Limits to Gene Regulation due to Crosstalk

Contents

Chapter 1

Introduction

1.1 Basic model

We adapt the basic model from [Friedlander et al., 2015]. In its most general form, we model the gene regulation according to the following assumptions:

- We assume N *target* genes, each regulated by a single *binding site* (BS). Each gene is activated whenever a *transcription factor* (TF) is bound to its binding site. We assume that the length of all binding sites is equal to, say, L *base pairs* (bp).
- We assume a total of T distinct transcription factors responsible for activating the genes.
- We assume a fixed set of *environments* $\mathcal{E}$, where each environment $E \in \mathcal{E}$ is a subset $E \subset [N] := \{1, \dots, N\}$ of all genes that should be activated in this environment. Moreover, we assume some fixed probability distribution P on $\mathcal{E}$ assigning probabilities to the different environments.
- To achieve the regulation we assume that depending on the environment E, the transcription factors are present in concentrations
$$C_1(E), C_2(E), \dots, C_T(E).$$
- We assume that the energy of a single binding site $i \in [M]$ bound to some TF $j \in [T]$ is equal to $\epsilon d_{i,j}$ where $\epsilon > 0$ is the per-nucleotide binding energy and $d_{i,j}$ is the number of mismatches between the binding site and the TF. The unbound state of a BS is assumed to have energy $E_a > 0$.
- We call a binding with zero mismatches a *cognate* binding. Non-perfect matches are called *noncognate* bindings.
- We employ a thermodynamic model to calculate the equilibrium binding probabilities of cognate and noncognate bindings in a fixed environment. That is, the probability of finding gene i in an erroneous state is assumed to be
$$x_i(E) := \begin{cases} \frac{e^{-E_a}}{e^{-E_a} + \sum_j C_j(E) e^{-\epsilon d_{i,j}}} & \text{if } i \in E \\ \frac{\sum_j C_j(E) e^{-\epsilon d_{i,j}}}{e^{-E_a} + \sum_j C_j(E) e^{-\epsilon d_{i,j}}} & \text{if } i \notin E \end{cases}$$
and the joined probabilities of multiple genes are assumed to be independent of each other. We emphasize that this definition of error differs from the model considered in [Friedlander et al., 2015] in the sense that the correct activation of a gene by a noncognate TF was considered as erroneous therein.

- The goal of this model is to estimate the so called *crosstalk*. We define the crosstalk $X = X(E)$ as the expected proportion of genes in an erroneous state according to the thermodynamic binding probabilities. Since the binding states of individual genes are independent from each other, the expected number of genes in an erroneous state is just the sum of the probabilites that some gene is in an erroneous state, i.e.,

$$X(E) = \frac{1}{N} \sum_{i \in [N]} x_i(E).$$

  We will be mainly interested in the crosstalk averaged over the environments, that is

$$\mathbf{E}\, X = \sum_{E \in \mathcal{E}} P(E) X(E).$$

The model considered in [Friedlander et al., 2015] is a specialization of this, besides the consideration of rightful activation by noncognate TFs as an error. Specifically in their "base" model they assumed

- a one-to-one correspondence between binding sites and cognate transcription factors, in particular $T = N$.
- environments of the form $\mathcal{E} = \binom{[N]}{Q}$, i.e., all Q-element subsets of $[N]$, equipped with a uniform probability distribution $P(E) = \binom{N}{Q}^{-1}$.
- equal concentration $C_i(E) = C > 0$ for all $i \in E$ and $C_i = 0$ otherwise.

As a first step, let us try to understand how precise the approximated analytical solution from [Friedlander et al., 2015] is to numerical reality. We therefore still assume a one-to-one correspondence between binding sites and transcription factors. We shall also assume that the concentration of the transcription factors is stoichiometric in the sense that $C_i(E) = C/|E|$ for $i \in E$, i.e. all TFs corresponding to genes that should be active, and $C_i(E) = 0$ for all other TFs. Thus the summation over the thermodynamic weights simplifies to

$$\sum_j C_j(E) e^{-\epsilon d_{i,j}} = \frac{C}{|E|} \sum_{j \in E} e^{-\epsilon d_{i,j}}$$

