BIOINFORMATICS & COMPUTATIONAL GENETICS MSc
PROBABILITY AND STATISTICS

Chapter 4. Probabilistic Models and Statistical Analysis in Bioinformatics

The statistical analysis of biological sequences utilizes the ideas previously introduced in the context of general probability theory, such as conditional probability, entropy, covariance and order statistics, and statistical tools such as maximum likelihood estimation, hypothesis testing and p-value calculation. In many ways, biological sequences provide a relatively simple data analysis environment, in that the raw data themselves are usually discrete items. The main difficulty lies in the scale of the analysis as genomic/proteomic data sets are typically vast

4.1 Models and Analyses for Dependence in Biological Sequences : Stochastic Processes

A *stochastic process*, denoted here by $\{X_t\}$ or $X(t)$, is a sequence of discrete or continuous random variables indexed by a time parameter t that itself may be *discrete* (so that $t = 0, \pm 1, \pm 2, ...$ say) or *continuous* (so that $t \geq 0$ say). For example, a nucleotide sequence can be thought of as a simple stochastic process where time variable t indexes base position, and random variables $X_1, X_2, X_3, ...$ are discrete random variables taking values on the integers $\{1(=A), 2(=C), 3(=G), 4(=T)\}$ say, that correspond to the bases in consecutive positions. We can treat the sequence of variables $\{X_1, X_2, X_3, ...\}$ merely as a collection of random variables and specify and study their joint distribution. Usually, the most simple type of stochastic process is a sequence of *discrete* variables observed in *discrete* time, as then we can merely use the chain rule to write down the joint probability mass function

$$\mathrm{P}\left[X_1 = x_1, X_1 = x_2, X_1 = x_3, ...\right] = \mathrm{P}\left[X_1 = x_1\right] \times \mathrm{P}\left[X_2 = x_2 | X_1 = x_1\right] \times \mathrm{P}\left[X_3 = x_3 | X_1 = x_1, X_2 = x_2\right] ...$$

Such a model will form the basis of much of the statistical analysis of biological sequences.

4.1 1. The Poisson Process

Consider a sequence of events occurring in *continuous* time according to some probabilistic rules that are to be defined, and consider random variables $\{X(t), t \geq 0\}$ that count the **number** of events that occur in the intervals $(0, t]$ for $t \geq 0$, so that $X(t)$ takes values $0, 1, 2, ...$

Suppose that

1. $X(0) = 0$

2. For all $0 < s \leq t$, $h > 0$, and non-negative integers n and m,

$$\mathrm{P}[X(t+h) - X(t) = n | X(s) = m] = \mathrm{P}[X(t+h) - X(t) = n]$$

that is, the numbers of events occurring in disjoint intervals are probabilistically **independent**.

3. For $\triangle t > 0$ small,

$$\mathrm{P}[X(t + \triangle t) - X(t) = 1] \approx \lambda \triangle t$$

for some $\lambda > 0$, that is, the probability of **exactly one** event occurring in the small interval $(t, t + \triangle t]$ is, for small $\triangle t$, **proportional to the length of the interval**, $\triangle t$. The parameter λ is known as the **rate** parameter.

4. For $\triangle t > 0$ small,

$$\mathrm{P}[X(t + \triangle t) - X(t) \geq 2] \approx 0$$

for some $\lambda > 0$, that is, the probability of **more than one** event occurring in a small interval $(t, t + \triangle t]$ is essentially zero.

It can be shown that the sequence of discrete variables $\{X(t), t \geq 0\}$ each follow a discrete **Poisson** distribution, that is, if $P_n(t)$ =P[Precisely n events occur in $(0, t]$] it can be shown that

$$P_n(t) = P[X(t) = n] = \frac{e^{-\lambda t}(\lambda t)^n}{n!} \qquad n = 0, 1, 2, ...$$

The sequence $\{X(t), t \geq 0\}$ form a **homogeneous Poisson Process with rate** λ (that is, a Poisson process with constant rate λ)

Distributional Results relating to the Poisson process

We have studied the Poisson process in previous sections; the relevant distributional results can be summarized as follows:

