Biostatistics, Univ. of Makerere: Auxiliaries, Excercises & Homework
Timo Koski, KTH Royal Institute of Technology, Stockholm
13/08/2018 - 18/08/2018

1 Contents

1.1 Course material

Textbook Warren J. Ewens & Gregory R. Grant: Statistical Methods in Bioinformatics. An Introduction. Second Edition (EG) Springer Verlag, New York 2005.

Other Handouts, Lecture slides

1.2 Course topics

- **Probability (1)** Chapter 1.1− 1.10, 1.13 in EG
- **Probability (2)** Chapter 1.12, 2.1 −2.11 in EG
- **Probability (3)** Chapter 2.12−2.13 in EG
- **Probabilistic inference and measures of information** Chapter 3.9−3.10, kapitel 1.14, 8.3 in EG
- **Markov chains** Chapter 4 in EG
- **Markovian models of DNA, Hidden Markov chains** Chapter 4, Chapter 5. 5.2 - 5.3 in EG chapter 12 in EG
- **Random Walks and BLAST in EG** Chapter 7

- **Markov chains in continuous time** Chapter 11.7, 4.1
- **Molecular evolution and learning of trees** Chapter 14 Chapter 15 Chapter 13
- **Additional** Power laws, Dynamic Programming, Viterbi Algorithm, McMC

1.3 Formulas

A number of formulae and results in probability calculus are quickly recapitulated. This constitutes a very brief summary of the probability calculus to the extent required for these lectures. In a restricted sense this summary makes this self contained. Of course, it does no harm to have more extensive knowledge about the issues touched upon here.

1.4 Formulae

A number of formulae and results in probability calculus are quickly recapitulated. This constitutes a very brief summary of the probability calculus to the extent required by these lectures. In a restricted sense this summary makes this self contained. Of course, it does no harm to have more extensive knowledge about the issues touched upon here.

In addition there is a set of additional exercises, i.e., exercises not found in the Texbook by Ewens and Grant.

2 Recommended Exercises and Homework:

2.1 In the Textbook (Ewens and Grant)

- 1.17, 1.18, 1.19, 1.27, 1.29, 1.31 To be submitted: 1.13, 1.14, 1.25
- 2.1., 2.2, 2.3, 2.10, 2.15, 2.22 To be submitted: 2.17, 2.20
- 3.3., 3.10, 8.2, 8.3 To be submitted: 3.1
- 4.4, 4.5, 4.9, 4.10, 5.5 (i) -(iii) To be submitted: 4.6
- 12.1, 12.4, 11.1, 11.2, 11.4 To be submitted: 12.3, 12.5
- 7.2, 7.3, 7.4, 14.2, 14.3, 14.5 To be submitted: 7.5

2.2 In the Set of Problems in Section 19

TBA

3 Formulae and Definitions

3.1 Alphabet, Sequence

$\mathcal{X}$ and $\mathcal{Y}$ are two discrete *alphabets*, whose generic elements are called *symbols* and denoted by x_i and y_j, respectively.

$$\mathcal{X} = \{x_1, \cdots, x_L\}, \mathcal{Y} = \{y_1, \cdots, y_J\}.$$

$\mid \mathcal{X} \mid :=$ the number of elements in $\mathcal{X}$ = cardinality of $\mathcal{X} = L \leq \infty$, $\mid \mathcal{Y} \mid = J \leq \infty$. Unless otherwise stated the alphabets considered here are finite.

Example 3.1. In analysis and modelling of DNA sequences the alphabet is

$$\mathcal{X} = \{A, T, C, G\}.$$

These are the four bases present in subunits of DNA called *nucleotides*: adenine (A), thymine (T), cytosine (C) and guanine (G). The particular order of these bases arranged along the sugarphosphate backbone is called the *DNA sequence*.

□

More formally we define a sequence as follows.

Definition 3.1. A *sequence* or a *string* $\mathbf{x}$ of length m is an ordered list of m symbols from an alphabet $\mathcal{X}$ written contiguously from left to right

$$\mathbf{x} = x_{l_1} x_{l_2} \ldots x_{l_i} \ldots x_{l_m}; x_{l_i} \in \mathcal{X}, i = 1, \ldots, m.$$

The length of the symbol is denoted by $|\mathbf{x}| = m$. The *empty sequence* $\emptyset$ has length 0 and contains no symbols taken from $\mathcal{X}$.

□

A DNA sequence of length 15 is thus for example

$$\mathbf{x} = \text{ AAAACGTTTAAAAAA}$$

DNA's genetic code can be represented as a single alphabetic sequence composed of these four symbols. It is by means of this code that the gene controls the formation of other substances in the cell. Progress of biotechnology as well as of computer hardware and software have made it possible to determine the nucleotide

sequences for large and increasing number of orgganisms. Such data are now available in computer-readable form, so it is possible to look for and analyze patterns within sequences using probabilistic modelling and computer algorithms.

Tehre are numerous *model families* assigning a probability $P(\mathbf{x})$ to a string (or by extension to a set of strings) that have been applied in bioinformatics. The model families and the methods of probabilistic learning are in this text presented in an order of increasing complexity and flexibility from the multinomial processes to the Markov chains to hidden Markov models. For each of the model families some of its applications in biological sequence analysis are presented.

3.2 Random Variables and their Distributions

X is a (**discrete**) random variable that assumes values in $\mathcal{X}$ and Y is a (**discrete**) random variable that assumes values in $\mathcal{Y}$.

Remark 3.1. These are measurable maps $X(\omega)$, $\omega \in \Omega$, from a basic probability space $(\Omega, \mathcal{F}, P)$ (= outcomes, a sigma field of of subsets of Ω and probability measure P on $\mathcal{F}$). For the most part in the current text this foundational piece of mathematics is suppressed. A very readable survey of these ideas is given by Gray and Davisson (1986), available at
http://ee.stanford.edu/-gray/sp.html.

□

Events of the form $\{\omega \in \Omega \mid X(\omega) = x_i\}$ are thus more simply written as

$$\{X = x_i\} \quad : X \text{ assumes the value } x_i$$

$$\{Y = y_j\} \quad : Y \text{ assumes the value } y_j.$$

Then the probability of the event $\{X = x_i\}$ is

$$f_X(x_i) := P(X = x_i). \tag{3.1}$$

The probability of the event $\{Y = y_j\}$ is

$$f_Y(y_j) := P(Y = y_j). \tag{3.2}$$

Furthermore

$$f_X := (f_X(x_1), \cdots, f_X(x_L))$$

designates a discrete probability distribution on $\mathcal{X}$ and

$$f_Y := (f_X(y_1), \cdots, f_Y(y_J))$$

designates a discrete probability distribution on $\mathcal{Y}$. This means that

$$f_X(x_i) \geq 0 \tag{3.3}$$

$$\sum_{i=1}^{L} f_X(x_i) = 1 \tag{3.4}$$

and similarly for f_Y. If $A \subseteq X$ then

$$P_X(A) := \sum_{x_i \in A} f_X(x_i) \tag{3.5}$$

is the probability of the event that X assumes a value in A, a subset of $\mathcal{X}$. From (3.5) one easily finds the complement rule

$$P\left(A^c\right) = 1 - P(A), \tag{3.6}$$

where A^c is the complement of A, i.e. those outcomes that do not lie in A. Also

$$P\left(A \bigcup B\right) = P(A) + P(B) - P(A \bigcap B), \tag{3.7}$$

is immediate. This constitutes a powerful calculus, since it applies ad verbatim also to compound alphabets with elements that are in turn sequences from finite alphabets or products of other alphabets.