It is now assumed that the binding site mismatches $d_{i,j}$ are distributed symmetrically in the sense that the above expression is roughly independent from i. Therefore we can approximate it by

$$\frac{C}{|E|} \sum_{j \in E} e^{-\epsilon d_{i,j}} \approx \begin{cases} \frac{C}{|E|} |E| \sum_d P(d) e^{-\epsilon d} =: CS & \text{if } i \notin E \\ \frac{C}{|E|} + \frac{C}{|E|} (|E|-1) \sum_d P(d) e^{-\epsilon d} =: \frac{C}{|E|} + \frac{C}{|E|} (|E|-1) S & \text{if } i \in E \end{cases}$$

where $P(d)$ is the distribution of binding site mismatches and S is a new parameter, which we call the *average binding site similarity*. The simplified expression for the probability of finding gene i in an erroneous state then reads

$$x_i'(E) := \begin{cases} \frac{e^{-E_a}}{e^{-E_a} + CS + C(1-S)/|E|} & \text{if } i \in E \\ \frac{CS}{e^{-E_a} + CS} & \text{if } i \notin E \end{cases}.$$

It should be emphasized that these two expressions are not the averages of $x_i(E)$ over all possible choices of enviroments of some fixed size. Indeed, as Figure 1.1 shows, $x_i(E)$ averaged over all $i \notin E$ and 1000 random $E \subset [N]$ with $|E| = Q$ is pretty far away from

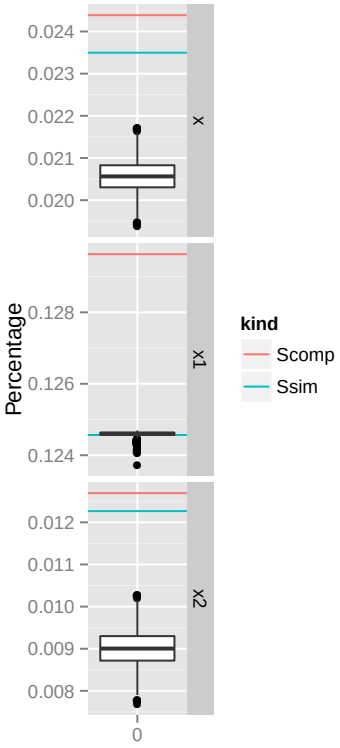


FIGURE 1.1: Numerical Computation of $X(E)$, $x_1(E) := \frac{1}{|E|}\sum_{i\in E} x_i(E)$ and $x_2(E) := \frac{1}{N-|E|}\sum_{i\notin E} x_i(E)$ for 1000 random choices of $E \subset [N]$ with $L = 10$, $N = 5000$, $|E| = 500$, $E_a = 10$, $\epsilon = 3$ and uniformly sampled binding codes, where S_{comp} is the theoretical binding site similarity $S_{comp} = \left(\frac{1}{4} + \frac{3}{4}e^{-\epsilon}\right)^L$ and S_{sim} is the actual value obtained from the sampled binding codes.

the value obtained through the approximation above. To understand why this happens let us write out $x_i(E)$ averaged over all $i \notin E$, averaged over all E with $|E| = Q$ fixed:

$$\mathbf{E}_E \frac{1}{N-|E|}\sum_{i\notin E} x_i(E) = 1 - \binom{N}{Q}^{-1} \sum_{E\subset[N],|E|=Q} \frac{1}{N-Q}\sum_{i\notin E} \frac{1}{1+e^{E_a}\frac{C}{Q}\sum_{j\in E} e^{-\epsilon d_{i,j}}}$$