(a) For $t \geq 0$, $X(t) \sim Poisson(\lambda t)$

(b) If $T_1, T_2, T_3, ...$ define the inter-event times of events occurring as a Poisson process, that is, for $n = 1, 2, 3, ...$

$$T_n = \text{“time between } (n-1)\text{st and } n\text{th event”}$$

then $T_1, T_2, T_3, ...$ are a sequence of **independent** and **identically distributed** random variables with

$$T_n \sim Exponential(\lambda)$$

(c) If $Y_1, Y_2, Y_3, ...$ define the times of events occurring as a Poisson process, that is, for $n = 1, 2, 3, ...$

$$Y_n = \text{“time at which } n\text{th event occurs”}$$

then $Y_1, Y_2, Y_3, ...$ are a sequence of random variables with

$$Y_n \sim Gamma(n, \lambda)$$

(d) Consider the interval of length L, $(0, L]$, and suppose that k events occur (according to the Poisson process rules) in that interval. If $V_1, V_2, ..., V_k$ define the (ordered) event times of the k events occurring in that interval, then, given L and k, $V_1, V_2, ..., V_k$ are the **order statistics** derived from an **independent** and **identically distributed** random sample $U_1, U_2, ..., U_k$ where

$$U_i \sim Uniform(0, L)$$

The homogeneous Poisson process is the standard model for *discrete* events that occur in *continuous* time. It can be thought of as a limiting case of an independent Bernoulli process (a model for *discrete* events occurring in *discrete* time); let

$$X_t \sim Bernouilli\,(\theta) \qquad t = 1, 2, 3,$$

be an i.i.d. sequence: a realization of this sequence might look like

0010101000100101000010010001.....

For such a process, we have seen (in chapter 2) that

(a) the number of 1s that occur in any finite and disjoint subsequences are independent random variables
(b) the number of 1s that occur in any finite subsequence of n time points is a $Binomial\,(n, \theta)$ random variable
(c) the numbers of trials between successive 1s are i.i.d. $Geometric\,(\theta)$ random variables.

Now consider the a large sequence where θ is small; the Bernoulli sequence approximates a continuous time sequence, where the events (i.e. the 1s) occur at a constant rate of $\lambda = n\theta$ per n trials, and the Poisson process requirements are met.

The homogeneous Poisson process is a common model that is used often in many scientific fields. For example, in genetics, the Poisson process is used to model the occurrences of crossings-over in meiosis.

4.1 2. Markov Chains

Markov chains are a special type of *discrete* valued, *discrete* time stochastic process that play a central role in the modelling of biological sequences. Consider a sequence of discrete random variables $\{X_t\}$ indexed in discrete time by $t = 1, 2, 3, ...$, and each having a range of possible values (or **states**) $\mathbb{X} = \{1, 2, 3, ..., S\}$ say (with S finite).

Suppose that the joint distribution of the $\{X_1, X_2, X_3, ...\}$ is specified (using chain-rule ideas) entirely via the one-step ahead conditional probabilities

$$\mathrm{P}\left[X_{t+1} = x_{t+1} | X_1 = x_1, ..., X_t = x_t\right] = \mathrm{P}\left[X_{t+1} = x_{t+1} | X_t = x_t\right]$$

(known as the **Markov** or **memoryless property)** so that

$$\mathrm{P}\left[X_{t+1} = j | X_t = i\right] = \Pr\left(\text{State } i \text{ at time } t \to \text{State } j \text{ at time } t+1\right) = p_{ij} \qquad i, j \in \{1, 2, 3, ..., S\}$$

say, that **does not depend on** t. The probabilistic specification can be encapsulated in the $S \times S$ matrix

$$P = \begin{bmatrix} p_{11} & p_{12} & p_{13} & \cdots & p_{1S} \\ p_{21} & p_{22} & p_{23} & \cdots & p_{2S} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_{S1} & p_{S2} & p_{S3} & \cdots & p_{SS} \end{bmatrix}$$

which is called the **transition matrix** where the element in row i and column j defines the probability of moving from state i to state j. Note that the row totals must equal 1.

