Optional course M2 IMSD 2025-2026 : Statistics applied to biology

Introduction to statistical modelization of informational polymers

Jérémie Unterberger

Table of contents

Chapitre 1

Introduction

1.1 Informational polymers : introduction

Most applications in this course are about inferring a statistical model from data. Data are in the form of N observed *sequences* $\mathbf{x}_i$, $i = 1, \ldots, N$. Sequences are very diverse, and can be very long : from a few hundreds to a few thousands for proteins, from a few hundreds for a gene – since a gene codes for a protein, and it takes three nucleobases to code for an aminoacid – to a few billions for DNA molecules in the genome of a multicellular organism. Sequences in biology are **polymers**, i.e. monomer units bonded by covalent (chemical) bonds.

Mathematicians tend to think of them as words of length $L \geq 1$ in an alphabet $\mathcal{A}$, making up a kind of language. This language is based on physics (force fields) and chemistry (reactivity) ; however, the utter complexity of these laws for long sequences makes simple predictions from first principles very difficult. Also, these polymers are copied, or "transcribed" or "translated" through definite rules, from *templates* (the process is called *transcription* for the production of an RNA sequence from a DNA template, and *translation* for the production of a protein from an RNA template). The process has been repeated billions and billions of times in the course of evolution, and is not exempt from errors ; actually, "errors" (called generically : **mutations**) are one of the fundamental processes in evolution. As a result, each molecule comes with a huge number of variants, which cannot be exhausted.

Instead, the biologist looks for a 'clustering' of the space of all observed polymers. Naive clusterings based on distances easy to formulate on a mathematical level are practically clueless about biology. It is useful to have in mind first that the 'space' $\mathcal{E}$ on which polymers of length L live can be thought of either as $\mathcal{A}^L$ (word space) or something like $\mathbb{R}^{3L}$ (space of **embedded** positions of monomers in physical space). Thus $\mathcal{E}$ is, in general, unimaginably large. The reason for the failure of naive clusterings is twofold :

(1) physics and chemistry produce an extraordinarily complex energy landscape ; as a result, the space $\mathcal{E}$ is highly structured ;

(2) random evolution over billions of years may have allowed exploration of a significant part of $\mathcal{E}$. Yet only a tiny part of this huge space is observed today, as a result of *natural selection*.

The basis of natural selection, brought forward as one of the main driving forces in biology by Darwin in the 19th century, despite huge progress, has remained very elusive

up to now. Biology is still hampered by its tendency to think in terms of **function** – a protein is classified as globine if it is involved in binding or transporting oxygen, and therefore acts as a constituent of blood, whose *aim* is to bind or transport oxygen. So biologists are interested in *classifying* proteins according to identified functions – this looks more like supervised classification than clustering. However, the very mechanisms of mutation and natural selection do not operate according to function. Rather, they involve **structure** at a fundamental level – the letters of the 'word', making up what is called *primary structure*, and the folding of the polymer when embedded in physical space, making up *higher-order structures*, called *secondary*, *tertiary*, sometimes even *quaternary*, though these terms are sometimes simplified descriptions of an elusive reality. See Wikipedia article en.wikipedia.org/wiki/Structural_biology for a short notice on **structural biology**, experimental techniques (X-ray diffraction, NMR, cryogenic electron microscopy) and more recent computational methods (molecular dynamics, QM/MM, ML...), and applications in medicine (diseased tissues, interaction between pathogens and hosts, drug discovery).

As a consequence, one of the main research themes for structural biologists is the search for a *structure-to-function mapping*. In less expert terms, this is taught in secondary school as rules determining the phenotype in terms of the genotype for multicellular organisms. But this mapping also makes sense for RNA, which has diverse functions in the cell that are not directly observable through a cell phenotype.

