Computational Biology I: Quantitative Data Analysis
Frühjahrsemester 2019

Exercise 9

Athos Fiori and Pascal Grobecker

How to submit the exercise

Submit your completed `Exercise9.py` file by email to biocomp1-bioz@unibas.ch. The submission deadline is Thursday, May 10th, midnight.

Maximum entropy and transcription factor binding

The binding of transcription factors to DNA is an important factor in the regulation of gene expression. In the following exercise we take expression measurements from two genes whose promoters contain binding sites for three transcription factors, and attempt to infer the activity levels of these transcription factors. It is immediately obvious that if we assume a linear model to relate the transcription factor activity to the gene expression level that this is an under-determined problem - we cannot infer three parameters from two measurements without further constraints in our model. However, we can use the principle of maximum entropy to derive a set of expressions which tell us how our estimates for each parameter is related to our estimates for the other two.

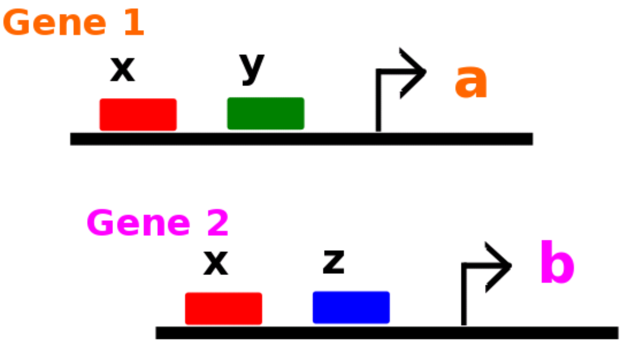

Figure 1: Schematic of the two gene system. Gene 1 is regulated by factors x and y to produce the gene product a (mRNA of gene 1), while gene 2 is regulated by x and z to express mRNA b.

1. To begin with, we will assume that in log-space the expression level of gene 1 is $a = x + y$ (the sum of the concentrations of transcription factors X and Y), and the expression level of gene 2 is $b = x + z$ (the sum of the concentrations of transcription factors X and Z)

   that leads to two constraints for our data:

$$f_1(x, y, z) = x + y \tag{1}$$
$$f_2(x, y, z) = x + z \tag{2}$$

   and

$$\langle a \rangle = \langle f_1(x, y, z) \rangle \tag{3}$$
$$\langle b \rangle = \langle f_2(x, y, z) \rangle \tag{4}$$

   from the slides you know that the maximum entropy distribution takes the form

$$p(x, y, z) = \frac{\exp(-\sum_{k=1}^{n} \lambda_k f_k(x, y, z))}{Z(\lambda_1, \lambda_2, ..., \lambda_k)} \tag{5}$$

   with

$$Z(\lambda_1, \lambda_2, ..., \lambda_k) = \int_x \exp(-\sum_{k=1}^{n} \lambda_k f_k(x)) \tag{6}$$

   Note that in the lecture was a sum instead of the integral used to go over all x. Here we use the integral as our x are continuous.

   Write a function Z that takes λ_1 and λ_2 as parameters and returns the value of the partition function Z. *Hint*: You may need the integral: $\int_0^\infty e^{-ax} = 1/a$.

2. Now we need to find λ_1 and λ_2 in terms of the measured a and b. Given that

$$-\frac{\partial \log Z}{\partial \lambda_1} = \langle a \rangle, \tag{7}$$
$$-\frac{\partial \log Z}{\partial \lambda_2} = \langle b \rangle, \tag{8}$$

   (a) write down equations for $\langle a \rangle$ and $\langle b \rangle$ in terms of λ_1 and λ_2.

