Exam preparation solution

BIO-369, Randomness and information in biological data

Prof. Anne-Florence Bitbol
EPFL

1 Luria-Delbrück experiment

a) Each bacteria exposed to phage has a probability μ of becoming resistant and $1-\mu$ of remaining sensitive to phage under the Lamarckian hypothesis. The probability distribution of the random variable X characterizing whether one bacteria is resistant or not is a Bernoulli distribution with parameter μ.

b) In a culture containing N bacteria that is exposed to phage, the probability that m of them become resistant under the Lamarckian hypothesis is given by the binomial distribution:

$$P(m) = \binom{N}{m} \mu^m (1-\mu)^{N-m} \,. \tag{1}$$

c) If $\mu \ll 1$ and $N \gg 1$ but $N\mu = \lambda$ is of order one, this distribution becomes the Poisson distribution with parameter $\lambda = N\mu$,

$$P(m) = e^{-\lambda} \frac{\lambda^m}{m!} \,. \tag{2}$$

The mean value of m reads:

$$\langle m \rangle = \sum_{m=0}^{\infty} m \frac{\lambda^m e^{-\lambda}}{m!} = \sum_{m=1}^{\infty} m \frac{\lambda^m e^{-\lambda}}{m!} = \sum_{m=1}^{\infty} \frac{\lambda^m e^{-\lambda}}{(m-1)!} = \lambda e^{-\lambda} \sum_{m=1}^{\infty} \frac{\lambda^{m-1}}{(m-1)!} \,. \tag{3}$$

Introducing $k = m-1$, we obtain

$$\langle m \rangle = \lambda e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = \lambda e^{-\lambda} e^{\lambda} = \lambda \,, \tag{4}$$

where we have used the series expansion of the exponential function

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!} \,. \tag{5}$$

In addition, for a Poisson distribution, the variance is equal to the mean, so it is also equal to λ.

d) Under the Darwinian hypothesis where phage resistance mutations can be acquired randomly at any time during growth of the bacterial cultures (before exposure to phage), if a mutation occurs early in the growth process, then the culture will comprise many mutants as all descendants of the mutant are mutant too. Thus, the number of colonies with large numbers of mutants (albeit small) is much larger under the Darwinian hypothesis than under the Lamarckian one. These events are called "jackpots".

e) If the number m of resistant bacteria is counted in many replicate cultures, we can use the maximum likelihood method to quantitatively compare the two hypotheses and decide which of these two models best describes the data. It amounts to comparing the probability that the data actually observed can be obtained under the two hypotheses. In practice, we would study the (logarithm of the) ratio of likelihoods for these two models,

$$\log \left[\frac{P(\text{data}|\text{model})}{P(\text{data}|\text{model}')} \right] \,. \tag{6}$$

2 Protein sequences

a) We can compute the covariance from such data using the definition of covariance, $\mathrm{cov}(X_i, X_j) = \langle x_i x_j \rangle - \langle x_i \rangle \langle x_j \rangle$. To construct the covariance matrix of sites, we can employ this definition, where X_i is column i (site i) of the data. This yields a matrix C with L rows and L columns and elements

$$C_{ij} = \mathrm{cov}(X_i, X_j) = \langle x_i x_j \rangle - \langle x_i \rangle \langle x_j \rangle \,, \tag{7}$$

where averages denoted by $\langle . \rangle$ are over all sequences (all rows) of the alignment.

b) Sequences are rows of the alignment. To construct the covariance matrix of sequences, we can employ the definition of covariance, where Y_i is row i (site i) of the data. This yields a matrix M with N rows and N columns and elements

$$M_{ij} = \mathrm{cov}(Y_i, Y_j) = \langle y_i y_j \rangle - \langle y_i \rangle \langle y_j \rangle \,, \tag{8}$$

where averages denoted by $\langle . \rangle$ are over all sites (all columns) of the alignment.

c) If we applied PCA to the covariance matrix of sites, and used this to represent the data in two dimensions as we did in the lectures and in problem series 11, each marker in the plot would represent a sequence. Looking at this data, we may identify clusters of sequences that tend to be correlated and to have common points.

d) If we applied PCA to the covariance matrix of sequences, and used this to represent the data in two dimensions as we did in the lectures and in problem series 11, each marker in the plot would represent a site. Looking at this data, we may identify clusters of sites that tend to be correlated and to have common points.