Due to the large amount of data and its complexity, computational biology deals with this mapping at a statistical level. In particular, it treats sequences as **informational polymers**. Considering the word description to keep simple, imagine that every letter $a \in \mathcal{A}$ has empirical frequency p_a in a given subspace of $\mathcal{E} = \mathcal{A}^L$ characterized by a given higher-order structure and/or function. Then *Shannon's entropy* $S_1 = -\sum_a p_a \log p_a$ gives a quantitative measure of the *information* contained in one letter. Since $p \log(p) =$ when $p = 0, 1$, Shannon's entropy is zero for a deterministic distribution $P(a) = \delta_{a,a_0}$. Conversely, it is maximal when all $P(a) = \frac{1}{|\mathcal{A}|}$ are equal. It is in this maximal uncertainty case that the *choice* of a letter a gives the maximal information. Shannon's entropy is (up to normalization) the only function that is additive for product probabilities describing independent variables. Modelling a polymer of length L as independent letters therefore yields a very simple Shannon entropy $S_L = LS_1$.

Sequences are mainly of two types.

Nucleotide bases (for short, nucleobase/base nucléique, or simply base) sequences. For DNA, nucleotides are A (adenine), T (thymine), G (guanine) or C (cytosine). In RNA, T is replaced by U (uracile), which is very similar.

Protein sequences (also called peptides, or polypeptides, depending on the length), made up of amino-acid units. Amino-acids are chemically a carbon atom bonded to a carboxyl group, an amine group, and a distinctive group called *side chain*. Amino-acids can be bonded by dehydration through a *peptide bond*; side chains of protein sequences are commonly called **residues**, which characterize the aminoacid. Although residues could in principle be any carbon chain (C,H,O) with some nitrogen (N) atoms

and metallic atom insertions, DNA codes only (through RNA transcription, and then translation) for 22 types, of which only 20 are used commonly. Residues are designated with 3-letter acronyms and 1-letter codes, and can be grouped into 4-5 categories :

(A) *positively charged residues,* Arg(inine) **R**, His(tidine) **H**, Lys(ine) **K** ;
(A') *negatively charged residues,* Asp(artic acid) **D**, Glu(tamic acid) **E** ;
(B) *polar uncharged residues*, Ser(ine) **S**, Thr(eonine) **T**, As(paragi)n(e) **N**, GL(utami)n(e) **Q** ;
(C) special cases (very short residues with particular chemical properties) : Cys(teine) **C**, Se(leno)c(ysteine) **U**, Gly(cine) **G**, Pro(line) **P** ;
(D) and the largest category, **hydrophobic residues**, Ala(nine) **A**, Val(ine) **V**, I(so)le **I**, Leu(cine) **L**, Met(hionine) **M**, Phe(nyalanine) **F**, Tyr(osine) **Y**, Tr(y)p(tophan) **W** .

Hydrophilic residues, i.e. residues of type (A,A',B), induce strong electrostatic effects, which bind them to the water environment present in the cell. **Hydrophobic residues** of type (D), on the contrary, do not bind to water, which has the effect to bury them in the interior of the protein, as far as possible from the solvent. On the other hand, hydrophilic residues often form weak hydrogen bonds involving their hydroxyl, carbonyl and amine groups. These physical forces strongly influence the **structure** of protein sequences.

Statistical models. As explained above, the sequence space is partitioned into **families** – sequences believed to have the same higher-order structure and function. We consider sequences as words in the whole course, and understand the structure-function mapping as a **statistical model** $P(\mathbf{x})$ for sequences in a given family. These models serve a discriminative as well as generative objective. A typical discriminative use may be : given a newly observed protein, which family does it belong to ? (is it a mutated sequence made difficult to recognize through some malfunctioning, e.g. cancer ?) Generative uses are more typical of theoretical questions ; we will see some in the last part of the course.

Parametric, vs. non-parametric statistics. Models fall generally into two categories. **Parametric models** are based on the *ordering* of the letters. Typical models of this sort are **profile models**, in which the essential feature is the probability p_a^i of finding letter a at position i, or Markov chain models, in which the main parameters are transition probabilities $p_{a,b}$ from letter a to letter b. One also finds HMM (Hidden Markov Models) in this category, by assuming some position-dependent features not directly accessible to observation.

By contrast, **non-parametric models** make little or no use of the letter ordering. The main reason for such more complex models is that long-range correlations are in practice observed, both for RNA sequences and protein sequences. The usual interpretation is that x_i and x_j, $|i-j| \gg 1$ are correlated when their positions in physically embedded space $\mathbb{R}^3$ are close, making up what is called a *contact*. Two letters in contact usually change simultaneously, implying a large covariance. Non-predictable contacts impose a very large class of models, so large that the number of parameters is often not fixed in

advance, but during the inference procedure. In this sense, these models may be called non-parametric.