The sequence of random variables described by the matrix P form a **Markov Chain**. Thinking back to the chain rule, in order to complete the specification, a probability specification for the initial state random variable X_1 is required; we can denote this discrete probability distribution by row vector of probabilities $\pi^{(1)} = \left(\pi_1^{(1)}, \pi_2^{(1)}, ...\pi_S^{(1)}\right)$. To compute the (marginal) probability of random variable X_t taking the value i, we can use matrix algebra and an iterative calculation as follows; let $\pi^{(t)} = \left(\pi_1^{(t)}, \pi_2^{(t)}, ...\pi_S^{(t)}\right)$ denote the probability distribution of X_t. First, using the Theorem of Total Probability (chapter 1), conditioning in turn on the different possible values of X_{t-1}

$$\mathrm{P}\left[X_t = j\right] = \sum_{i=1}^{S} \mathrm{P}\left[X_t = j | X_{t-1} = i\right] \mathrm{P}\left[X_{t-1} = i\right]$$

which can be re-written

$$\pi_j^{(t)} = \sum_{i=1}^{S} p_{ij} \pi_i^{(t-1)} \qquad \text{or in matrix form} \qquad \pi^{(t)} = \pi^{(t-1)} P$$

Using this definition recursively, we have

$$\pi^{(t)} = \pi^{(t-1)} P = \pi^{(t-2)} P^2 = \pi^{(t-3)} P^3 = ... = \pi^{(2)} P^{t-2} = \pi^{(1)} P^{t-1}$$

which gives a mechanism for computing the marginal probability after t steps.

EXAMPLE The random walk with absorbing states.

Suppose that the Markov chain is defined on the states $\mathbb{X} = \{0, 1, 2, 3, ..., S\}$

$$\mathrm{P}\left[X_{t+1} = j | X_t = i\right] = p_{ij} = \begin{cases} p & j = i + 1 \\ 1 - p & j = i - 1 \end{cases}$$

and zero otherwise, unless $i = 0$ when

$$\mathrm{P}\left[X_{t+1} = 0 | X_t = 0\right] = 1 \qquad \mathrm{P}\left[X_{t+1} > 0 | X_t = 0\right] = 0$$

so that the chain gets "stuck" in state 0, or unless $i = S$ when

$$\mathrm{P}\left[X_{t+1} = S | X_t = S\right] = 1 \qquad \mathrm{P}\left[X_{t+1} < S | X_t = S\right] = 0$$

so the chain gets "stuck" in state S. Here the states 0 and S are termed **absorbing** states.

Stationary distributions

If the Markov chain does not have any absorbing states, or absorbing subsets of states, then it is of interest to try to determine what happens to the marginal probability distributions of the variables $\{X_t\}$ as t becomes large. It can be shown that, in most cases, the Markov chain eventually "settles down", that is, that the marginal distribution stabilizes. That is, for the Markov chain with transition matrix P, there exists a **stationary** or **equilibrium** probability distribution $\pi = (\pi_1, \pi_2, ..., \pi_S)$ that satisfies

$$\pi = \pi P$$

so that,

$$\pi^{(t)} \to \pi \qquad \text{or} \qquad \left(\pi_1^{(t)}, \pi_2^{(t)}, ..., \pi_S^{(t)}\right) \to (\pi_1, \pi_2, ..., \pi_S)$$

as $t \to \infty$. This equilibrium distribution can be computed algebraically, or using numerical computation.

EXAMPLE

Consider the Markov chain with transition matrix for $S = 4$

$$P = \begin{bmatrix} 0.6 & 0.1 & 0.2 & 0.1 \\ 0.1 & 0.7 & 0.1 & 0.1 \\ 0.2 & 0.2 & 0.5 & 0.1 \\ 0.1 & 0.3 & 0.1 & 0.5 \end{bmatrix}$$

Then, the equilibrium distribution π can be obtained by solving the system of equations

$$\begin{aligned} 0.6\pi_1 + 0.1\pi_2 + 0.2\pi_3 + 0.1\pi_4 &= \pi_1 \\ 0.1\pi_1 + 0.7\pi_2 + 0.1\pi_3 + 0.1\pi_4 &= \pi_2 \\ 0.2\pi_1 + 0.2\pi_2 + 0.5\pi_3 + 0.1\pi_4 &= \pi_3 \\ \pi_1 + \pi_2 + \pi_3 + \pi_4 &= 1 \end{aligned}$$

