Numerical project (Problem set 10): Inference on evolutionary data

BIO-369

Prof. Anne-Florence Bitbol
EPFL

This numerical project will be graded and count for 40% of your final grade. ***Each student*** *should hand in their* ***personal*** *solution as a* ***Jupyter notebook using Python 3****, by uploading it on* ***Moodle*** *on* ***May 9*** *at the latest. Please clearly label question numbers using markdown cells. Please answer all questions (including those requiring sentences and not code as an answer) in the same Jupyter notebook, using markdown cells for text. Please name your Jupyter notebook* `LASTNAME_FirstName.ipynb`.

Three problem classes (April 28, May 5 and May 7) will be dedicated to this project, and during them, you can ask questions to the teaching assistants and discuss with other students as usual. However, in the end, you must hand in your personal solution. Detected plagiarism will result in a reduction of your grade.

This numerical project contains two parts that are independent from one another.

1 Synthetic evolution of protein sequences

In this problem, we will study data coming from a synthetic evolution process aimed to be realistic. The starting point is a natural multiple sequence alignment, and then each sequence was artificially evolved via mutations and selection. Specifically, we will consider the data in the 4 following files: `timepoint-0.fasta`, `timepoint-1.fasta`, `timepoint-10.fasta` and `timepoint-0-b.fasta`. Each of these files contains an alignment of amino acid sequences. As you can check, there is the same number of sequences (rows) and sites (columns) in each of the 4 files.

The files `timepoint-0.fasta` and `timepoint-0-b.fasta` each contain a different sample of natural sequences from the same protein family (AAA ATPases, which act as chaperones). The files `timepoint-1.fasta` and `timepoint-10.fasta` contain sequences that were synthetically evolved starting from the sequences in the file `timepoint-0.fasta` as respective ancestors, kept in the same order (i.e., the first sequence in `timepoint-1.fasta` has for ancestor the first sequence in `timepoint-0.fasta`). The evolution process lasted for a short time in the file `timepoint-1.fasta`, and substantially longer in the file `timepoint-10.fasta`.

Our goal will be to analyze how the synthetically evolved sequences diverge from their ancestors.

a) In Python, load the file `timepoint-0.fasta` and extract all the sequences in it as an array of strings. We will not need the headers for this analysis, so you can discard them. Next, transform the array of strings you obtained into a `Numpy` array of integer numbers, preferably by using the following mapping: A is 0, C is 1, ... [use alphabetic order for standard amino acids represented by their one-letter code]..., Y is 19 and – is 20. *Note: here and throughout this project, alignment gaps (–) are considered just as an extra character.*

b) Calculate the entropy of each column of the alignment. Do the same for the files `timepoint-1.fasta` and `timepoint-10.fasta`, naming the entropy arrays differently for each file so that you can reuse them.

c) Make a plot that shows (on the same plot) the entropy versus the site (column) of the alignment for these 3 files. *Note: for better visualization, we recommend regular plots (with markers and lines) and not bar plots. Besides, to make your plot wider horizontally, you can change the aspect ratio using* `plt.figure(figsize=(a,b))` *where a and b are two numbers representing width and height, respectively.*

d) Comment on the plot you obtained: how similar or different are the overall patterns of the curves showing the entropies of columns across the three files? What does this tell us about the synthetic evolution process, regarding entropy?

e) Produce a `Numpy array` that contains the mutual information between each column i in the data from `timepoint-0.fasta` and the same column i in the data from `timepoint-1.fasta`. Do the same for the mutual information between each column i in the data from `timepoint-0.fasta` and the same column i in the data from `timepoint-10.fasta`.

f) Make a plot that shows (on the same plot) the mutual information versus the site (column) of the alignment for these 2 pairs of files (one curve for times 0-1 and one for times 0-10). *Note: see question c) for style recommendations.*

g) Comment on the plot you obtained: based on these timepoints, how does the mutual information between a column at time 0 and the same column at time t generally vary with time t? Why is it the case?

h) Now use the file `timepoint-0-b.fasta`, and produce a `Numpy array` that contains the mutual information between each column i in the data from `timepoint-0.fasta` and the same column i in the data from `timepoint-0-b.fasta`. Calculate the mean of these mutual information values.

i) Recall that these two files contain different natural sequences that are homologous (i.e., from the same protein family). What do you *a priori* expect for the mutual information values calculated in the previous question? Why? Do the results align with that expectation?

j) Estimate the theoretical magnitude of (leading order) finite size effects on the mutual information values that you calculated in the question before. Compare it with the average value of mutual information over columns, and comment on the result. Given the number of sequences available in these files, how would you propose to modify the theoretical formula to estimate these (leading order) finite size effects? (*Hint: Consider how the number of observable states is impacted by the number of available sequences.*) How do these finite size effects help to interpret the results of the previous question? More generally, what can be misleading about uncorrected mutual information values estimated from limited numbers of samples?

k) Make a plot that shows (on the same plot) the mutual information versus the site (column) of the alignment for the following 3 pairs of files: one curve for times 0-1, one for times 0-10 (as before), and one for the two samples at time 0 (time 0 and time 0-b). *Note: see question c) for style recommendations.*

l) Comment on the plot: overall, how do the mutual information values between timepoints 0 and 1 from synthetic evolution compare to those between the two natural samples? Same question for the mutual information values between timepoints 0 and 10. What does this tell us about the synthetic evolution process?