1.2 Introductory material

Bound extrema. Consider a hypersurface $\mathcal{S}$ in $\mathbb{R}^N$ defined by a single equation $F(\mathbf{x}) = 0$, satisfying the non-degeneracy condition

$$\forall \mathbf{x}\ F(\mathbf{x}) = 0, \vec{\nabla} F \neq \vec{0}, \tag{1.2.1}$$

and a C^1 function $f : \mathbb{R}^N \to \mathbb{R}$ defined on a neighborhood of $\mathcal{S}$.

One calls $\mathbf{x} \in \mathcal{S}$ a bound extremum of f if

$$\exists \lambda \in \mathbb{R}, \vec{\nabla} f(\mathbf{x}) = \lambda \vec{\nabla} F(\mathbf{x}) \tag{1.2.2}$$

Namely, a vector $\vec{v} \in \mathbb{R}^N$ is tangent to $\mathcal{S}$ at $\mathbf{x}$ iff $dF_{\mathbf{x}}(\vec{v}) = 0$. Since $dF_{\mathbf{x}}(\vec{v}) = \langle \vec{\nabla} F(\mathbf{x}), \vec{v} \rangle$, the condition (1.2.2) is equivalent to stating that $df_{\mathbf{x}}(\vec{v}) = 0$ for every $\vec{v}$ tangent to $\mathcal{S}$.

Bound extrema may be obtained by looking at the extrema of an "augmented" function

$$f(\mathbf{x}; \lambda) := f(\mathbf{x}) - \lambda F(\mathbf{x}) \tag{1.2.3}$$

where $\lambda \in \mathbb{R}$ is called a **Lagrangian parameter**. Namely, if $\mathbf{x}$ is in a neighbourhood of $\mathcal{S}$, $\vec{\nabla} f(\mathbf{x}; \lambda) = \vec{0}$ iff $F(\mathbf{x}) = 0$, i.e. $\mathbf{x} \in \mathcal{S}$, and (1.2.2) holds.

The above computation easily extends to the case of a surface defined by several equations $\mathbf{F}(\mathbf{x}) = \mathbf{0}$, $\mathbf{F} : \mathbb{R}^N \to \mathbb{R}^n$. Then bound extrema may be obtained by looking at the extrema of a function with n Lagrangian parameters, $f(\mathbf{x}; \lambda_1, \ldots, \lambda_n) := f(\mathbf{x}) - \sum_{i=1}^n \lambda_i F_i(\mathbf{x})$.

Maximum likelihood estimation (MLE) of Markov chains. Consider N observed i.i.d. sequences $x^{(i)} = (x_1^{(i)}, \cdots, x_{T_i}^{(i)})$, $i = 1, \ldots, N$, of variable lengths T_i. The statistical model for these sequences is a Markov chain with initial probability $\pi = (\pi_a)_{a \in \mathcal{A}}$ and transition probabilities $p = (p_{ab})_{a,b \in \mathcal{A}}$. The problem of maximizing the likelihood of $\mathcal{D} = \{x^{(1)}, \ldots, x^{(N)}\}$ w.r. to the parameters $\theta = (\pi, p)$ is solvable. Namely,

$$p(x^{(i)}|\theta) = \prod_a \pi_a^{\mathbf{1}[x_1^{(i)} = a]} \times \prod_{t=2}^{T_i} \prod_{a,b} p_{ab}^{\mathbf{1}[x_{t-1}^{(i)} = a, x_t^{(i)} = b]} \tag{1.2.4}$$

Thus

$$\log p(\mathcal{D}|\theta) = \sum_{i=1}^N \log p(x^{(i)}|\theta) = \sum_a N_a^1 \log \pi_a + \sum_{a,b} N_{ab} \log p_{ab}, \tag{1.2.5}$$

where we define the *counts*,

$$N_a^1 = \sum_{i=1}^N \mathbf{1}[x_1^{(i)} = a], \qquad N_{ab} = \sum_{i=1}^N \sum_{t=2}^{T_i} \mathbf{1}[x_{t-1}^{(i)} = a, x_t^{(i)} = b] \tag{1.2.6}$$