Note that the last row of P is not used in the calculation, but is replaced by the probability distribution constraint that the probabilities must sum to 1. This system can be solved to find

$$\pi = (\pi_1, \pi_2, ..., \pi_S) = (0.2414, 0.3851, 0.2069, 0.1667)$$

Here is a MAPLE calculation for this matrix: it computes the matrix P^{50}, and demonstrates that the marginal distribution for the next state is independent of the current state.

```
> with(linalg): P:=Matrix(1..4,1..4,
  [[0.6,0.1,0.2,0.1],
  [0.1,0.7,0.1,0.1],
  [0.2,0.2,0.5,0.1],
  [0.1,0.3,0.1,0.5]]);

                           [.6    .1    .2    .1]
                           [                    ]
                           [.1    .7    .1    .1]
                      P := [                    ]
                           [.2    .2    .5    .1]
                           [                    ]
                           [.1    .3    .1    .5]
> evalf(P^50,6);

            [.241379    .385057    .206897    .166667]
            [                                        ]
            [.241380    .385058    .206897    .166667]
            [                                        ]
            [.241380    .385058    .206897    .166667]
            [                                        ]
            [.241379    .385057    .206897    .166667]
```

4.1 3 Markov Models for DNA Sequences

The Markov process models described in the previous section can be used to model DNA sequences. Genomic sequences are comprised of coding and non-coding regions, and small-scale features such as splice sites, and thus some generalization of the models is required. However, if a homogeneous segment can be identified, a Markov model might seem plausible. Here the $S = 4$ states are the nucleotide codes (A, C, G, T) and the transition matrix (to be identified numerically) is

$$P = \begin{bmatrix} p_{AA} & p_{AC} & p_{AG} & p_{AT} \\ p_{CA} & p_{CC} & p_{CG} & p_{CT} \\ p_{GA} & p_{GC} & p_{GG} & p_{GT} \\ p_{TA} & p_{TC} & p_{TG} & p_{TT} \end{bmatrix}$$

The S-Plus Exercise 1 was concerned with estimating the parameters in this matrix; in fact Maximum Likelihood Estimation gives a formal justification for the technique used there. However, there is no reason to believe that the assumption of homogeneity will hold across large-scale genomic regions.

4.1 4 Hidden Markov Models for DNA Sequences – introduction

Here we note that Markov process models can be used to model the *hidden* or *latent* structure in DNA sequences that determines whether a segment is a coding region, intron, splice site, C+G islands etc. So far we have only modelled the observed nucleotide sequence using the Markov structure. However, we can couple the Markov model for the observed structure with a **Hidden Markov Model (HMM)** for the latent structure.

Suppose that a nucleotide position can be classified as being in a **region** of one of a number, H. Then, we might assume that, within a homogeneous region labelled by h the observed nucleotide sequence follows a Markov chain with transition matrix P_h To complete the specification, we assume that there is a sequence in parallel to the nucleotide sequence, comprising region label random variables $H_1, H_2, ...$ which itself follows a Markov chain governed by transition matrix P_θ where

$$\mathrm{P}\left[H_{t+1} = j | H_t = i\right] = \mathrm{Pr}\left(\text{Region type } i \text{ at time } t \to \text{Region type } j \text{ at time } t+1\right) = \theta_{ij} \qquad i, j \in \{1, 2, 3, ..., H\}$$

We will study the HMM specification in more detail in later sections.

4.2 Tests for a single DNA sequence

The discovery of structure in a DNA sequence is one of the primary goals of Bioinformatics. This structure can take the form of nucleotide frequency changes, changes in dependence (as in the Markov process construction), the appearance of large scale and small scale structures. We will now study several types of analysis.