On the other hand, if we just closed our eyes and exchanged the averaging and the division we would find

$$1 - \frac{1}{1+e^{E_a}\frac{C}{Q}\binom{N}{Q}^{-1}\sum_{E\subset[N],|E|=Q}\frac{1}{N-Q}\sum_{i\notin E}\sum_{j\in E} e^{-\epsilon d_{i,j}}}$$
$$= 1 - \frac{1}{1+e^{E_a}\frac{C}{(N-Q)Q}\binom{N}{Q}^{-1}\sum_{i\neq j}\sum_{E\subset[N],i\notin E\ni j} e^{-\epsilon d_{i,j}}} = 1 - \frac{1}{1+e^{E_a}\frac{C}{N(N-1)}\sum_{i\neq j} e^{-\epsilon d_{i,j}}}$$

which is nothing but $\frac{CS}{e^{-E_a}+CS}$, i.e., precisely the approximation from above. The problem here is, that the error made by exchanging summation and division is not negligible since $\frac{1}{Q}\sum_{j\in E} e^{-\epsilon d_{i,j}}$ is not sharply concentrated around its mean and the unlikely big values have an unproportional big effect on the result. On the other hand the, for the error of the first kind we have approximately $\frac{e^{-E_a}}{e^{-E_a}+CS+C(1-S)/Q}$ where CS is very small compared to $C(1-S)/Q$ and therefore the the resulting probabilites are indeed rather sharply concentrated. This effect is illustrated in Figure 3. The plots from Figure 1.2 show histograms of

$$\sum_{j\in E} e^{-\epsilon d_{i,j}}, \frac{1}{1+e^{E_a}\frac{C}{Q}\left(1+\sum_{j\in E} e^{-\epsilon d_{i,j}}\right)} \quad \text{and} \quad 1-\frac{1}{1+e^{E_a}\frac{C}{Q}\sum_{j\in E} e^{-\epsilon d_{i,j}}}$$

for randomly chosen $E \subset [N]$ and $i \notin E$, together with correctly applied mean and the mean just applied in the denuminator.

To work out a more precise result we have to compute the expectation of

$$1-\frac{1}{1+e^{E_a}\frac{C}{Q}\sum_{j\in E} e^{-\epsilon d_{i,j}}}.$$

This is not feasible analytically but the paticular distribution of $\sum_{j\in E} e^{-\epsilon d_{i,j}}$ allows a good approximation. The basic idea is to compute the conditional expectation conditioned on the number of occurences of 0's and 1's beneath the $d_{i,j}$'s. The next highest order terms would be $e^{-2\epsilon}$ which should be small compared to 1 and can reasonably well be approximated by its conditional mean. Applied to our concrete example this would mean that we average the five spikes and then compute the above expectation according to this simpler distribution.

1.2 Ising model inspired environment distribution

We shall go in a slightly different direction now and assume a Ising model inspired distribution on the environments. Concretely, we assume a distribution of the form

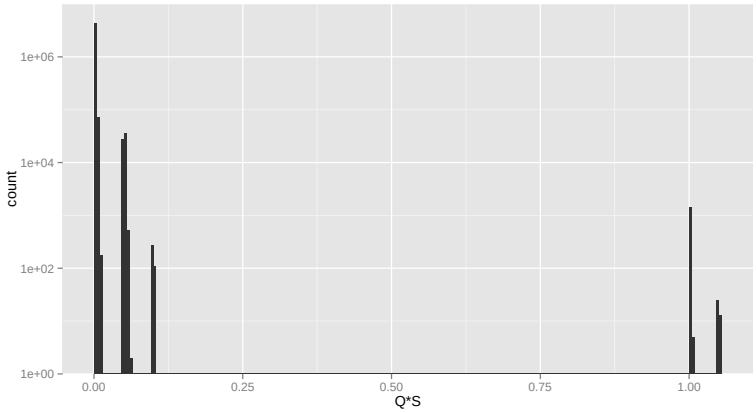
$$P(E) := \frac{\exp\left(\sum_{i=1}^N h_i \mathbb{1}_E(i)\right)}{\sum_{\delta\in\{0,1\}^N}\exp\left(\sum_{i=1}^N h_i\delta_i\right)}$$

for some parameters $h_1,\dots,h_N$. These parameters have a natural interpretation in terms of the proportion of environments in which a certain gene is active. Concretely,