The notation f_X encompasses *tabular* probability distributions, which are simply tables of numbers with the stated properties, and *algorithmic* distributions, which are algorithms for computing probabilities. There are both sorts of probability distributions in bioinformatics as well as combinations thereof. A *parametric* distribution is a special kind of algorithmic distribution: it consists of a few numerical parameters and a relatively simple algorithm, usually a formula for computing probabilities given some specific values of these parameters. The examples immediately following belong to the parametric type.

Example 3.2 (Bernoulli random variables)**.** Consider the binary alphabet $\mathcal{X} = \{0, 1\}$. Let p be a number between zero and one. If X is a random variable assuming values in $\{0, 1\}$ and if

$$f_X(1) = P\left(X = 1\right) = p$$

then we call X a Bernoulli random variable with the *probability of success* p. We write

$$X \in Be(p).$$

We refer to p as a *parameter* of the distribution f_X. The parameter p will be called the *chance of success.* □

Example 3.3 (Geometric random variables)**.** Consider an infinite sequence of Bernoulli random variables $\{X_n\}_{n=1}^{\infty}$ with a common probability of success p. This models an infinite sequence of zeros and ones. Let

$$X = \text{ the first time } (= n) \text{ at which success occurs.}$$

Then $\mathcal{X} = \{1, 2, 3, \ldots, \infty\}$. It can be proved (Helms 1997) that X is a well defined random variable with respect to a suitable probability space. We say that X has *a geometric probability distribution* if

$$f_X(k) = p \cdot (1-p)^{k-1}; k = 1, 2, \ldots, \tag{3.8}$$

This is written as

$$X \in Ge(p).$$

□

Example 3.4 (Uniform random variables)**.** Let $\mathcal{X} = \{x_1, x_2, \ldots, x_L\}$ be a finite alphabet and

$$f_X(x_k) = \frac{1}{L}; k = 1, 2, \ldots, L. \tag{3.9}$$

This is called the *uniform probability distribution*. If X is a random variable assuming values in this alphabet and $P(X = k) = \frac{1}{L}$, then we say that

'X is selected at random'

and we write

$$X \in U(1, L).$$

□

Example 3.5 (Poisson distributed random variable)**.** Let

$$f_X(k) = e^{-\lambda} \cdot \frac{\lambda^k}{k!}, k = 0, 1, 2, \ldots.$$

This is a probability distribution on the non-negative integers called the *Poisson distribution*. A random variable X with this distribution is said to be Poisson distributed and we write

$$X \in Po(\lambda).$$

□

3.3 Joint Probability Distributions

A two dimensional *joint (simultaneous) probability distribution* is a probability defined on the alphabet $\mathcal{X} \times \mathcal{Y}$

$$f_{X,Y}(x_i, y_j) := P(X = x_i, Y = y_j). \tag{3.10}$$

Hence $0 \leq f_{X,Y}(x_i, y_j)$ and $\sum_{i=1}^{L} \sum_{j=1}^{L} f_{X,Y}(x_i, y_j) = 1$.
Marginal distribution for X:

$$f_X(x_i) = \sum_{j=1}^{J} f_{X,Y}(x_i, y_j). \tag{3.11}$$

Marginal distribution for Y:

$$f_Y(y_j) = \sum_{i=1}^{L} f_{X,Y}(x_i, y_j). \tag{3.12}$$

These notions can be extended to define the joint (simultaneous) probability distribution and the marginal distributions of n random variables.

3.4 Conditional Probability Distributions

The conditional probability for $X = x_i$ given $Y = y_j$ is

$$f_{X|Y}(x_i \mid y_j) := \frac{f_{X,Y}(x_i, y_j)}{f_Y(y_j)}. \tag{3.13}$$

The conditional probability for $Y = y_j$ given $X = x_i$ is

$$f_{Y|X}(y_j \mid x_i) := \frac{f_{X,Y}(x_i, y_j)}{f_X(x_i)}. \tag{3.14}$$

Here we assume $f_Y(y_j) > 0$ and $f_X(x_i) > 0$. If for example $f_X(x_i) = 0$, we can make the definition of $f_{Y|X}(y_j \mid x_i)$ arbitrarily through $f_X(x_i) \cdot f_{Y|X}(y_j \mid x_i) = f_{X,Y}(x_i, y_j)$. In other words

$$f_{Y|X}(y_j \mid x_i) = \frac{\text{prob. for the event } \{X = x_i, Y = y_j\}}{\text{prob. for the event } \{X = x_i\}}. \tag{3.15}$$

Hence

$$\sum_{i=1}^{L} f_{X|Y}(x_i \mid y_j) = 1.$$

3.5 A Chain Rule

Let Z be a (discrete) random variable that assumes values in $\mathcal{Z} = \{z_1, z_2, \cdots, z_K\}$. In view of (3.15) we have also, if $f_Z(z_k) > 0$,

$$f_{X,Y|Z}(x_i, y_j \mid z_k) = \frac{f_{X,Y,Z}(x_i, y_j, z_k)}{f_Z(z_k)}.$$

Then we get as an identity

$$f_{X,Y|Z}(x_i, y_j \mid z_k) = \frac{f_{X,Y,Z}(x_i, y_j, z_k)}{f_{Y,Z}(y_j, z_k)} \cdot \frac{f_{Y,Z}(y_j, z_k)}{f_Z(z_k)}$$

and again by definition of conditional probability the right hand side equals

$$= f_{X|Y,Z}(x_i \mid y_j, z_k) \cdot f_{Y|Z}(y_j \mid z_k).$$

In other words,

$$f_{X,Y|Z}(x_i, y_j \mid z_k) = f_{X|Y,Z}(x_i \mid y_j, z_k) \cdot f_{Y|Z}(y_j \mid z_k). \tag{3.16}$$

This chain rule (with extensions) will be invoked in numerous situations.

3.6 Independence

X and Y are *independent* random variables if and only if

$$f_{X,Y}(x_i, y_j) = f_X(x_i) \cdot f_Y(y_j) \tag{3.17}$$

for *all pairs* (x_i, y_j) in $\mathcal{X} \times \mathcal{Y}$. In other words all events $\{X = x_i\}$ and $\{Y = y_j\}$ are to be independent. We say that $X_1, X_2, \ldots, X_n$ are **independent** random variables if and only if the joint distribution

$$f_{X_1,X_2,\ldots,X_m}(x_{i_1}, x_{i_2} \ldots, x_{i_m}) = P\left(X_1 = x_{i_1}, X_2 = x_{i_2}, \ldots, X_m = x_{i_m}\right) \tag{3.18}$$

equals

$$f_{X_1,X_2,\ldots,X_m}(x_{i_1}, x_{i_2} \ldots, x_{i_m}) = f_{X_1}(x_{i_1}) \cdot f_{X_2}(x_{i_2}) \cdots f_{X_m}(x_{i_m}) \tag{3.19}$$

for every m, $2 \leq m \leq n$, and every $x_{i_1}, x_{i_2} \ldots, x_{i_m} \in \mathcal{X}^m$. We are here assuming for simplicity that $X_1, X_2, \ldots, X_n$ take values in the same alphabet. One way of assigning a probability to a sequence $\mathbf{x} = x_{i_1} x_{i_2} \ldots x_{i_n}$ is thus by

$$P\left(\mathbf{x}\right) = f_{X_1}(x_{i_1}) \cdot f_{X_2}(x_{i_2}) \cdots f_{X_n}(x_{i_n}) = \prod_{l=1}^{n} f_{X_l}(x_{i_l}).$$

3.7 Conditional Independence

The random variables X, YY are called *conditionally independent* given Z if

$$f_{X,Y|Z}(x_i, y_j|z_k) = f_{X|Z}(x_i|z_k) \cdot f_{Y|Z}(y_j|z_k) \cdot \tag{3.20}$$

for all triples $(z_k, x_i, y_j) \in \mathcal{Z} \times \mathcal{X} \times \mathcal{Y}$ (cf. (3.16)). We write (3.20) symbolically as

$$X \perp Z|Y.$$

We may also say that $X \longrightarrow Z \longrightarrow Y$ forms a *Markov Chain*. Conditional independence is a core element of the definition of a hidden Markov model.