The MLE $\hat{\theta}$ of (π, p) is obtained by maximizing in π, p, which can be done separately. Focusing on the maximization in π,

$$\begin{aligned}\hat{\pi} &= \text{argmax}_\pi \sum_a N_a^1 \log \pi_a \\ &= \text{argmax}_\pi \sum_a q_a^{MLE} \log \pi_a \\ &= \text{argmax}_\pi \Big(- H(q^{MLE}) - D_{KL}(q^{MLE}||\pi) \Big) \\ &= \text{argmin}_\pi D_{KL}(q^{MLE}||\pi) = q^{MLE} \end{aligned} \tag{1.2.7}$$

where

$$q_a^{MLE} = \frac{N_a^1}{\sum_{a'} N_{a'}^1} \tag{1.2.8}$$

(normalized counts) are probabilities ;

$$H(q) = -\sum_q q \log(q) \tag{1.2.9}$$

is the Shannon entropy of q ;

$$D_{KL}(q||\pi) := \sum_a q_a \log(\frac{q_a}{\pi_a}) \tag{1.2.10}$$

is the **Kullback-Leibler divergence** (aka : **relative entropy**) of q w.r. to π. Thus the ML estimator $\hat{\pi}$ for the initial probability is the set of empirical frequencies for the first letter. A more direct computation goes through the maximization of the function $\mathcal{L}(\pi, \lambda) = \sum_a N_a^1 \log \pi_a \ - \lambda(\sum_a \pi_a - 1)$, with added Lagrangian multiplier λ ensuring the equality $\sum_a \pi_a = 1$.

By a similar computation (exercise), one can prove that the ML estimator for the transition probabilities

$$\hat{p}_{ab} = \frac{N_{ab}}{\sum_{b'} N_{ab'}} \tag{1.2.11}$$

are the empirical transition frequencies.

Posterior distribution , entropy, relative entropy. Assume that data $Y = \{\mathbf{y}_i\}_{i=1,\dots,N}$ are in the form of a i.i.d. sample of N data points $\mathbf{y}_i \in \mathbb{R}^L$ generated from a probability distribution $p_\theta(\mathbf{y})$ (called : **model**) depending on parameters $\theta \in \mathbb{R}^D$. The purpose of inference is to infer the best possible value of θ given the data.

Bayesian inference assumes some previous information on θ *prior* to measurements. This information iq given the form of a probability distribution $p(\theta)$, called the **prior**. Now that θ has turned into a random variable, we may reinterpret $p_\theta(\mathbf{y})$ as the distribution $p(\mathbf{y}|\theta)$ of Y conditional to θ. Bayes' rule then yields a conditional distribution of θ conditional to the data Y, called the **posterior**,

$$p(\theta|Y) = \frac{p(Y|\theta)p(\theta)}{p(Y)} \tag{1.2.12}$$

with (by independence)

$$p(Y|\theta) = \prod_{i=1}^{N} p(\mathbf{y}_i|\theta) \tag{1.2.13}$$

and (assuming p represents densities)

$$p(Y) = \int p(Y|\theta)\, p(\theta) d\theta. \tag{1.2.14}$$

Note that, for a given dataset Y,

$$p(\theta|Y) \propto p(Y|\theta)p(\theta) \tag{1.2.15}$$

where "$\propto$" means : equal up to an undisclosed proportionality coefficient (here, $1/p(Y)$) ; thus the prior and posterior essentially differ by a multiplicative factor $p(Y|\theta)$, describing the supplementary information brought about by the data.

