Computational Biology I: Quantitative Data Analysis
Frühjahrsemester 2019

Exercise 10

Pascal Grobecker and Athos Fiori

Detecting dependencies between adjacent bases within human promoters

Genetic (i.e. DNA sequence) as well as epi-genetic information controls gene regulation. In mammals an important epi-genetic gene regulatory mechanism is the methylation of cytosines within CpG di-nucleotides (i.e. a cytosine nucleotide occurs next to a guanine nucleotide). Methylation of CpGs in gene promoters is used to silence gene expression. However, not every promoter within mammalian genomes is regulated by this mechanism. According to a common view, mostly house-keeping genes, i.e., genes which are expressed in a broad range of cell types, are regulated by CpG-high promoters, while many developmental genes, and genes that are specific to particular tissues, are regulated by CpG-poor promoters. However, this picture has been questioned based on recent data that derives from the increase in the number of sequenced genomes. Is it possible to distinguish different functional gene classes based on their promoter sequence?

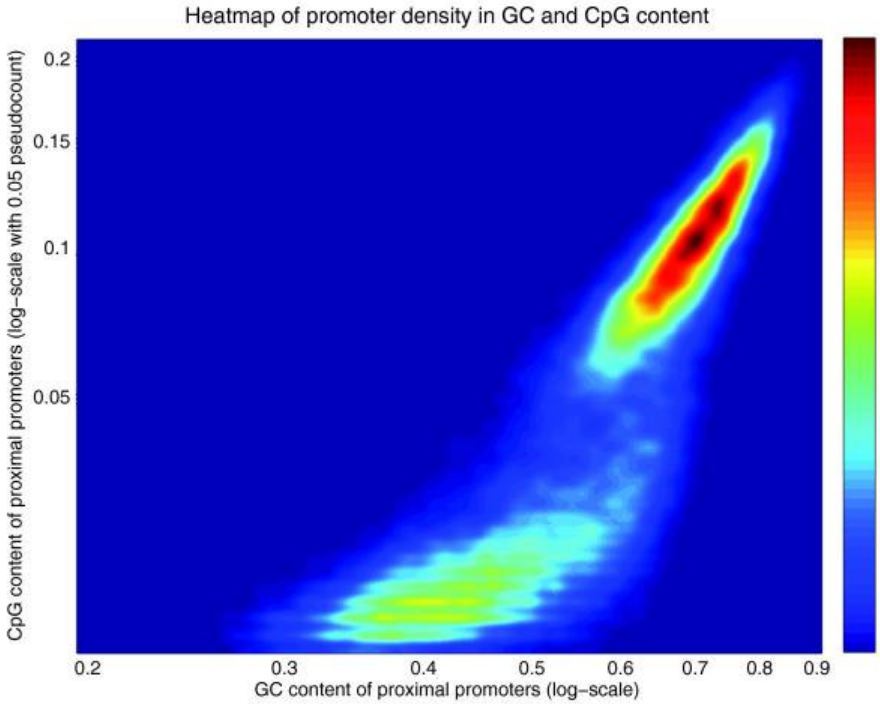

Figure 1: Two-dimensional histogram (shown as a heatmap) of the CG base content (horizontal axis) and CpG di-nucleotide content (vertical axis) of all human promoters. There appear to be two promoter classes: CpG-poor and CpG-high. From: Balwierz et al. 2009.

In this exercise we will investigate these questions by looking at the dependency between base pair positions within human promoter regions. To this end we will focus on a simple di-nucleotide model which assumes dependency between adjacent bases and contrast it with a model where every base position is independent. Can we classify human promoters according to these two models?

1. To begin with, have a look at the promoter sequences in the file `sequences.fa`. Note that the file is in

FASTA format. Each sequence is preceded by a header (identified by the `>` symbol) which contains the sequence ID.

```
>Seq_id1                # header with sequence ID
ATGC...                 # sequence
```

Complete the function `readFasta` which returns a dictionary with all sequences accessible by their ID.

2. Complete the function `countWord` which counts the number of possibly overlapping occurrences of a sub-string (i.e. word) in a string.

3. Complete the function `countMatrix` which returns a matrix containing the following four di-nucleotide counts:

		Position i	
		G or C	A or T
Position $i+1$	G or C	c_1	c_2
	A or T	c_3	c_4

E.g. the count c_1 is the number of occurrences of the di-nucleotides GpG,CpC,CpG and GpC within the input sequence. Note, that the sum of all counts $c_1 + c_2 + c_3 + c_4 = n$ is the total number of di-nucleotides in a sequence of length $n + 1$.