3.8 Probability Models with Independence

Example 3.6 (Independent Bernoulli random variables)**.** Let $\{X_i\}_{i=1}^n$ be independent Bernoulli random variables $Be\,(p)$ (example 3.2). Then

$$f_{X_1,X_2,\ldots,X_n}(x_{i_1},x_{i_2}\ldots,x_{i_n}) = P\,(X_1 = x_{i_1}, X_2 = x_{i_2},\ldots,X_n = x_{i_n}) =$$

$$= p^{\sum_{j=1}^n x_{i_j}} \cdot (1-p)^{n-\sum_{j=1}^n x_{i_j}} = p^k \cdot (1-p)^{n-k}\,.$$

if $\sum_{j=1}^n x_{i_j} = k$. We may reformulate this statement as $X_1,\ldots,X_n$ being conditionally independent given p.

□

Example 3.7 (Binomial random variables)**.** Let A be some event defined in terms of some basic experiment. Introduce the random variable X= *the number of times the event A occurs in n repetitions of some basic experiment*. Then the alphabet of X is $\mathcal{X} = \{0,1,\cdots,n\}$. If the repetitions are modeled as independent, then X has the *binomial distribution*:

$$f_X(k) = P(X=k) = \left(\begin{array}{c} n \\ k \end{array}\right) p^k (1-p)^{n-k}. \tag{3.21}$$

(Recall that $\left(\begin{array}{c} n \\ k \end{array}\right) = \frac{n!}{(n-k)!k!}$ (the *binomial coefficients*), $0! := 1$). We say that X is a binomial random variable with parameters p and n and write

$$X \in Bin(n,p).$$

Clearly the distibution of X is such that

$$X = X_1 + X_2 + \ldots + X_n,$$

as an equality in distribution, where $X_1, X_2, \ldots, X_n$ are independent $Be\,(p)$ - variables as in the preceding example.

□

If np is small, i.e. if the event A occurs 'rarely', the binomial distribution can be approximated by a Poisson distribution $Po(\lambda)$ with $\lambda = np$ (*the law of small numbers*). The approximation $Bin(n,p) \approx Po(np)$, or more appropriately its generalizations have turned out to be useful for a number of problems in computational biology (Waterman 1995). Poisson approximation is the topic of the exercises 10 - 12.

3.9 Multinomial Probability Distribution

Let $X_1, X_2, \ldots, X_n$ be independent random variables assuming values in

$$\mathcal{X} = \{x_1, \cdots, x_L\}$$

with the common distribution

$$p_l = P(X_i = x_l), l = 1, 2, \ldots, L.$$

Thus

$$P(X_1 = x_{i_1}, X_2 = x_{i_2}, \ldots, X_n = x_{i_n}) = p_{i_1} \cdot p_{i_2} \cdots p_{i_n}.$$

Let for $l = 1, 2, \ldots, L$

$$n_l = \text{ the number of times the symbol } x_l \text{ is found in } \mathbf{x} = x_{i_1} x_{i_2} \ldots x_{i_n}.$$

Thus $n_1 + n_2 + \ldots + n_L = n$. Then from the preceding

$$P(\mathbf{x}) = P(X_1 = x_{i_1}, X_2 = x_{i_2}, \ldots, X_n = x_{i_n}) = p_1^{n_1} \cdot p_2^{n_2} \cdots p_L^{n_L}.$$

If we introduce the random variables $Y_1, Y_2, \ldots, Y_L$ as functions of the random variables $X_1, \ldots, X_n$ by

$$Y_l = \text{ the number of times } X_i = x_l,\ i = 1, 2, \ldots, n,$$

then the joint distribution of $Y_1, Y_2, \ldots, Y_L$ is

$$P(Y_1 = n_1, Y_2 = n_2, \ldots, Y_L = n_L) = C(n_1, n_2, \ldots, n_L) \cdot p_1^{n_1} \cdots p_L^{n_L}. \tag{3.22}$$

This is by independence (symmetry) just the sum of probabilities of all those outcomes of $X_1, \ldots, X_n$ that have exactly $n_1, n_2, \ldots, n_L$ as their frequency counts. Therefore a combinatorial argument (Briggs 1993, p. 97) shows that

$$C(n_1, n_2, \ldots, n_L) = \frac{n!}{n_1! n_2! \ldots n_L!}, \tag{3.23}$$

which is called the *multinomial coefficient*. The probability in (3.22) is called the *multinomial distribution*. Note that the binomial distribution is a special case. The *Whittle distribution* is a generalization of the multinomial distribution for Markov chains.

In the sequel we shall frequently regard the probability assignment

$$P(\mathbf{x}) = P(X_1 = x_{i_1}, X_2 = x_{i_2}, \ldots, X_n = x_{i_n}) = p_1^{n_1} \cdot p_2^{n_2} \cdots p_L^{n_L},$$

as a statement of conditional independence

$$P(\mathbf{x} \mid \underline{p}) = P(X_1 = x_{i_1}, X_2 = x_{i_2}, \ldots, X_n = x_{i_n} \mid \underline{p}) = p_1^{n_1} \cdot p_2^{n_2} \cdots p_L^{n_L},$$

where $\underline{p} = (p_1, p_2, \cdots, p_L)$ is seen as an outcome of a continuous random variable. Since $\underline{p}$ is an unknown parameter, it is supposed constant and is therefore in some literature called a random element.

3.10 A Weight Matrix Model for a Family of Sequences

A weight matrix $\mathcal{M}$ is a simple model often used in biological modelling as a representation for a *family of sequences*. The sequences are supposed to have equal length (=n) and to have no *gaps* (i.e. no positions are blank). A weight matrix $\mathcal{M}$ has as entries the tabular probabilities $p_i(x_j)$ of a string having the symbol x_j at position i:

$$\mathcal{M}: \quad \begin{array}{ccc} p_1(x_1) & \dots & p_n(x_1) \\ p_1(x_2) & \dots & p_n(x_2) \\ \vdots & \vdots & \vdots \\ p_1(x_L) & \dots & p_n(x_L) \end{array}. \tag{3.24}$$

Thus $\sum_{j=1}^{L} p_i(x_j) = 1$ for every column i. The assignment of probabilities for a string $\mathbf{x} = x_{l_1} x_{l_2} \dots x_{l_n}$ conditioned on $\mathcal{M}$ is now given by

$$P(\mathbf{x}|\mathcal{M}) = \prod_{j=1}^{L} \prod_{i=1}^{n} p_i(x_j)^{I_{i,x_j}(\mathbf{x})}, \tag{3.25}$$

where the indicator $I_{i,x_j}(\mathbf{x})$, a function of $\mathbf{x}$, is 0 if $x_j \neq x_{l_i}$ i.e. if the symbol x_j does not appear in position i in the string $\mathbf{x}$ and is 1 otherwise. Thus the symbols in the different positions are conditionally independent given the model $\mathcal{M}$.