3 Evolution and SARS-CoV-2 variants

a) In the Moran model, the number i can take values from 0 to N. Upon each step, the number i can either:

- increase by 1, if an individual of type A divides while a B dies;
- decrease by 1, if an individual of type B divides while an A dies;
- stay constant, if the individual that divides and the one that dies are of the same type.

b) At each time step, i either increases or decreases by one, or stays constant, with given probabilities. Thus it is a rugged trajectory with steps of size one up or down, and plateaus. Such a trajectory is a random walk (biased by fitness differences between A and B). A protein diffusing in the cell also undergoes a random walk due to the multiple hits of smaller molecules that constantly move due thermal fluctuations. The global behavior is similar to the Moran process but there are differences, especially the origin of randomness. For the protein in the cell, it arises from thermal fluctuations while for the Moran process it arises from the finite size of the system and the randomness in the birth-death process. (Another difference is dimension, the Moran process is 1D, the protein in the cell performs a 3D random walk.) A bacteria performing chemotaxis also undergoes a random walk, etc.

c) The fitness difference between A and B biases the random walk, e.g. if A has a higher fitness than B, then it will be more likely that i increases than that it decreases. After a sufficiently long amount of time, either A or B will take over and fix in the population. The fixation probabilities can be computed analytically. The two types cannot coexist indefinitely. Fixation probabilities are impacted by fitnesses, e.g. if A has higher fitness than B, then the fixation probability of A will be larger than if A has lower fitness than B (for the same N and initial i).

d) In Fig. 1, a mutation occurred each time a new variant appeared. Mutation appearance is not described in the framework of the Moran process. Mutations affect the composition of the population by increasing diversity, while in the Moran process diversity can only decrease (fixation, cf. before).

e) In Fig. 1, the frequency of variant Alpha on 2020-11-21 was roughly 0.12. In February 2021 it had substantially increased to about 0.9. This time evolution may mean that variant Alpha is fitter than the other strains existing during this time period. The trajectory of the frequency of variant Alpha is quite different from that of the variants that were present on 2020-08-09, as it grew and became dominant while those earlier variants never became dominant and decayed. Those variants may not have had substantial fitness advantages, contrary to variant Alpha.

f) On 2021-05-17, variant Delta had frequency 0.73, and variant Alpha had frequency 0.26.

g) The definition of entropy is

$$H(X) = -\sum_{x\in\Omega} P(x)\log_2[P(x)]\,, \tag{9}$$

where Ω is the set of possible values x that can be taken by the random variable X, and where $P(x)$ is the probability that X takes the value x. Here, frequencies are used instead of probabilities, which is appropriate in the limit of large populations.
On 2020-08-17, the entropy value was 1.74 bit. On 2021-05-17, it was 0.90 bit. For a population of viruses, entropy quantifies diversity. Consistently, the higher value on 2020-08-17 reflects the higher diversity at that time point, which was then lost as variants Alpha and Delta progressed through the population.

4 Diagnosing the strep throat disease

The population can be partitioned in two different ways:

- people who are sick with strep throat (s) and people who are healthy (h),
- people who are tested positive (p) and people who are tested negative (n).

a) The sensitivity α is the probability that a person is tested positive given that they are really sick. Thus, we can write

$$\alpha = P(p|s)\,, \tag{10}$$

which is the probability that one is tested p knowing that one is s. Similarly, the specificity β is the probability that a person is tested negative given that they are really healthy, and we have

$$\beta = P(n|h)\,, \tag{11}$$

which is the probability of being tested n conditioned being h.
For the throat swab, their values are $\alpha = 0.7$ and $\beta = 0.9$.

b) We know $P(s)$, $\alpha = P(p|s)$ and $\beta = P(n|h)$, and we want to calculate $P(s|p)$. Using Bayes' theorem, we can write

$$P(s|p) = \frac{P(p|s)\,P(s)}{P(p)}\,. \tag{12}$$

But, because an individual tested positive can be either sick or healthy (no third choice), we can write

$$P(p) = P(p,s) + P(p,h) = P(p|s)P(s) + P(p|h)P(h) = P(p|s)P(s) + P(p|h)\,[1-P(s)]\,. \tag{13}$$

