{\overline{\left(\mu_1^{\mathbf{I}}\right)^2} - \mu_1^2}{\mu_1^2}. \tag{10.14}$$

10.4 Computing noise from experiments

Let's say we are monitoring copy number by fluorescence. Let c be the fluorescence intensity, say for a CFP channel, and we assume $c = n/a$; i.e., the fluorescence intensity is proportional to the copy number. In this case, we may use c as a proxy for n in computing the noise, since the constant of proportionality will cancel out. Computing the total noise is easy, since our experiment naturally contains all of the intrinsic and extrinsic variables. Say we have N cells. Then,

$$\mu_1 \approx \frac{a}{N}\sum_{k=1}^{N} c_k \tag{10.15}$$

$$\mu_2 \approx \frac{a^2}{N}\sum_{k=1}^{N} c_k^2, \tag{10.16}$$

where c_k is the integrated intensity for cell k. The problem is that we cannot compute $\overline{\left(\mu_1^{\mathbf{I}}\right)^2}$ directly from measurements. We cannot separate out extrinsic from intrinsic variables.

This conundrum was tackled in a clever experiment by Elowitz, et al., *Science*, **197**, 1183–1186, 2002. They realized that if each cell has two different genes coding for fluorescent proteins, say CFP and YPF, with identical regulatory sequences that can be measured simultaneously, we can compute $\overline{\left(\mu_1^{\mathbf{I}}\right)^2}$. This can be seen as follows. Because both copies of the gene are in the same cell, they experience the same extrinsic variables. Then, we can write

$$\begin{aligned}
\overline{\left(\mu_1^{\mathbf{I}}\right)^2} &= \int \mathrm{d}\mathbf{E} \left(\int \mathrm{d}\mathbf{I} \sum_{n=0}^{\infty} n\, P(n, \mathbf{E}, \mathbf{I}) \right)^2 \\
&= \int \mathrm{d}\mathbf{E} \left(\int \mathrm{d}\mathbf{I}_c \sum_{n_c=0}^{\infty} n_c\, P(n_c, \mathbf{E}, \mathbf{I}_c) \right) \left(\int \mathrm{d}\mathbf{I}_y \sum_{n_y=0}^{\infty} n_y\, P(n_y, \mathbf{E}, \mathbf{I}_y) \right) \\
&= \int \mathrm{d}\mathbf{E} \int \mathrm{d}\mathbf{I}_c \int \mathrm{d}\mathbf{I}_y \sum_{n_c=0}^{\infty} \sum_{n_y=0}^{\infty} n_c\, n_y\, P(n_c, n_y, \mathbf{E}, \mathbf{I}_c, \mathbf{I}_y) \\
&\approx \frac{1}{N}\sum_{k=1}^{N} n_{c,k}\, n_{y,k} = \frac{ab}{N}\sum_{k=1}^{N} c_k y_k,
\end{aligned} \tag{10.17}$$

where c_k is the integrated fluorescence intensity of cell k in the CFP channel with y_k similarly defined. So, if we have two different fluorescent reporters on identical genes, we can compute the necessary average from the product of the intensities of each reporter.

Elowitz and coworkers built strains of *E. coli* that contained CFP and YFP reporter genes in the chromosomal genome with identical promoters. The promoters were equidistant from the origin of

replication, so their expression level should be exactly the same in the absence of noise. If there is no intrinsic noise, the levels of CFP and YFP should be identical throughout time, and all of the cells should be the same color. In the presence of intrinsic noise, the CFP and YFP signals will be different.

To quantify the intrinsic versus extrinsic noise, we measure many cells in a population and compute the integrated CFP and YFP intensity for each cell. To ease notation, we define

$$\langle c \rangle = \frac{1}{N} \sum_k c_k, \tag{10.18}$$

$$\langle c^2 \rangle = \frac{1}{N} \sum_k c_k^2, \tag{10.19}$$

$$\langle cy \rangle = \frac{1}{N} \sum_k c_k y_k, \tag{10.20}$$

with $\langle y \rangle$ and $\langle y^2 \rangle$ similarly defined. There should be no notational confusion with the angled brackets of the previous section; the present angled brackets indicate averaging over a population of cells. Then, we have plug-in estimates for the moments,

$$\mu_1 \approx \langle n_c \rangle = \langle n_y \rangle = a\langle c \rangle = b\langle y \rangle, \tag{10.21}$$

$$\mu_2 \approx \langle n_c^2 \rangle = \langle n_y^2 \rangle = a^2\langle c^2 \rangle = b^2\langle y^2 \rangle. \tag{10.22}$$

Here, we have defined b analogously to a; it is the constant of proportionality between YFP copy number and measured YFP fluorescence intensity. Because the genes are identical, the distributions of the copy numbers should be identical, meaning that the moments and their plug-in estimates are also identical. That is, $a^m\langle c^m \rangle = b^m\langle y^m \rangle$, which we express in the equalities in the above expressions. We can then write the contributions to the noise as

$$\eta_{tot}^2 = \frac{a^2\langle c^2 \rangle - a^2\langle c \rangle^2}{a^2\langle c \rangle^2} = \frac{\frac{1}{2}\left(a^2\langle c^2 \rangle + b^2\langle y^2 \rangle\right) - ab\langle c \rangle\langle y \rangle}{ab\langle c \rangle\langle y \rangle}, \tag{10.23}$$

$$\eta_{int}^2 = \frac{a^2\langle c^2 \rangle - ab\langle cy \rangle}{a^2\langle c \rangle^2} = \frac{\frac{1}{2}\left(a^2\langle c^2 \rangle + b^2\langle y^2 \rangle\right) - ab\langle cy \rangle}{ab\langle c \rangle\langle y \rangle}, \tag{10.24}$$

$$\eta_{ext}^2 = \eta_{tot}^2 - \eta_{int}^2 = \frac{\langle cy \rangle - \langle c \rangle\langle y \rangle}{\langle c \rangle\langle y \rangle}. \tag{10.25}$$