A sequence of strings $\mathbf{x}^1, \dots, \mathbf{x}^t$, which are also taken as mutually independent conditioned on $\mathcal{M}$, is by multiplication of the expressions in (3.25) assigned the probability

$$P\left(\mathbf{x}^1, \dots, \mathbf{x}^t|\mathcal{M}\right) = \prod_{s=1}^{t} P(\mathbf{x}^s|\mathcal{M}) = \prod_{j=1}^{L} \prod_{i=1}^{n} p_i(x_j)^{n_i(x_j)}, \tag{3.26}$$

where $n_i(x_j)$ is the number of times the symbol x_j appears on position i in $\mathbf{x}^1, \dots, \mathbf{x}^t$.

The weight matrix model is in bioinformatics often known as a *profile* (Gusfield 1997, ch. 14). In practice the probabilities $p_i(x_j)$ are *estimated* or *learned* using a set of sequences known to belong to a family of sequences e.g. a family of proteins ($L = 20$). One important application is in *database searching* (Gusfield 1997, ch. 15).

3.11 Simplifying Notations

For ease of writing we are sometimes going to drop the symbols in the argument and to write for (3.13)

$$f_{X|Y} = \frac{f_{X,Y}}{f_Y}, \tag{3.27}$$

for (3.14)

$$f_{Y|X} = \frac{f_{X,Y}}{f_X}, \tag{3.28}$$

and for (3.17)

$$f_{X,Y} = f_Y \cdot f_X. \tag{3.29}$$

Eq. (3.29) implies thus

$$X \text{ and } Y \text{ are independent } \Longleftrightarrow \left\{ \begin{array}{ll} f_{X|Y} = f_X & \text{and} \\ f_{Y|X} = f_Y. & \end{array} \right.$$

Conditional independence (3.20) becomes thus

$$f_{Z,X|Y} = f_{Z|Y} \cdot f_{X|Y}. \tag{3.30}$$

Furthermore this is equivalent to

$$f_{Z|X,Y} = f_{Z|Y}. \tag{3.31}$$

Independence of $X_1, X_2, \ldots, X_n$ gives

$$\prod_{l=1}^{n} f_{X_l}.$$

4 Learning and Bayes' Rule

4.1 Bayes' Rule

Since

$$f_{X,Y} = f_{X|Y} \cdot f_Y = f_{Y|X} \cdot f_X$$

we have in a formal way

$$f_{X|Y} = \frac{f_{Y|X} \cdot f_X}{f_Y}.$$

But the marginal distribution f_Y is from (3.12) and (3.28) written as

$$f_Y(y_j) = \sum_{i=1}^{L} f_{Y|X}(y_j \mid x_i) f_X(x_i). \tag{4.1}$$

Hence we have obtained

Bayes' Rule

$$f_{X|Y}(x_i \mid y_j) = \frac{f_{Y|X}(y_j \mid x_i) \cdot f_X(x_i)}{\sum_{i=1}^{L} f_{Y|X}(y_j \mid x_i) f_X(x_i)}. \tag{4.2}$$

Bayes' rule gives a fundamental operation for *up-date of probability distributions* in response to observed information. The rule shows how knowledge about the occurrence of the event $Y = y_j$ is to be used to

transform probabilities on $\mathcal{X}$. In this context we note the possibility of interpreting probability as a degree of rational belief, which makes (4.2) a rule for inductive reasoning.

Terminology for Bayes' Rule :

f_X : A **Prior Distribution** on $\mathcal{X}$.
$f_{X|Y}$: A **Posterior Distribution** on $\mathcal{X}$. If X and Y are independent, then the prior distribution and posterior distribution are identical and there is no *learning* about X from Y and vice versa. Bayes' rule can be seen as just a formal identity derived from certain premises and definitions. One of the stages of probabilistic modelling and learning is to find explicit forms of the distributions, appropriate in some specific application, to be used in Bayes' rule.

4.2 A Missing Information Principle and Inference

Let us suppose that we have a pair of random variables (X, Y). These are thought to occur simultaneously and to assume values in $\mathcal{X} \times \mathcal{Y}$. Let us now suppose that we have observed the event $\{Y = y_j\}$ but are missing information about the corresponding outcome of X (so that X is *hidden* from us).

The *missing information principle* says for our purposes that the posterior $f_{X|Y}(x \mid y_j)$ contains all the probabilistic information about X. This missing information principle is a cornerstone of probabilistic modelling of biological sequences as will become evident in the sequel.

In many situations we want to summarize the posterior $f_{X|Y}(x \mid y_j)$ by a single symbol ('*estimate*') in the alphabet $\mathcal{X}$. This operation of probabilistic inference is called *(point) estimation*. The following two methods of estimation would seem natural:
Maximum Posterior (**MAP**) Estimate, $\widehat{X}_{MAP}$, of X given $Y = y_j$:

$$\widehat{X}_{MAP} = arg \max_{x \in \mathcal{X}} f_{X|Y}(x \mid y_j). \tag{4.3}$$

Maximum Likelihood (**ML**) Estimate, $\widehat{X}_{ML}$, of X given $Y = y_j$:

$$\widehat{X}_{ML} = arg \max_{x \in \mathcal{X}} f_{Y|X}(y_j \mid x). \tag{4.4}$$

5 Some Distributions for DNA Analysis

5.1 Fragment Accuracy

One interesting application of the probability calculus presented so far is the method for evaluating the accuracy of shotgun sequencing given in (Churchill and Waterman 1992). In shotgun sequencing a large DNA molecule is broken into a collection of *fragments*. The fragments are cloned and sequenced individually. The fragments are then assembled by determining their relative orientations and overlaps and aligned

(the technical meaning of this phrase is made more precise in exercise 9. to form a column-by-column matrix.

Suppose that a collection of fragment sequences $\mathbf{x}^1, \ldots, \mathbf{x}^m$ has been aligned by some procedure (more on alignment is found in example 9 in the exercises). The symbol $\emptyset$ is used as a place holder for nonaligned positions beyond the ends of the fragment. The depth of coverage of position i is defined as the number of fragments contributing sequence information at position i i.e.

$$d = \sum_{l=1}^{t} I\left(x_i^l \neq \emptyset\right),$$

where $I\left(x_i^l \neq \emptyset\right) = 1$ if $x_i^l \neq \emptyset$ and zero otherwise. Let in addition

$$\overline{p} = \sum_{x_j \neq x_l, x_j \in \mathcal{X}} P\left(x_i^l = x_j \mid x_l\right),$$

where $P\left(x_i^l = x_j \mid x_i\right)$ is the conditional probability that the shotgun sequenced fragment $\mathbf{x}^l$ equals x_j in position i, given that the true sequence symbol is x_l. It is assumed that $\overline{p}$ is the same for all x_l and i.