Using Eq. 13, Eq. 12 becomes

$$P(s|p) = \frac{P(p|s)\,P(s)}{P(p|s)P(s) + P(p|h)\,[1-P(s)]}\,. \tag{14}$$

Because an individual that is healthy is either tested positive or negative (no third choice), we have

$$P(p|h) = 1 - P(n|h)\,, \tag{15}$$

and thus

$$P(s|p) = \frac{P(p|s)\,P(s)}{P(p|s)P(s) + [1-P(n|h)]\,[1-P(s)]}\,. \tag{16}$$

Using $\alpha = P(p|s)$ and $\beta = P(n|h)$, this yields

$$P(s|p) = \frac{\alpha\, P(s)}{\alpha P(s) + [1-\beta]\,[1-P(s)]}\,. \tag{17}$$

With $\alpha = 0.7$, $\beta = 0.9$ and $P(s) = 0.8$, this gives $P(s|p) = 0.97$. This value is quite high, and much higher than that before the test result was known (0.8): the test has improved our knowledge, and now it sounds OK to prescribe antibiotics (only 3% chance of them being useless), but still the possibility of mis-diagnosing is not completely negligible.

c) Using the definition of conditional probability and Bayes' theorem, we can write

$$P(t_1, t_2|s) = \frac{P(t_1, t_2, s)}{P(s)} = \frac{P(t_1|s, t_2)P(s, t_2)}{P(s)} = P(t_1|s, t_2)P(t_2|s)\,. \tag{18}$$

d) Because successive test results (performed on the same patient) can be considered independent from one another, we have $P(t_1|s, t_2) = (t_1|s)$: indeed, t_2 gives no information on t_1 (but s does). Thus, Eq. 18 becomes

$$P(t_1, t_2|s) = P(t_1|s)P(t_2|s)\,. \tag{19}$$

e) Using Eq. 19, we can write

$$P(p, n, p, n, p|s) = P(p|s)^3 P(n|s)^2 = \alpha^3(1-\alpha)^2 = 0.031\,. \tag{20}$$

The probability of obtaining these (extremely mixed!) test results is rather small if the patient is sick (3.1%).

f) Similarly, using Eq. 19, we can write

$$P(p, n, p, n, p|h) = P(p|h)^3 P(n|h)^2 = (1-\beta)^3\beta^2 = 0.00081\,. \tag{21}$$

The probability of obtaining these test results is much smaller if the patient is healthy (0.081%) than if the patient is sick. This is because β is closer to 1 than α.

g) Denoting by t the test results ($t = p, n, p, n, p$) and using Bayes' theorem, we can write as before

$$P(s|t) = \frac{P(t|s)\,P(s)}{P(t)} = \frac{P(t|s)\,P(s)}{P(t|s)P(s) + P(t|h)\,[1-P(s)]}\,. \tag{22}$$

In the previous questions, we calculated $P(t|s)$ and $P(t|h)$, and we know $P(s)$. Then, we obtain: $P(s|t) = 0.993$. Thus, despite the apparently mixed test results, these tests provided a lot of information and essentially allow us to say that the patient is sick. We should definitely recommend the antibiotic treatment at this point.

h) Here we need to redo the same calculation as above but with $P(s) = 0.5$ instead of 0.8. In this case, we obtain $P(s|t) = 0.97$, which is quite large too. However, the probability that the patient is not sick is far less negligible, and this is due to our less certain prior.

5 Analyzing neuroscience data

5.1 Activity of one neuron

a) Based on Fig. 2, and assuming that activity is never more frequent than in the right panel, the discretized data is not going to be binary if we choose $\Delta\tau = 100$ ms, but it is going to be binary if we choose $\Delta\tau = 5$ ms. Indeed, the smallest time intervals between spikes appear to be of the order of 0.025 s or 25 ms on the right panel of Fig. 2.

b) The definition of entropy applied to binary data reads

$$H(X) = -\sum_{x\in\{0,1\}} P(x)\log_2[P(x)] = -p\log_2(p) - (1-p)\log_2(1-p)\,, \tag{23}$$

where we denote by p the probability that the neuron is active in the time bin considered. The maximum possible value that this entropy can take is 1 bit, and it is obtained for $p = 1/2$, i.e. in the case of the uniform distribution. The minimum possible value is 0, obtained for $p = 0$ and $p = 1$, i.e. when one outcome is certain.