We can arbitrarily choose units of fluorescence such that $a = b$. In practice, this involves setting intensity units such that $\langle c \rangle = \langle y \rangle = 1$. Then, the expressions simplify to

$$\eta_{tot}^2 = \frac{\frac{1}{2}\left(\langle c^2 \rangle + \langle y^2 \rangle\right) - \langle c \rangle\langle y \rangle}{\langle c \rangle\langle y \rangle}, \tag{10.26}$$

$$\eta_{int}^2 = \frac{\frac{1}{2}\left(\langle c^2 \rangle + \langle y^2 \rangle\right) - \langle cy \rangle}{\langle c \rangle\langle y \rangle} = \frac{\langle (c-y)^2 \rangle}{2\langle c \rangle\langle y \rangle}, \tag{10.27}$$

$$\eta_{ext}^2 = \frac{\langle cy \rangle - \langle c \rangle\langle y \rangle}{\langle c \rangle\langle y \rangle}. \tag{10.28}$$

Note that the extrinsic noise is independent of the choice of fluorescence units.

Looking at these expressions, we see that the extrinsic noise is proportional to the covariance of the two signals. This makes sense, since correlation between the two should depend on external considerations. The intrinsic noise is proportional to $\langle (c-y)^2 \rangle$, the deviation between CFP and YFP.

10.5 Experimental results

In Elowitz, et al., *Science*, **197**, 1183–1186, 2002, the authors performed experiments to study intrinsic versus extrinsic noise in *E. coli*. The two identical promoters we discussed in the last section were repressed by LacI, which itself was modulated by IPTG, which inhibits LacI. They could therefore dial the level of expression up and down to see how this affects noise.

In Fig. 15, we see measurements of fluorescence intensity for a group of cells. Because the genes are in the same cell, they are subject to the same extrinsic fluctuations, which are manifest in scatter of the point along the diagonal. Orthogonal scatter is due to intrinsic fluctuations. We see scatter in both directions, indicating the presence of both intrinsic and extrinsic noise.

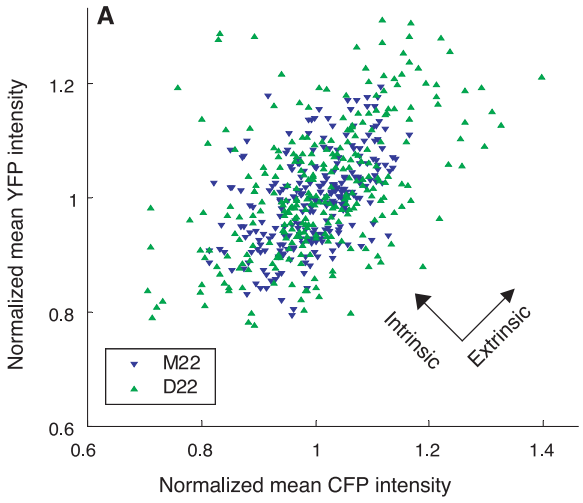

Figure 15: Measured fluorescent intensities of fluorophores from two identical promoters in single cells of two strains of *E. coli*. Taken from Elowitz, et al., *Science*, **197**, 1183–1186, 2002.

In Fig. 16, we see measured intrinsic, extrinsic, and total noise as a function of total fluorescence level. For very high copy numbers (high fluorescence), we see low noise, both intrinsic and extrinsic. At large copy numbers, noise tends to be low, since fluctuations only comprise a small fraction of the total number of molecules. At low copy number, we have higher intrinsic noise; fluctuations are more pronounced at low copy numbers. We also have higher extrinsic noise, but the extrinsic noise exhibits a maximum at intermediate copy number. This is possibly due to repression by LacI. At high copy number, we expect low noise. At very low copy number, we have very little ITPG present, so we have large copy numbers of LacI. This results in strong repression by large amount of LacI, resulting in more immunity to extrinsic fluctuations (e.g., fluctuations in the already large copy number of LacI). In between, we have moderate copy number of both the fluorescent protein and LacI, resulting in greater extrinsic noise.

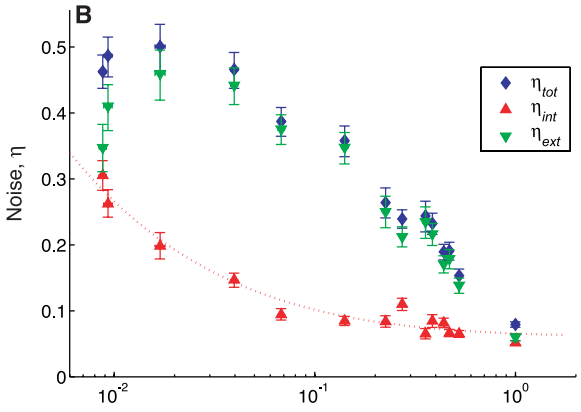

Figure 16: Measured noise in the M22 strain. The *x*-axis is the relative fluorescence level. Taken from Elowitz, et al., *Science*, **197**, 1183–1186, 2002.