For a position covered at depth d, we compute the probability that at most one half of the bases are correct. This is the binomial probability (see example 3.7) given by

$$P\left(\text{error}|d\right) = \sum_{k=0}^{\lfloor d/2 \rfloor} \left(\begin{array}{c} d \\ k \end{array} \right) \overline{p}^{d-k} (1-\overline{p})^k,$$

where the probability of success is taken as $1-\overline{p}$ and $\lfloor d/2 \rfloor$ is largest integer smallaer than or equal to $d/2$. In shotgun sequencing the distribution of d can according to (Churchill and Waterman, 1992) be taken as $Po\left(\lambda\right)$. Positions not covered by any fragments are ignored and thus d is at least 1. Then we get

$$P\left(\text{error}\right) = \frac{1}{1-e^{-\lambda}} \sum_{d=1}^{\infty} \frac{\lambda^d}{d!} e^{-\lambda} P\left(\text{error}|d\right) \tag{5.5}$$

as a practical measure of the accuracy of shotgun sequencing.

5.2 The Distribution of the Number of Fragments

The distribution of fragment lengths from restriction digests in DNA is important in many algorithms used in bioinformatics. The fragment lengths are revealing for DNA *polymorphisms* (roughly speaking, variations in different DNA sequences (Gusella 1986)) and have thus been used for constructing gene maps. In (Bishop et.al. 1983) the distribution of the *number of fragments* is presented. Let N designate the number of fragments of length greater than x base pairs. Fragments shorter than this will not be

detected. Let the number of cuts of the DNA occurring within the probe used (Gusella 1986) be a random variable with the distribution $Po(P/m)$. Here P is the length of the probe and m is the mean length of the fragments. We condition on there being $N = n - 1$ cuts within the probe. Then the probability of seeing i fragments of length greater than x is given (by a well known result independently due to many different authors) as

$$P(i \text{ gaps } |n) = \left(\begin{array}{c} n \\ i \end{array} \right) \sum_{j=0}^{n-i} (-1)^j \left(\begin{array}{c} n-i \\ j \end{array} \right) \left(1 - (i+j)\frac{x}{P}\right)_+^{n-1} \tag{5.6}$$

where

$$a_+ = \max(a, 0).$$

A short proof of (5.6) is found e.g. in (Holst 1980 p.624). Therefore, by the preceding rules of probability calculus we get the probability of i fragments of length greater than x as

$$\begin{aligned} P(N = i) &= \sum_{n=1}^{\infty} \frac{(P/m)^{n-1} e^{-P/m}}{(n-1)!} P(i \text{ gaps } |n) \\ &= \sum_{j=0}^{\infty} \frac{(-1)^j}{i!j!m^{i+j}} \left[P - (i+j)(x-m)\right] (1 - (i+j)x)_+^{i+j-1} e^{-(i+j)/m}. \end{aligned} \tag{5.7}$$

6 Expectation

If $\mathcal{X}$ is a discrete alphabet with a finite set of real numbers as symbols (i.e. $x_i \in \mathcal{X} \subset R$) we define the expectation, $E(X)$, of the random variable X as

$$E(X) := \sum_{i=1}^{L} x_i f_X(x_i). \tag{6.8}$$

In case $L = \infty$ we require absolute convergence of the series. If $\mathcal{X}$ is an arbitrary alphabet and $g(\cdot)$ is a map (function) of $\mathcal{X}$ to R, $g : \mathcal{X} \longmapsto R$,

$$E(g(X)) := \sum_{i=1}^{L} g(x_i) f_X(x_i). \tag{6.9}$$

A speciality of information theory is the use of the following particular case of (6.9) with $g(\cdot) = -\log_2(f_X(\cdot))$,

$$H(X) := E[-\log_2(f_X(X))] := -\sum_{i=1}^{L} \log_2(f_X(x_i)), f_X(x_i) \tag{6.10}$$

$H(X)$ is the *entropy* of X or of f_X, and

$$H(Y) := E[-\log_2(f_Y(Y))] := -\sum_{j=1}^{J} \log_2(f_Y(y_j)) f_Y(y_j). \tag{6.11}$$

Note even the same principle in (the simultaneous entropy)

$$\begin{aligned} H(X,Y) &:= E\left[-\log_2\left(f_{X,Y}(X,Y)\right)\right] \\ &= -\sum_{i=1}^{L}\sum_{j=1}^{J} f_{X,Y}(x_i,y_j)\log_2\left(f_{X,Y}(x_i,y_j)\right) \end{aligned} \tag{6.12}$$

and in (*mutual information*)

$$\begin{aligned} I(X;Y) &:= E\left(\log_2\left(\frac{f_{X,Y}(X,Y)}{f_X(X)f_Y(Y)}\right)\right) \\ &= \sum_{i=1}^{L}\sum_{j=1}^{J} f_{X,Y}(x_i,y_j)\log_2\left(\frac{f_{X,Y}(x_i,y_j)}{f_X(x_i)f_Y(y_j)}\right). \end{aligned} \tag{6.13}$$

7 Jensen's Inequality

Let $\phi(\cdot)$ is a *convex* function on a convex subset of the real line, $\phi : \mathcal{X} \longmapsto R$. Convexity means by definition that

$$\phi\left(\lambda_1 x_1 + \lambda_2 x_2 + \ldots + \lambda_n x_n\right) \leq \lambda_1 \phi\left(x_1\right) + \lambda_2 \phi\left(x_2\right) + \ldots + \lambda_n \phi\left(x_n\right),$$

where $\lambda_1 \geq 0$, $\lambda_2 \geq 0$,..., $\lambda_n \geq 0$ and $\lambda_1 + \lambda_2 + \ldots + \lambda_n = 1$. The basic analytic properties of convex functions and criteria of convexity are found in (Khuri 1993 pp. 78 - 87). But using the language of expectation and the formula (6.9) this is

Jensen's Inequality

$$\phi\left(E\left(X\right)\right) \leq E\left(\phi(X)\right), \tag{7.14}$$

if X is a discrete random variable assuming values in $\mathcal{X}$.

Example 7.1. $\phi(x) = -\log x$ is a convex function. Then, if X is a random variable with positive real values, we have

$$\log\left(E\left(X\right)\right) \geq E\left(\log(X)\right).$$

□

8 Conditional Expectation

Conditional Expectation of X (X is real valued) given $Y = y_j$ is

$$E\left[X|Y = y_j\right] = \sum_{i=1}^{L} x_i f_{X|Y}(x_i|y_j). \tag{8.15}$$

From the preceding defintions we obtain the useful formula

$$E(X) = \sum_{j=1}^{J} E\left[X|Y = y_j\right] \cdot f_Y(y_j).$$

Conditional Expectation of $g(X)$, $g : \mathcal{X} \longmapsto R$, given $Y = y_j$ is

$$E\left[g(X)|Y = y_j\right] = \sum_{i=1}^{L} g(x_i) f_{X|Y}(x_i|y_j). \tag{8.16}$$

Since this is another expectation we have for convex ϕ Jensen's inequality

$$\phi\left(E[X|Y = y_j]\right) \leq E\left[\phi(X)|Y = y_j\right]. \tag{8.17}$$

An information theoretical example of this is given with $g(\cdot) = -\log_2 f_{X|Y}(\cdot|y_j)$,

$$H(X|Y = y_j) := -\sum_{i=1}^{L} \log_2\left(f_{X|Y}(x_i|y_j)\right) \cdot f_{X|Y}(x_i|y_j),$$

which is called the *conditional entropy* of X given $Y = y_j$.