It is interesting to study the large N asymptotics ; unsurprisingly, the dependence of the posterior upon the prior disappears, and the posterior concentrates around the true parameter value $\hat{\theta}$ from which data are drawn (mind the notation $\hat{\theta}$, which is unusual for mathematicians, but convenient and widely used). Namely, from the law of large numbers,

$$\frac{1}{N} \log p(Y|\theta) \to_{N\to\infty} \int d\mathbf{y}\, p(\mathbf{y}|\hat{\theta})\, \log p(\mathbf{y}|\theta). \tag{1.2.16}$$

Thus the posterior distribution $p(\theta|Y) \propto p(Y|\theta)p(\theta)$ is of the form $\approx e^{-NS_c(\hat{\theta},\theta)} \times p(\theta)$, where

$$S_c(\hat{\theta}, \theta) := -\int d\mathbf{y}\, p(\mathbf{y}|\hat{\theta})\, \log p(\mathbf{y}|\theta) \tag{1.2.17}$$

is called the **cross-entropy** (of θ w.r. to $\hat{\theta}$). This quantity is always ≥ 0 ; actually, the following fundamental identity holds,

$$S_c(\hat{\theta}, \theta) = S(\hat{\theta}) + D_{KL}(\hat{\theta}||\theta), \tag{1.2.18}$$

where :

1. $S(\hat{\theta}) := -\int d\mathbf{y}\, p(\mathbf{y}|\hat{\theta})\, \log p(\mathbf{y}|\hat{\theta})$ is the **Shannon entropy** of $p(\cdot|\hat{\theta})$;
2. $D_{KL}(\hat{\theta}||\theta) = \int d\mathbf{y}\, p(\mathbf{y}|\hat{\theta})\, \log(\frac{p(\mathbf{y}|\hat{\theta})}{p(\mathbf{y}|\theta)})$ is the **relative entropy** or **Kullback-Leibler divergence** of $p(\cdot|\hat{\theta})$ w.r. to $p(\cdot|\theta)$.

From the more general definitions,

$$S(p) = -\mathbb{E}_p \log(p) = \begin{cases} -\sum_i p_i \log(p_i) \\ -\int dx\, p(x) \log(p(x)) \end{cases}, \; D_{KL}(p||q) = \begin{cases} \sum_i p_i \log(\frac{p_i}{q_i}) & \text{(discrete case)} \\ \int dx\, p(x) \log(\frac{p(x)}{q(x)}) & \\ \text{(continuous case)} & \end{cases} \tag{1.2.19}$$

and Jensen's formula, it is apparent that these two functions are always ≥ 0 (exercise). In the continuous case, p, q denote densities. (The relative entropy makes sense as long as the distribution p has a density w.r. to q). Furthermore, $D_{KL}(p||q) = 0$ iff $p \overset{a.s.}{=} q$; thus (if the distribution $p(\cdot|\theta)$ are discernible) $D(\hat{\theta}||\theta) = 0$ iff $\theta = \hat{\theta}$, implying that

the posterior distribution is indeed concentrated around the value $\theta = \hat{\theta}$. The prior distribution $p(\theta)$ in factor (see below (1.2.16)) gives an additive contribution to the cross-entropy of order $O(1/N)$, which is negligible when N is large enough.

Bayesian PME (Posterior Mean Estimate) and pseudocounts. Case of trivial Markov chains (i.i.d. r.v., $p_{ab} = \pi_b$). There is a related procedure based on **PME** or (more or less equivalently, as we shall see) **MAP (Maximum a posteriori estimation)**, giving a slightly different result, that is particularly useful in practice for short sequences. We illustrate it on the case of i.i.d. r.v., so that the length of chains is $T_i = 1$. One wants to estimate π. Choose $\alpha = (\alpha_a)_{a\in\mathcal{A}} > 0$ and a prior **Dirichlet distribution**

$$Dir(\pi|\alpha) = Z^{-1}(\alpha) \prod_a \pi_a^{\alpha_a - 1} \ \delta(\sum_a \pi_a - 1). \tag{1.2.20}$$