c) Using the formula above, we obtain for $p = 110/120$ an entropy of 0.41 bit, and we obtain the exact same result for $p = 10/120$. The results are the same because of the symmetry of the above formula when p and $1 - p$ are exchanged.

d) The value of the entropy in a given time bin is telling us whether the activity of the neuron considered at this particular time of the movie is stereotypical (almost always the same) or almost uniformly random, depending whether it is small or large. The heterogeneity of the values of entropy in different time bins is informing us about how the activity of this neuron is impacted by the movie (if there was no impact of the stimulus, entropy would remain the same throughout the movie).

e) The probability $P(m)$ that the neuron is active in this time bin in m repetitions of the movie out of N total repetitions is given by the binomial distribution:

$$P(m) = \binom{N}{m} p^m (1-p)^{N-m} \,. \tag{24}$$

f) The likelihood that the neuron is active in this time bin in m repetitions of the movie out of N total repetitions given the model above and the value p is $P(\text{data}|\text{model}, p) = P(m)$. In order to find the maximum-likelihood estimate of p as a function of m and N, we need to differentiate the log-likelihood versus p at N and m constant. Thus, we need to solve

$$0 = \frac{d}{dp} \ln \left[P(\text{data}|\text{model}, p) \right] = \frac{d}{dp} \left[m \ln p + (N-m) \ln(1-p) \right] = \frac{m}{p} - \frac{N-m}{1-p} \,, \tag{25}$$

which gives $p = m/N$. This is the natural expectation for p.

5.2 Activities of two neurons

a) The mutual information between two binary random variables X and Y reads

$$I(X;Y) = \sum_{x \in \{0,1\}, y \in \{0,1\}} P(x,y) \log_2 \left[\frac{P(x,y)}{P(x)P(y)} \right] \,. \tag{26}$$

b) The minimum possible value of $I(X;Y)$ is zero. It would be reached if X and Y were independent variables. Indeed, then, $P(x,y) = P(x)P(y)$ for all x, y.

c) The value of $I(X;Y)$ informs us about the statistical dependence between X and Y, and thus here between the activities of the two neurons considered. In particular, a large value of $I(X;Y)$ would mean that the activities of the two neurons considered are highly statistically dependent.

5.3 Maximum entropy model of neuron activity

a) We have measured the average value of X, obtaining m. We need to find a distribution that satisfies the constraints

$$\sum_{x \in \{-1,1\}} P(x) = P(1) + P(-1) = 1 \,, \tag{27}$$

and

$$\sum_{x \in \{-1,1\}} x P(x) = P(1) - P(-1) = m \,. \tag{28}$$

We thus need to maximize the function

$$\tilde{H}(X) = - \sum_{x \in \{-1,1\}} P(x) \log_2 P(x) + \frac{\lambda}{\ln(2)} \left[1 - \sum_{x \in \{-1,1\}} P(x) \right] + \frac{\mu}{\ln(2)} \left[m - \sum_{x \in \{-1,1\}} x P(x) \right] \,, \tag{29}$$

which gives for all $x \in \{-1, 1\}$:

$$P(x) = \frac{e^{-\mu x}}{e^{1+\lambda}}\,, \tag{30}$$

and thus explicitly

$$P(1) = \frac{e^{-\mu}}{e^{1+\lambda}}\,, \tag{31}$$

and

$$P(-1) = \frac{e^{\mu}}{e^{1+\lambda}}\,. \tag{32}$$

Next, we need to use the constraints in Eqs. 27 and 28 to find the values of λ and μ.

b) Denoting by λ the parameter associated to the normalization constraint, by μ and ν those associated respectively to the means of X and Y and my ξ that associated to the mean of XY, we have, by analogy with the previous formula, for all $x \in \{-1, 1\}$ and $y \in \{-1, 1\}$:

$$P(x, y) = \frac{e^{-\mu x - \nu y - \xi xy}}{e^{1+\lambda}}\,. \tag{33}$$

c) The formula given for N neurons indexed by i ranging from 1 to N is a generalization of Eq. 33 to N random variables. In order to compute the values of Z, h_i and J_{ij}, we would need to solve the system of equations given by the constraints (normalization, mean of each random variable and mean of each product of two random variables). This calculation is difficult in practice as all these equations are coupled and they are strongly non-linear.

d) A large absolute value of J_{ij} mean for a pair of neurons i, j means that they are strongly coupled and that their activities are statistically dependent. A small one means that their coupling is small.