9 Law of Large Numbers

Let $\mathcal{X}$ be a real and finite alphabet and let the *variance*, $Var(X)$, of X be

$$\begin{aligned} Var(X) &:= E\left([X - E(X)]^2\right) \\ &= \sum_{i=1}^{L} (x_i - E(X))^2 f_X(x_i) \end{aligned} \tag{9.18}$$

(by (6.9)). Hence $Var(X) = E(X^2) - (E(X))^2$ ($(E(X)$ is given in (6.8)). Chebysjev's inequality is another partial intuitive interpretation of variance.

Proposition 9.1 (Chebysjev's inequality)**.** *For $k > 0$,*

$$P\left(\mid X - E(X) \mid > k\right) \leq \frac{Var(X)}{k^2}. \tag{9.19}$$

Proof of (9.19):

$$\begin{aligned} Var(X) &= \sum_{i=1}^{L}(x_i - E(X))^2 f_X(x_i) = \\ &= \sum_{\{x_i \mid\mid x_i - E(X)\mid > k\}} (x_i - E(X))^2 f_X(x_i) + \\ &+ \sum_{\{x_i \mid\mid x_i - E(X)\mid \leq k\}} (x_i - E(X))^2 f_X(x_i) \\ &\geq k^2 \sum_{\{x_i \mid\mid x_i - E(X)\mid > k\}} f_X(x_i) \\ &= k^2 P(\mid X - E(X) \mid > k), \end{aligned} \tag{9.20}$$

where we have used the fact that $\sum_{\{x_i \mid\mid x_i - E(X)\mid \leq k\}}(x_i - E(X))^2 f_X(x_i) \geq 0$ and the formula (3.5). □

Let X_k be independent identically distributed copies of X, in the sense that $f_{X_k} = f_X$, for $k = 1, 2, \cdots$, and let $\mathcal{X}$ be a real alphabet. In other words $f_{X_k, X_l} = f_{X_k} \cdot f_{X_l}$ as well as for all higher order distributions. Set $\mu = E(X)$, $Var(X) = \sigma^2$. Let

$$\overline{x}_N := \frac{1}{N}\sum_{k=1}^{N} X_k. \tag{9.21}$$

Then, $E(\overline{x}_N) = \mu$ and

$$Var(\overline{x}_N) = \frac{1}{N^2}\sum_{k=1}^{N} Var(X_k) = \frac{1}{N^2} N\sigma^2 = \frac{1}{N}\sigma^2.$$

Hence Chebysjev's inequality yields

$$\begin{aligned} P\left(|\overline{x}_N - \mu| > k\right) &\leq \frac{Var(\overline{x}_N)}{k^2} \\ &= \frac{1}{N}\frac{\sigma^2}{k^2}. \end{aligned} \tag{9.22}$$

As $N \longrightarrow \infty$ we have thus obtained the following proposition.

Proposition 9.2. *[Law of Large Numbers] If X_k are independent identically distributed copies of X, for $k = 1, 2, \cdots$, and if $\mathcal{X}$ is a real and finite alphabet and $\mu = E(X)$, $Var(X) = \sigma^2$. Then*

$$P\left(|\overline{x}_N - \mu| > k\right) \longrightarrow 0. \tag{9.23}$$

for any $k > 0$.

□

In words this means that *the arithmetic mean of a sequence of independent, equally distributed random variables converges to their common expectation* $= \mu$. The Law of Large Numbers (9.23) can also be stated as for every $\epsilon > 0$ there exists a $N_\epsilon = N(\epsilon, k, \sigma^2)$ such that $N > N_\epsilon$ implies

$$P\left(|\overline{x}_N - \mu| > k\right) < \epsilon."$$

The law of large numbers for binomial random variables means here that

$$P\left(|\frac{X}{n} - p| > k\right) \longrightarrow 0. \tag{9.24}$$

as n increases to $+\infty$ for any $k > 0$.

10 Some Formulas for Dirichlet Densities

10.1 Euler's gamma function

The *gamma* function $\Gamma(z)$ is defined for complex numbers z, whose real part is positive, by the definite integral

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx. \tag{10.1}$$

A special case, obtained by the substitution $x = u^2/2$ is

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}.$$

The recursion formula is

$$\Gamma(z) = (z-1)\Gamma(z-1). \tag{10.2}$$

Hence, if $z = n$, where n is a positive integer, we have the factorial

$$\Gamma(n) = (n-1)!. \tag{10.3}$$

10.2 The Dirichlet density

Let $S_L \subset R^k$ be the *simplex*

$$S_L = \left\{ (\theta_1, \ldots, \theta_L) \,|\theta_i \geq 0, i = 1, \ldots, L, \sum_{i=1}^{L} \theta_i = 1 \right\}. \tag{10.4}$$

Let for $\alpha_i > 0$

$$\phi(\theta_1, \ldots, \theta_L) = \begin{cases} \frac{\prod_{i=1}^{L} \theta_i^{\alpha_i - 1}}{Z}, & \text{if } \theta_1, \ldots, \theta_L \in S_L \\ 0 & \text{otherwise.} \end{cases} \tag{10.5}$$

Here

$$\frac{1}{Z} = \frac{\Gamma\left(\sum_{i=1}^{L} \alpha_i\right)}{\prod_{i=1}^{L} \Gamma(\alpha_i)}. \tag{10.6}$$

The density $\phi(\theta_1, \ldots, \theta_L)$ is called a *Dirichlet density*. We designate it symbolically by

$$Dir(\alpha_1, \ldots, \alpha_L). \tag{10.7}$$

If $\alpha_1 = \alpha_2 = \ldots = \alpha_L = \kappa$, then we talk about a *symmetric Dirichlet density*.

$$\int_{S_L} \phi(\theta_1, \ldots, \theta_L)\, d\theta_1 \ldots d\theta_L = 1 \tag{10.8}$$

This means also that

$$\int_{S_L} \prod_{i=1}^{L} \theta_i^{\alpha_i - 1} d\theta_1 \ldots d\theta_L = \frac{\prod_{i=1}^{L} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{L} \alpha_i\right)}. \tag{10.9}$$

10.3 Beta density

As a special case for $L = 2$ we obtain in (10.9) the *Beta integral*

$$\int_0^1 \theta^{\alpha_1 - 1}(1 - \theta)^{\alpha_2 - 1} d\theta = \frac{\Gamma(\alpha_1) \cdot \Gamma(\alpha_2)}{\Gamma(\alpha_1 + \alpha_2)}. \tag{10.10}$$

Thus

$$f(\theta) = \begin{cases} \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1) \cdot \Gamma(\alpha_2)} \theta^{\alpha_1 - 1} (1 - \theta)^{\alpha_2 - 1} & 0 \le \theta \le 1 \\ 0 & \text{elsewhere.} \end{cases} \tag{10.11}$$

is a probability density called the *Beta density* and denoted by

$$\mathcal{B}e(\theta; \alpha_1, \alpha_2).$$

Note the difference in the heuristic notation between Beta and Bernoulli $Be(p)$. If $\theta = (\theta_1, \ldots, \theta_L)$ is a random variable that assumes values in S_L in (10.4) and has the symmetric $Dir(\alpha, \ldots, \alpha)$ distribution, then the marginal density of any θ_i is given by

$$\theta_i \in \mathcal{B}e(\theta; \alpha, (L - 1)\alpha). \tag{10.12}$$

11 Miscellaneous Probability Results

11.1 Chebysjev's inequality

Chebysjev's inequality is another partial intuitive interpretation of variance. For $k > 0$,

$$P\left(\mid X - E(X) \mid > k\right) \leq \frac{Var(X)}{k^2}.$$

Proof :

$$\begin{aligned} Var(X) &= \sum_{i=1}^{L} (x_i - E(X))^2 f_X(x_i) = \\ &= \sum_{\{x_i \mid\mid x_i - E(X) \mid > k\}} (x_i - E(X))^2 f_X(x_i) + \\ &+ \sum_{\{x_i \mid\mid x_i - E(X) \mid \leq k\}} (x_i - E(X))^2 f_X(x_i) \\ &\geq k^2 \sum_{\{x_i \mid\mid x_i - E(X) \mid > k\}} f_X(x_i) \\ &= k^2 P(\mid X - E(X) \mid > k), \end{aligned} \tag{11.1}$$

where we have used the fact that $\sum_{\{x_i \mid\mid x_i - E(X) \mid \leq k\}} (x_i - E(X))^2 f_X(x_i) \geq 0$. □