2 Fitness effects of beneficial mutations

If the reference fitness of the wild-type organism is 1, then the fitnesses of a beneficial mutant can be written as $1 + s$, were $s > 0$ is the relative fitness effect of the mutation.

a) The file `fitness_effects.csv` contains a list of relative fitness effects of beneficial mutations that could be observed starting from a given organism. Load the data and plot it in a histogram. Comment on the histogram.

We would like to model the probability distribution of relative fitness effects s by a gamma distribution, which is a way to generalize over the exponential distribution. The gamma distribution has the following probability density:

$$p(s) = \frac{s^{\alpha-1}\beta^{\alpha}e^{-\beta s}}{\Gamma(\alpha)}, \tag{1}$$

where α and β are two parameters of the model, and Γ is the Euler Gamma function (`math.gamma` in Python).

b) Assuming that fitness effects are independent from one another, write the likelihood of observing N fitness effects $s_1, \ldots, s_N$ under the model given by the gamma distribution. Express its logarithm.

c) To find the gamma distribution that best describes the data, we would like to estimate α and β using the maximum likelihood approach. In principle, what derivative(s) of the log-likelihood should we compute to do this? What should be held constant in this process? Why?

d) Calculate the derivative of the log-likelihood with respect to β, and express the maximum-likelihood estimate of β as a function of α and of the s_i. What simple quantity involving the s_i do you recognize in this expression?

e) By using your result from the question above, express the log-likelihood as a function of α and of the s_i (but not β). Divide it by the number N of measurements to obtain a per-measurement log-likelihood that we will call $\mathcal{L}$. Show that it can be written as $\mathcal{L} = \alpha \ln(\alpha) - \alpha \ln(\langle s_i \rangle) - \ln(\Gamma(\alpha)) + (\alpha - 1)\langle \ln(s_i) \rangle - \alpha$, and provide the explicit definitions of $\langle s_i \rangle$ and $\langle \ln(s_i) \rangle$.

f) Plot $\mathcal{L}$ versus α for 50 values of α linearly spaced between 0.1 and 2. Comment on the variations of $\mathcal{L}$ with α.

g) Determine the maximum likelihood value of α with two significant digits. Using your previous results, also determine the maximum likelihood value of β with two significant digits.

h) Using the Python fitting function `stats.gamma.fit` from `scipy`, propose another estimate for α and for β. (*Note: `stats.gamma.fit` outputs 3 parameters, which represent respectively α, "loc", and β, where "loc" is a shift we will not discuss here.*) Compare them with your own maximum-likelihood estimates, and comment.

Beyond the distribution of fitness effects, an interesting question is how beneficial mutations accumulate in an evolving population. For the last questions, we will consider data from a long evolution experiment performed on *Escherichia coli* bacteria. The relative fitness of an evolved strain with respect to a reference ancestral strain can be obtained by growing them together in a competition experiment. The number of bacteria N of the evolved strain and N_{anc} of the ancestral strain are determined at the beginning (t_0) and at the end (t_f) of a growth phase. We will admit that the relative fitness advantage of the evolved strain compared to the ancestral one can be expressed as:

$$s = \frac{1}{t_f - t_0} \ln\left(\frac{N(t_f)\ N_{anc}(t_0)}{N_{anc}(t_f)\ N(t_0)}\right), \tag{2}$$

and that $t_f - t_0 = 7$ generations. *Note: this time should be kept in generations, and should be kept constant at 7 throughout. In particular, it is different from the "generation" column in the data which indicates how much the evolved strain considered in a given competition experiment has evolved before that competition experiment is started.*

The files `LTEE_1.csv`, `LTEE_2.csv` and `LTEE_3.csv` each contain the results of competition experiments conducted between the ancestral strain (the same one in each row) and evolved strains taken at various times of the experiment (one time point per row).

i) Separately for each of the 3 files, extract the data and compute for each time point the relative fitness advantage of the evolved strain compared to the ancestral one using Eq. 2.

j) Make a plot that shows (on the same plot) the relative fitness advantage of the evolved strain versus time (indicated in the "generation" column in the data) in each of the 3 files considered.

k) Comment on the plot: how does fitness evolve overall during the evolution experiment? Does the trend become stronger or weaker as time goes by?

l) Out of the 3 files considered here, 2 of them are coming from different competition experiments performed from the same evolved strains (same evolution experiments), and can be considered biological replicates of the measurement of the same fitness values. Meanwhile, one files comes from a different evolution experiment (different evolved strains, but they evolved in the same conditions). Which file do you think is from a different evolution experiment? Comment on the precision of fitness measurements via competition experiments. Also comment on the reproducibility of evolution in the same conditions.