The normalizing constant $Z(\alpha)$ is equal to the multinomial beta-function $\frac{\prod_a \Gamma(\alpha_a)}{\Gamma(\sum_a \alpha_a)}$. When all $\alpha_a = 1$, $Dir(\cdot|\alpha)$ is the uniform distribution on the simplex $\sum_a \pi_a = 1$. This distribution looks like the multinomial distribution

$$Mult(n|\pi) = \frac{n!}{\prod_a n_a!} \prod_a \pi_a^{n_a}, \tag{1.2.21}$$

with the formal identification, $a_\alpha \equiv n_a + 1$, except that we let π vary instead of n. Also, the α's are not necessarily integers. Note $A := \sum_a \alpha_a$. The mean of a Dirichlet distribution is $\mathbb{E}[\pi_a] = \frac{\alpha_a}{A}$; in particular, $\mathbb{E}[\pi_a] = \frac{1}{|\mathcal{A}|}$ when all $\alpha_a = 1$. The mode, $\text{argmax}_\pi Dir(\pi|\alpha) = (\frac{\alpha_a - 1}{A - |\mathcal{A}|})_a$, is well defined only if all $\alpha_a > 0$, and close to the mean when all α_a are large. When all α_a go to infinity at the same speed, $\alpha_a \sim N p_a$ for some probability distribution $p = (p_a) > 0$ and $N \to \infty$, π concentrates around its mean (p_a). It is thus natural to call the α_a "pseudo-counts" : the prior behaves as would be expected if one used the priori information drawn from tossing a dice N times and getting empirical frequencies p.

With this specific prior, the MPE/MAP may be solved ; namely, letting $N = (N_a)$ be the observed counts, the Bayes rule yields

$$\begin{aligned} P(\pi|N) &\propto P(N|\pi) Dir(\pi|\alpha) \\ &\propto \prod_a \pi_a^{n_a} \times \pi_a^{\alpha_a - 1} = \prod_a \pi_a^{n_a + a_{\alpha - 1}} \\ &\propto Dir(\pi|N + \alpha) \end{aligned} \tag{1.2.22}$$

so that the posteriori distribution is a modified Dirichlet distribution. The PME is equal to

$$\pi^{PME} = q^{PME}, \qquad q_a^{PME} = \frac{N_a + \alpha_a}{N + A}. \tag{1.2.23}$$

Comparing to q^{MLE}, one sees again that the α_a acts as additive "pseudo-counts". The conclusion is the same for MAP up to the replacement $\alpha_a \to \alpha_a - 1$.

Exercises for Chapter 1.2

Ex. 1.2.1. Maximum likelihood estimation of Markov chains.

1. Prove that MLE $\hat{\pi}$ of initial distribution $\hat{\pi}$ of a Markov chain are equal to normalize counts q_a^{MLE} by maximizing $\mathcal{L}(\pi, \lambda) = \sum_a N_a^1 \log \pi_a \ - \lambda(\sum_a \pi_a - 1)$.
2. Similarly, prove that MLE $\hat{p}$ of transition probabilities $p = (p_{ab})_{a,b\in\mathcal{A}}$ are equal to $\hat{p}_{ab} = \frac{N_{ab}}{\sum_{b'} N_{ab'}}$ by maximizing $\mathcal{L}(\mathbf{p}, \lambda) = \sum_{a,b} N_{ab} \log p_{ab} - \sum_a \lambda(\sum_b p_{ab} - 1)$.

Ex. 1.2.2. Entropy.

1. (Shannon entropy). Prove that Shannon entropy $S(p) = -\sum_i p_i \log(p_i), i \in \Omega$ is maximal for the uniform distribution $p_i = \frac{1}{|\Omega|}$. Prove that $S(p \otimes q) = S(p) + S(q)$ for a product distribution $p \otimes q$.
2. **(Maximum Entropy Principle, MEP)**. Let $\mathbf{y}$ be a discrete L-dimensional r.v. with distribution p and *fixed* average $\langle \mathbf{y} \rangle = \sum_{\mathbf{y}} p(y)\, \mathbf{y}$. Prove that $S(p)$ *restricted to distributions* p *with average* $\sum_{\mathbf{y}} p(\mathbf{y})\, \mathbf{y} = \langle \mathbf{y} \rangle$ is maximum for p of the form
$$p(\mathbf{y}) = \frac{1}{Z}\, e^{\sum_i \mu_i y_i} \tag{1.2.24}$$
with normalizing constant Z. What conditions do μ_i satisfy ?