11.2 Factorial Moments

X is an integer-valued discrete R.V.,

$$\mu_{[r]} \stackrel{def}{=} E\left[X(X-1)\cdots(X-r+1)\right] =$$

$$= \sum_{x:\text{integer}} \left(x(x-1)\cdots(x-r+1)\right) f_X(x).$$

is called the r:th factorial moment.

11.3 Binomial Moments

X is an integer-valued discrete R.V..

$$E\left(\begin{array}{c} X \\ r \end{array} \right) = E\left[X(X-1)\cdots(X-r+1)\right]/r!$$

is called the binomial moment.

11.4 Probability Generating Function

11.4.1 Definition

Let X have values $k = 0, 1, 2, \ldots,$.

$$G(t) = E\left(t^X\right) = \sum_{k=0}^{\infty} t^k f_X(k)$$

is called the probability generating function.

11.4.2 Prob. Gen. Fnct: Properties

- $$\begin{aligned}\frac{d}{dt}G(1) &= \sum_{k=1}^{\infty} k t^{k-1} f_X(k) \mid_{t=1} \\ &= E[X]\end{aligned}$$

- $$\mu_{[r]} = E[X(X-1)\cdots(X-r+1)] = \frac{d^r}{dt^r}G(1)$$

11.4.3 Prob. Gen. Fnct: Properties

$Z = X + Y$, X and Y integer valued, independent (?),

- $$\begin{aligned}\mathbf{p}_Z(t) = E\left(t^Z\right) = \\ E\left(t^{X+Y}\right) = E\left(t^X\right) \cdot E\left(t^Y\right) = G_X(t) \cdot G_Y(t).\end{aligned}$$

11.4.4 Prob. Gen. Fnct: Examples

- $X \in \mathrm{Be}(p)$
$$\mathbf{p}_X(t) = 1 - p + pt.$$

- $Y \in \mathrm{Bin}(n, p)$
$$\mathbf{p}_Y(t) = (1 - p + pt)^n$$

- $Z \in \mathrm{Po}(\lambda)$
$$\mathbf{p}_Z(t) = e^{\lambda \cdot (t-1)}$$

11.5 Moment Generating Functions

11.5.1 Definition

$$\phi_X(s) \stackrel{def}{=} E\left(e^{sX}\right) = \begin{cases} \sum_{x_i} e^{sx_i} f_X(x_i) & X \text{ discrete} \\ \int_{-\infty}^{\infty} e^{sx} f_X(x)dx & X \text{ continuous} \end{cases}$$

is called the moment generating function.

11.5.2 Moment Gen. Fnctn: Properties

- $$\frac{d}{ds}\phi(0) = E[X]$$
- $$\phi(0) = 1$$
$$\frac{d^k}{ds^k}\phi(0) = E\left[X^k\right].$$

$S_n = X_1 + X_2 + \ldots + X_n$, X_i independent.

$$\phi_{S_n}(s) = E\left(e^{sS_n}\right) =$$
$$E\left(e^{s(X_1+X_2+\ldots+X_n)}\right) = E\left(e^{sX_1}e^{sX_2}\cdots e^{sX_n}\right) =$$
$$E\left(e^{sX_1}\right)E\left(e^{sX_2}\right)\cdots E\left(e^{sX_n}\right) = \phi_{X_1}(s)\cdot\phi_{X_2}(s)\cdots\phi_{X_n}(s)$$

X_i I.I.D.,

$$\phi_{S_n}(s) = (\phi_X(s))^n.$$

11.5.3 Moment Gen. Fnctn: Examples

- $X \in \mathrm{N}(\mu, \sigma^2)$
$$\phi_X(s) = e^{\mu s + \frac{1}{2}\sigma^2 s^2}$$
- $Y \in \exp(\lambda)$
$$\phi_Y(s) = \frac{\lambda}{\lambda - s}$$

11.5.4 A Thm for Bioinformatics

Let X be a discrete R.V. with moment generating function $\phi_X(s)$. Say that X can take at least one negative value (say $-a$) with positive probability $f_X(-a)$ and at least one positive value (say b) with positive probability $f_X(b)$, and that the expectation of X is nonzero. Then there exists a unique nonzero value s^* such that

$$\phi_X(s^*) = 1.$$

12 Some Distributions

12.1 Power Laws

A discrete probability mass function p_k has a power-law tail or is a **Power law**, if it holds that

$$p_k := P(X = k) \sim k^{-\gamma}, \quad \text{as } k \to \infty. \tag{12.1}$$

A probability density function can also have a power-law tail defined in an analogous manner.

Yuel-Simon Distribution

$$p_k = \delta B(\delta + 1, k), k = 1, 2, \ldots, . \tag{12.2}$$

Here $\delta > 0$ is real, $B(\delta + 1, k)$ is the *Beta function*

A probability distribution for a continuous random variable with the density

$$f(x) = \begin{cases} \frac{\alpha q^\alpha}{x^{\alpha+1}} & x > q, \\ 0 & x \leq q, \end{cases} \tag{12.3}$$

where $q > 0$, $\alpha > 0$, is called a **Pareto density** with parameters q and α. The distribution function is thus

$$F(x) = \int_{-\infty}^{x} f(u)du = \begin{cases} 1 - \frac{q^\alpha}{x^\alpha} & x \geq q, \\ 0 & x \leq q. \end{cases} \tag{12.4}$$

12.2 Asymptotic Extreme Value Distribution (EVD)

$X_{\max} = \max(X_1, X_2, \ldots, X_n)$, $X_1, X_2, \ldots, X_n$, I.I.D.

$$\lim_{n\to\infty} P\left(X_{\max} - \frac{\log n}{\lambda} \leq x\right) = e^{-e^{-\lambda x}}.$$

Here log is the natural logarithm.

$$F(x) := e^{-e^{-\lambda x}}$$

is a probability distribution function:

- $F(x) \to 1$, as $x \to \infty$,
- $F(x) \to 0$, as $x \to -\infty$,
- $F(x)$ is an increasing function, since $\frac{d}{dx}F(x) = \lambda e^{-\lambda x - e^{-\lambda x}} > 0$

The distribution is known as Gumbel's distribution.

12.3 Asymptotic Distributions

In general, if $F_1(x), F_2(x), F_3(x), F_4(x), \ldots$ are distribution functions, and

$$\lim_{n\to\infty} F_n(x) = F(x).$$

where $F(x)$ is a distribution function.