4.2.1 Long Sequences of Nucleotide Repeats

Suppose that interest lies in detecting whether a nucleotide is present in a genomic segment in long repeat stretches. The statistical objective is to process a segment of length N say, and to test the hypothesis H_0 that the sequence is **random and independent** against the hypothesis H_1 that there is evidence of non-randomness or a lack of independence

From the start of the segment, reading "left-to-right" along the nucleotide sequence we can gather data on *run-lengths.* Suppose that the nucleotide of interest, A say, is coded 1 (success) and the other nucleotides are coded 0 (failure). Then, the sequence

$$CGAGAAGATATAAATTCAAATA$$

codes as

$$0010110101011100011101$$

giving run lengths of successes of $0, 0, 1, 0, 2, 0, 1, 0, 3, 0, 0, 0, 3, 0, 1$.

Now, if the sequence **is** a realization of a random and independent sampling process (that is, under H_0), then it can be shown that the distribution of run-lengths should follow an adjusted $Geometric(1 - p_A)$ distribution, that is, if X_i is defined the run-length of run i, then

$$f_{X_i}(x) = (1 - p_A)p_A^x \qquad x = 0, 1, 2, ...$$

where p_A is a hypothesized marginal probability of nucleotide A. Furthermore, **under** H_0, the collection of run-length random variables, $X_1, ..., X_n$ derived from a sequence are independent and identically distributed.

Now suppose that

$$Y_n = \max\{X_1, ..., X_n\}$$

then using the **Order Statistics** results (chapter 2) it can be shown that (under H_0)

$$F_{Y_n}(y-1) = \mathrm{P}[Y_n < y] = (1 - p_A^y)^n$$

and hence

$$\mathrm{P}[Y_n \geq y] = 1 - (1 - p_A^y)^n$$

Hence, for a formal significance test of the hypothesis H_0, we may use the observed version of Y_n (that is, the sample maximum) as the test statistic, and compute a p-value p

$$p = \mathrm{P}[Y_n \geq y_n] = 1 - (1 - p_A^{y_n})^n$$

Now, note that n must be specified before this p-value can be computed (effectively, we need to choose n large enough) A recommended choice is $n \approx (1 - p_A) N$, giving that

$$p \approx 1 - (1 - p_A^{y_n})^{(1-p_A)N}$$

which using an exponential approximation gives

$$p \approx 1 - \exp\{-(1 - p_A) N p^{y_n}\}$$

Hence, for a test at the $\alpha = 0.05$ significance level, we must check whether the computed p is less than α. If it is, we reject the hypothesis H_0 that the sequence is random and independent.

4.2.2 r-scans

In r-scan analysis, interest lies in detecting short nucleotide "words" in a long genomic segment of length L. If L is very large, then the locations of the occurrences of the words, can be regarded as points in the (continuous) interval $(0, L)$, or, without loss of generality, the interval $(0, 1)$. Suppose that the word is detected a total of k times. An appropriate test of the hypothesis H_0 that the locations are uniformly distributed in $(0, 1)$ can be based on the test statistic random variable Y_{k+1} where

$$Y_{k+1} = \text{“the maximum inter-location length”}$$

(note that the k points segment the genomic interval into $k + 1$ regions of lengths $U_1, U_2, ..., U_{k+1}$. Then

$$Y_{k+1} = \max\{U_1, U_2, ..., U_{k+1}\}$$

Again, to construct the significance test and calculate a p-value, we seek the distribution of this maximum order statistic. However, previous techniques cannot be used, as the random variables $U_1, U_2, ..., U_{k+1}$ are **not independent.** With much work, it can be shown that, as k becomes large, **under** H_0,

$$p = \mathrm{P}[Y_{k+1} \geq y] \approx 1 - \exp\left\{-(k+2)e^{-(k+2)y}\right\}$$

which may be re-written

$$p = \mathrm{P}\left[Y_{k+1} \geq \frac{\log(k+2) + y}{(k+2)}\right] \approx 1 - \exp\{-e^{-y}\}$$

This may be generalized to enable tests based on other test statistics to be carried out, such as those based on "r-scan" values (maxima of the sums of r adjacent inter-point intervals. Finally, as an alternative test, we could instead use the **minimum** order statistic Y_1 Again, after much work, it can be demonstrated that **under** H_0 the p-value is given (approximately) by

$$p = P[Y_1 \leq y] \approx 1 - \exp\{-y(k+2)^2\}$$