   Similarly, assume given an **energy function** $E(\mathbf{y})$, and fix an energy level $\langle E \rangle$. Prove that $S(p)$ *restricted to distributions* p *with average energy* $\sum_{\mathbf{y}} p(\mathbf{y}) E(\mathbf{y}) = \langle E \rangle$ is maximum for p of the **Gibbs form**
$$p(\mathbf{y}) = \frac{1}{Z} e^{\mu E(\mathbf{y})} \tag{1.2.25}$$
   Physicists identify μ with $-1/k_B T$. The resulting distribution is known as *canonical ensemble* in statistical mechanics.
3. **(Mutual Information, MI)** Let (x, y) be two coupled random variables. Define
$$MI(x, y) = \sum_{x,y} p(x, y) \log \Big(\frac{p(x, y)}{p(x) p(y)} \Big). \tag{1.2.26}$$
   Prove that $MI(x, y) = D_{KL}(p(x, y) \,||\, p(x) \otimes p(y))$ is ≥ 0 and symmetric in x, y. Compute its value in the extreme cases when x, y are independent or when $x = f(y)$. Prove that $MI(x, y) = S[p(x)] - S[p(x|y)]$, where $S[p(x|y)] = -\sum_{x,y} p(x, y) \log(\frac{p(x,y)}{p(y)})$, the conditional entropy of x given y, is an average of $p(x|y)$.
4. *(relative entropy and entropy dissipation in Markov chains.)* Let $(X_t)_{t\geq 0}$ be a time-continuous Markov chain chain with irreducible generator $A = \frac{d}{dt} P_t \Big|_{t=0}$, and unique invariant probability measure π. Let μ_t, $t \geq 0$ the measure of X_t. Prove that the relative entropy $H_t := D(\mu_t|\pi)$ decreases with time. *Indication :* soit $(P_t)_{t\geq 0}$ le semi-groupe associé, de sorte que $\mu_t = \mu P_t$. On utilisera la convexité de la fonction $\phi(u) = u \log(u)$, et on utilisera l'équation de stationarité, $\sum_y \pi(y) P_t(y, x) = \pi(x)$.

Ex. 1.2.3. Conjugate priors. Let $p(\mathbf{x}|\theta)$ be a family of statistical models depending on a parameter θ. A conjugate prior is a prior distribution $p(\theta)$ belonging to some parametric family, for which the resulting posterior distribution $p(\theta|\mathbf{x})$ also belongs to the same family.

1. Fix $n \in \mathbb{N}^*$. Let $X|\theta \sim Bin(n, \theta)$, and $\theta \sim B(a, b)$ a beta-distribution, i.e. $p(\theta) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)}$, $B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$. Prove that $\theta|X \sim B(a', b')$ with $a' = a + x, b' = b + n - x$.
2. Fix $\sigma > 0$. Let $x_i|\theta \sim \mathcal{N}(\theta, \sigma^2)$, $i = 1, \ldots, n$ be i.i.d. normal variables, and $\theta \sim \mathcal{N}(a, b^2)$. Prove that $\theta|\mathbf{x} \sim \mathcal{N}(a', (b')^2)$ for $a' = \frac{nb^2}{nb^2+\sigma^2}\bar{\mathbf{x}} + \frac{\sigma^2 a}{nb^2+\sigma^2}$, $(b')^2 = \frac{\sigma^2 b^2}{nb^2+\sigma^2}$.
3. Fix $n \in \mathbb{N}^*$. Let $x_i|\theta \sim Exp(\theta)$, $i = 1, \ldots, n$ be i.i.d. exponential variables, and $\theta \sim Gamma(a, b)$, i.e. $p(\theta) = \frac{\theta^{a-1}e^{-\theta/b}}{\Gamma(a)b^a}$. Prove that $\theta|\mathbf{x} \sim Gamma(a', b')$, $a' = a + n$, $b' = b + n\bar{\mathbf{x}}$.

Chapitre 2

Quelques idées de recruteurs potentiels (hors académique)

Quelques idées franco-centrées (mais le domaine est dominé par les groupes anglo-saxons et suisses).