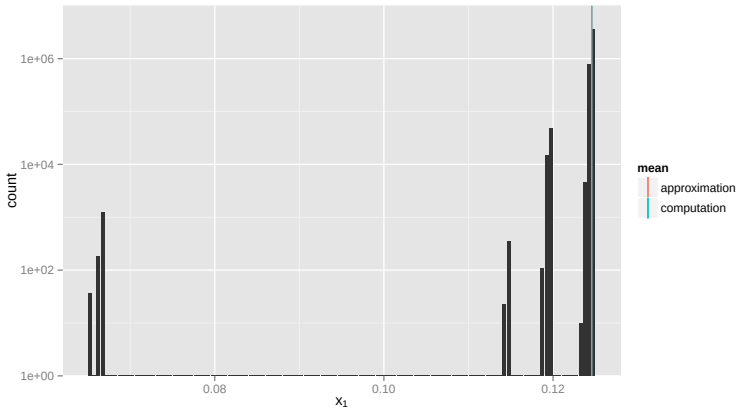
$$\begin{aligned} P(\text{Gene } i \text{ is active}) = \mathbf{E}_E\, \mathbb{1}_E(i) &= \frac{\sum_{E\subset[N]} \mathbb{1}_E(i)\exp\left(\sum_{i=1}^N h_i\mathbb{1}_E(i)\right)}{\sum_{\delta\in\{0,1\}^N}\exp\left(\sum_{i=1}^N h_i\delta_i\right)} \\ &= \frac{\sum_{\epsilon\in\{0,1\}^N,\epsilon_i=1}\exp\left(\sum_{i=1}^N h_i\epsilon_i\right)}{\sum_{\delta\in\{0,1\}^N}\exp\left(\sum_{i=1}^N h_i\delta_i\right)} = \frac{1}{1+\exp(-h_i)} =: \mu_i. \end{aligned}$$

This means, that to realize a gene in this model which should be active in $\mu_i \in (0,1)$ of the environments, this can be realized by choosing $h_i = \log\left(\frac{\mu_i}{1-\mu_i}\right)$.

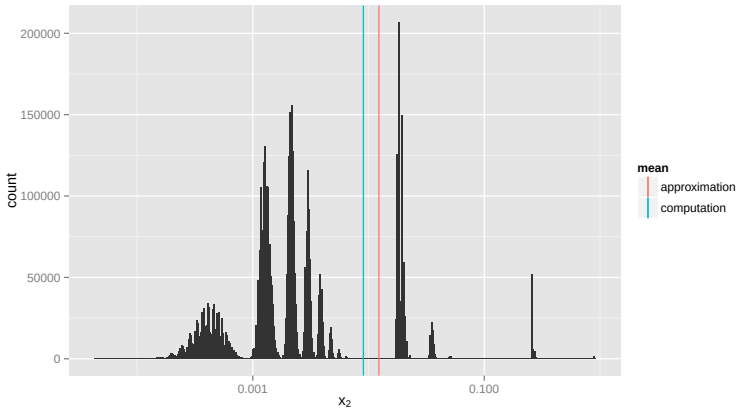
An analytic derivation of $\mathbf{E}\, X'$ is obviously only possible in the case $h_1 = \cdots = h_N = h$ since in this case the number of active genes is binomially distributed with parameter



(a) QS



(b) Expected error of erroneous non-activation



(c) Expected error of erroneous activation

Figure 1.2: Histograms

$\mu = \frac{1}{1+\exp(-h)}$. The expression for $\mathbf{E}\, X'$ then reads

$$\begin{aligned}\mathbf{E}\, X' &= \frac{1}{N}\sum_{n=0}^{N} \mu^n (1-\mu)^{N-n} \binom{N}{n} \left(\frac{n^2 e^{-E_a}}{n e^{-E_a} + nCS + C(1-S)} + \frac{(N-n)CS}{e^{-E_a} + CS} \right) \\ &= \frac{e^{-E_a}(1-\mu)^N \, {}_2F_1\left(-N, \frac{C(1-S)}{CS+e^{-E_a}}; \frac{C+e^{-E_a}}{CS+e^{-E_a}}; \frac{\mu}{\mu-1}\right)}{NC(1-S)} + 1 - \mu - \frac{1-\mu}{1+CSe^{-E_a}}.\end{aligned}$$

Bibliography

[Friedlander et al., 2015] Friedlander, T., Prizak, R., Guet, C. C., Barton, N. H., and Tkačik, G. (2015). Intrinsic limits to gene regulation by global crosstalk. *ArXiv e-prints*.