   (b) use the substitution $p = \frac{1}{\lambda_1}$ and $q = \frac{1}{\lambda_2}$ and show that q satisfies the following quadratic equation:

$$3(\langle a \rangle - \langle b \rangle)q^2 + (2\langle a \rangle^2 - 6\langle a \rangle\langle b \rangle + 4\langle b \rangle^2)q - \langle b \rangle\langle a \rangle^2 + 2\langle a \rangle\langle b \rangle^2 - \langle b \rangle^3 = 0 \tag{9}$$

   (c) derive a similar quadratic equation for p. *Hint*: The easiest way to do this is to notice that your initial equations are symmetric upon simultaneously swapping p with q and $\langle a \rangle$ and $\langle b \rangle$, so the solutions to these equations must obey the same symmetry.

   (d) hence show that the only biologically meaningful solution to these equations is as follows:

$$\lambda_1 = \frac{3}{2\langle a \rangle - \langle b \rangle + \sqrt{\langle a \rangle^2 - \langle a \rangle\langle b \rangle + \langle b \rangle^2}} \tag{10}$$
$$\lambda_2 = \frac{3}{2\langle b \rangle - \langle a \rangle + \sqrt{\langle a \rangle^2 - \langle a \rangle\langle b \rangle + \langle b \rangle^2}} \tag{11}$$

   Explain why any other solutions to these equations are rejected.

   Write a function *lambdas* that takes as parameters *mean_a* and *mean_b* which correspond to $\langle a \rangle$ or $\langle b \rangle$ and returns a **list** of **all** the solutions for the Lagrange Multipliers.

3. Ultimately we want to infer how x, y and z regulate both genes, given our observation on the expression of both, as stated in question 2. We are now able to attack this question in a more systematic manner by fixing $\langle b\rangle$ and examining how x, y and z vary as we vary $\langle a\rangle$. For this we need to obtain the marginal distributions of x, y and z.

   Obtain the marginal distributions of x, y and z and state their mean and variance as a function of λ_1 and λ_2. A marginal distribution for x would be for example:

   $$p(x|I) = \int_0^\infty p(x,y,z|I_2)\mathrm{d}y\mathrm{d}z \tag{12}$$

   Note, that these are therefore functions of $\langle a\rangle$ and $\langle b\rangle$ themselves. *Hint*: Don't write the Lagrange multipliers explicitly in terms of $\langle a\rangle$ or $\langle b\rangle$. Keep in mind that the mean and variance for an exponentially-distributed variable x with probability density function $P(x) = \lambda e^{-\lambda x}$ would be $\frac{1}{\lambda}$ and $\frac{1}{\lambda^2}$ respectively.

   Write a function *get_means_xyz* that takes as input $\langle a\rangle$ and $\langle b\rangle$ and returns the mean of x, y, z in a list. *Hint* : In the last two functions you must obtain the values of the Lagrange multipliers as functions of $\langle a\rangle$ and $\langle b\rangle$. You can reuse the function *lambdas* defined above.

4. The equations above also allow us to easily estimate the second-order moments for $\langle a\rangle$ and $\langle b\rangle$ under our maximum-entropy model.

   Calculate the following quantities (var(a), var(b) and cov(a,b)) to get expressions in terms of λ_1 and λ_2:

   $$\frac{\partial^2 \log Z}{\partial \lambda_1^2} = \mathrm{var}(a) \tag{13}$$

   $$\frac{\partial^2 \log Z}{\partial \lambda_2^2} = \mathrm{var}(b) \tag{14}$$

   $$\frac{\partial^2 \log Z}{\partial \lambda_1 \partial \lambda_2} = \langle ab\rangle - \langle a\rangle\langle b\rangle \tag{15}$$

   and comment on your results, relating them to the variances of x, y and z.

   Write a function *get_covariance* that takes $\langle a\rangle$ or $\langle b\rangle$ as parameters and returns the covariance matrix

5. Plot the graphs of the mean of x, y and z as you vary $\langle a\rangle$. Does it make sense? How does x,y and z behave if you increase $\langle a\rangle$? *Hint*: the code for the plot is already implemented!

Theoretical questions

1. Derive the maximum entropy distribution for our case, assuming that we have no prior information about cell-to-cell variability in concentrations a and b.

2. Follow the steps in exercise 2.

3. explain why the only “biologically meaningful” solutions for λ_1 and λ_2 are those given in 2.(d).