Industrie pharmaceutique : Sanofi, Servier, IPSEN

CRO (contract research organizations), sociétés de recherche contractuelle, auprès desquelles les pharmas sous-traitent : Novalix (Strasbourg, Oncodesign (Dijon), RCTs (Lyon), Aixial Group, Clin4all (Paris), Biotrial (Riom), ICTA (CRO internationale ayant des activités en France)

Grosses entreprises de biotechnologies :
https ://iktos.ai/, https ://www.aqemia.com/, https ://www.owkin.com/,
https ://www.bioptimus.com

Cf. https ://france-biotech.fr/annuaire/adherents/

Biotechs startups : https ://future4care.com/startups, ex. : https ://onebiosciences.fr/

Incubateurs d'entreprises biotech :

Genopole (Évry)
Paris Biotech Santé (Paris)
Agoranov (Paris)
BioLabs Hôtel-Dieu (Paris)
Station F – HealthTech Program (Paris)
IncubAlliance Paris-Saclay (Orsay)
Medicen Paris Région (Paris / IDF)
Eurasanté (Lille)
Eurasanté Bio-Incubateur (Lille)
Eurasanté Bio-Accélérateur (Lille)
Eurobiomed (Marseille / Montpellier)
The Camp / ZEBOX Health (Aix-Marseille)
Montpellier BIC – BioTech (Montpellier)
Lyonbiopôle (Lyon / Grenoble)
Pulsalys (Lyon)
I-Care Cluster (Auvergne-Rhône-Alpes)

SATT Paris-Saclay (tech transfer, incub support)
SATT AxLR (Montpellier)
SATT Conectus (Strasbourg)
SEMIA – Quest for Health (Strasbourg / Grand Est)
Toulouse White Biotechnology (Toulouse)
Catalyseur – Toulouse Tech Transfer (Toulouse)
Atlanpole Biotherapies (Nantes)
Biogenouest (Ouest)
Brest IRT BioCom / DeepTech (Bretagne, part biotech)
Normandie Incubation (Caen / Rouen)

References

General culture

[1] Alberts, Morgan et al. *Molecular biology of the cell* (6th ed.), Garland Science.

[2] C. Branden, J. Tooze. *Introduction to Protein Structure*, CRC Press, 1991.

[3] S. Cocco, R. Monasso, F. Zamponi. *From statistical physics to data-driven modelling, with applications to quantitative biolog*, Oxford University Press, 2022.

[4] Ken A. Dill, S. Bromberg. *Molecular driving forces. Statistical thermodynamics in biology, chemistry, physics and nanoscience*, Taylor & Francis, 2011.

[5] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of statistical learning : data mining, inference and prediction*. Springer, 2001.

[6] K. P. Murphy. *Machine learning. A probabilistic perspective*, MIT Press, 2002.

[7] Martin A. Nowak. *Evolutionary dynamics. Exploring the equations of life*, chap. 3, 4, 6. Harvard University Press, 2006.

[8] J. Pevsner. *Bioinformatics and functional genomics*, Wiley, 2015.

[9] D.M. Zuckerman. *Statistical Physics of Biomolecules : An Introduction*, CRC Press, 2010.

Main tutorials

[10] Simona Cocco, Rémi Monasson, Zamponi. *From statistical physics to data-driven modelling, with applications to quantitative biology*, Oxford Univ. Press.

[11] R. Durbin, S. Eddy, A. Krogh, G. Mitchison. *Biological sequence analysis. Probabilistic models of proteins and nucleic acids*, Cambridge University Press (1998).

Specialized topics

Main reference laboratories in Paris are IPGG (Institu Pierre-Gilles de Gennes), part of ESPCI (Ecole Supérieure de Physique et Chimie Industrielle), and LBCQ (Laboratoire de Biologie Computationnelle et Quantitative, Sorbonne Univeriseté).

EM.

[12] Sean Borman, *The Expectation Maximization Algorithm : A Short Tutorial*, online pdf tutorial (2004).

DCA.

[13] L. Rosset (LBCQ), R. Netti (LBCQ), A. P. Muntoni, M. Weigt (LBCQ), F. Zamponi. *ADABM DCA 2.0. A flexible but easy-to-use package for direct coupling analysis*, arXiv :2501.18456.

[14] F. Calvanese (LBCQ), C. Lambert (doct. P. Nghe), P. Nghe (IPGG), F. Zamponi, M. Weigt (LBCQ). *Towards parsimonious generative modeling of RNA families*, Nucl. Acids Research **52**, 5465-5477 (2024).