4. We now derive the likelihood of the data under the two models mentioned above. In the first case we assume all positions to be independent. Lets call the probability of observing a G or C at any given position ρ_{GC}. The likelihood of the count data under the independent model is

$$P(D|\rho_{GC}) = \binom{2n}{2c_1 + c_2 + c_3} (\rho_{GC})^{2c_1+c_2+c_3} (1 - \rho_{GC})^{2c_4+c_2+c_3} .$$

The likelihood of the count data under the dependent model is

$$P(D|\vec{\rho}) = \frac{n!}{c_1!c_2!c_3!c_4!} \left(\rho_{GC|GC}\right)^{c_1} \left(\rho_{GC|AT}\right)^{c_2} \left(\rho_{AT|GC}\right)^{c_3} \left(1 - \rho_{GC|GC} - \rho_{GC|AT} - \rho_{AT|GC}\right)^{n-(c_1+c_2+c_3)} .$$

Here, $\rho_{GC|AT}$ is the probability of observing a G or C at position $i + 1$ given an A or T at position i. Derive these likelihood equations yourself (answer not to be submitted)

5. In order to compare the two models we cannot use the likelihood alone. As mentioned in the lecture, the independent model is a sub-model of the dependent model. Therefore, the dependent model will always fit the data at least as well as the independent model. This implies that model comparison should evaluate if the data supports the more complex (i.e. dependent) model. This is possible by employing the *marginal* likelihood, i.e., the likelihood averaged over all parameters, for model comparison. A suitable prior for the probabilities $\vec{\rho}$ is the Dirichlet distribution

$$P(\vec{\rho}|\lambda) = \Gamma(m\lambda) \prod_{i=1}^{m} \frac{(\rho_i)^{\lambda-1}}{\Gamma(\lambda)}.$$

The marginal likelihood of the data under the independent model is

$$\begin{aligned} P(D|\text{indep}, \lambda) &= \int_{\rho_{GC}} P(D|\rho_{GC}) P(\rho_{gc}|\lambda) d\rho_{gc} \\ &= \frac{\Gamma(2\lambda)}{\Gamma(2n + 2\lambda)} \frac{\Gamma(2c_1 + c_2 + c_3 + \lambda)}{\Gamma(\lambda)} \frac{\Gamma(2c_4 + c_2 + c_3 + \lambda)}{\Gamma(\lambda)} \end{aligned}$$

where $n = \sum_i^4 c_i$. The marginal likelihood of the dependent model is

$$\begin{aligned} P(D|\text{dep}, \lambda) &= \int_{\vec{\rho}} P(D|\vec{\rho}) P(\vec{\rho}|\lambda) d\vec{\rho} \\ &= \frac{\Gamma(4\lambda)}{\Gamma(n+4\lambda)} \prod_i^4 \frac{\Gamma(c_i+\lambda)}{\Gamma(\lambda)} \end{aligned}$$

Complete the two functions `independentLML` and `dependentLML` which compute the logarithm of the marginal likelihoods of the independent and the dependent model, respectively. Note, that we will set the parameter λ, which plays the role of a pseudo-count, to 1 by default,

6. Complete the function `dependentPosterior` which calculates the posterior probability for the dependent model given the base pair counts D. If we assume both models to be equally likely *a priori*, then the posterior probability of the dependent model is

$$\begin{aligned} P(\text{dep}|D, \lambda) &= \frac{P(D|\text{dep}, \lambda)}{P(D|\text{dep}, \lambda) + P(D|\text{indep}\lambda)} \\ &= \frac{e^{\log(BF)}}{1 + e^{\log(BF)}}, \quad \text{with} \quad BF = \frac{P(D|\text{dep}, \lambda)}{P(D|\text{indep}, \lambda)}. \end{aligned}$$

Extra question: Calculate for each promoter sequence the posterior probability of the dependent model. How many promoter sequences can be better explained with the dependent model? How well does that correlate with the CpG-high/CpG-low promoter annotation? Note: each sequence ID contains a flag in the end indicating if it was classified as a CpG-high (flag=`gc`) or CpG-low (flag=`at`) promoter by an independent method. What do you think could be improved upon the model? You can submit your answers to these questions in written from, if you want to improve your overall score (by three points).

How to submit the exercise

Submit your completed `exercise10.py` file by email to biocomp1-bioz@unibas.ch. The submission deadline is Thursday, May 16nd, midnight.

1. Check that your functions are working properly before submitting them!
2. Don't forget to attach your exercise file!
3. Don't change the name of the file or the name of any function within the file!
4. Don't add any additional `import` statements within the file!
5. Please send your solutions from the same email address that you used to registering at the course.