We assume that

$$F(x) = 1 - V(x)e^{-\lambda x}$$

for large x, where $\lim_{x\to\infty} V(x) = 1$ is assumed for simplicity. In words, the tail of the distribution function

$$1 - F(x) = V(x)e^{-\lambda x}$$

vanishes exponentially.

12.3.1 Gumbel's distribution

- Gumbel Distribution Function

$$P\left(X_{\max} \leq x\right) \approx e^{-e^{-u}} = G(u)$$

- Gumbel density

$$f(u) = \frac{d}{du}G(u) = e^{-u-e^{-u}}$$

12.3.2 p-value

The p-value of an observed value x of $X_{\max}$ is the probability $P\left(X_{\max} > x\right)$ under some model for the underlying random variables. If these are exponential, then

$$p-\text{value} \approx 1 - e^{-ne^{-\lambda x}}.$$

The smaller the p-value, the less likely is the hypothesis that the model is the correct one.

13 Statistical Inference or Learning

There is one type of learning: inferring, analysing and using a family of models indexed by parameters.

Let $X_1, X_2, \ldots, X_n$ be independent random variables assuming values in

$$\mathcal{X} = \{x_1, \cdots, x_L\}$$

with the common distribution

$$\theta_l = P\left(X_i = x_l\right), l = 1, 2, \ldots, L.$$

Hence $\theta_1 + \theta_2 + \ldots + \theta_L = 1$. Let $\mathbf{x} = x_{i_1} x_{i_2} \ldots x_{i_n}$ be a string of symbols from $\mathcal{X}$ and let for $l = 1, 2, \ldots, L$

$$n_l = \text{ the number of times the symbol } x_l \text{ is found in } x_{i_1} x_{i_2} \ldots x_{i_n}.$$

We set

$$\underline{\theta} = (\theta_1, \theta_2, \ldots, \theta_L)$$

and consider $\underline{\Theta}$ as a random variable (element) that assumes values in the simplex

$$S_L = \{\underline{\theta} \mid \theta_1 + \theta_2 + \ldots + \theta_L = 1, \theta_l \geq 0, l = 1, \ldots, L\}.$$

THE MODEL FAMILY:

CONDITIONED ON $\underline{\Theta} = \underline{\theta}$, THE SYMBOLS IN $\mathbf{x}$ ARE INDEPENDENT.

Thus, as shown before,

$$P\left(\mathbf{x} \mid \underline{\theta}\right) = \theta_{i_1} \cdot \theta_{i_2} \cdots \theta_{i_n} = \theta_1^{n_1} \cdot \theta_2^{n_2} \cdots \theta_L^{n_L}.$$

Again we find a prior $\phi_{\underline{\Theta}}\left(\underline{\theta}\right)$ for Θ. Let us consider the *Dirichlet prior* given by

$$\phi_{\underline{\Theta}}\left(\underline{\theta}\right) = \begin{cases} \frac{\Gamma(\alpha)}{\Gamma\left(\prod_{j=1}^{L} \alpha q_j\right)} \prod_{j=1}^{L} \theta_j^{\alpha q_j - 1} & \underline{\theta} \in S_L \\ 0 & \text{elsewhere,} \end{cases}$$

where *the hyperparameters* are $\alpha > 0$, $q_j \geq 0$, $\sum_{j=1}^{L} q_j = 1$, $\Gamma\left(z\right)$ is Euler's gamma function as given in the appendix. The prior $\phi_{\underline{\Theta}}$ is in (10.7) in the appendix given the symbol

$$Dir\left(\alpha q_1, \ldots, \alpha q_L\right).$$

By extension of Bayes' rule we get the *posterior*

$$\phi_{\underline{\Theta}|\mathbf{x}}\left(\underline{\theta}|\mathbf{x}; \underline{\alpha}\right) = \frac{P\left(\mathbf{x} \mid \underline{\Theta} = \underline{\theta}\right) \cdot \phi_{\underline{\Theta}}\left(\underline{\theta}\right)}{\int_{S_L} P\left(\mathbf{x} \mid \underline{\Theta} = \underline{\theta}\right) \cdot \phi_{\underline{\Theta}}\left(\underline{\theta}\right) d\underline{\theta}}, \underline{\theta} \in S_L \tag{13.1}$$

and zero elsewhere. Using the Dirichlet integral

Proposition 13.1. *The posterior density* $\phi_{\underline{\Theta}|\mathbf{x}}\left(\underline{\theta}|\mathbf{x}; \underline{\alpha}\right)$ *is a Dirichlet density*

$$Dir\left(n_1 + \alpha q_1, \ldots, n_L + \alpha q_L\right)$$

or

$$\phi_{\underline{\Theta}|\mathbf{x}}\left(\underline{\theta}|\mathbf{x}; \underline{\alpha}\right) = \frac{\Gamma\left(n + \alpha\right)}{\prod_{i=1}^{L} \Gamma\left(\alpha q_i + n_i\right)} \prod_{i=1}^{L} \theta_i^{n_i + \alpha q_i - 1}. \tag{13.2}$$

□

This property says that the posterior density is in the same family of densities as the prior. Hence the prior is called *closed under sampling* or a *conjugate prior*.

13.1 Mean Posterior Estimate

One useful property of the Dirichlet density is that we can compute explicitly the expectation of any θ_i with respect to the posterior density. In fact this expectation is by (10.9) and (13.2)

$$\widehat{\theta}_i = \int_{S_L} \theta_i \phi\left(\theta_1, \ldots, \theta_L | \mathbf{x}; \underline{\alpha}\right) d\theta_1 \ldots d\theta_L = \frac{n_i + \alpha q_i}{n + \alpha}. \tag{13.3}$$

This result can be seen as a *regularization* adding pseudocounts αq_i to the vector of observed counts $\underline{n}$ and then normalising so that $\sum_{i=1}^{L} \widehat{\theta}_i = 1$. If we have $n = 0$, the estimate is simply q_i.

Wilson (1927) says that the value of α depends on 'our readiness to to gamble on the typicalness of our experience'.

13.2 Maximum Likelihood

The maximum likelihood estimate of $\underline{\theta}$ (a finite table of probabilities) is by a familiar principle given by

$$\widehat{\underline{\theta}}_{ML} = \operatorname{argmax}_{\underline{\theta} \in S_L} P\left(\mathbf{x} \mid \underline{\theta}\right) = \operatorname{argmax}_{\underline{\theta} \in S_L} \theta_1^{n_1} \cdot \theta_2^{n_2} \cdots \theta_L^{n_L}.$$

The presence of S_L imposes a constrained problem of maximization. We take the natural logarithm of $P\left(\mathbf{x} \mid \underline{\theta}\right)$, which gives us the *loglikelihood function*

$$l\left(\theta_1, \theta_2, \ldots, \theta_L\right) = \log P\left(\mathbf{x} \mid \underline{\theta}\right)$$

We may equivalently seek the maximum of $l\left(\theta_1, \theta_2, \ldots, \theta_L\right)$.

Proposition 13.2. *The maximum likelihood estimate $\widehat{\theta}_{ML}$ of $\underline{\theta}$ is*

$$\widehat{\underline{\theta}}_{ML} = \left(\frac{n_1}{n}, \frac{n_2}{n}, \ldots, \frac{n_L}{n}\right).$$

13.3 Kullback distance

$$D\left(f|g\right) \stackrel{def}{=} \begin{cases} \sum_{x \in \mathcal{X}} f(x) \log \frac{f(x)}{g(x)}, & f,\, g \text{ discrete} \\ \int_{-\infty}^{\infty} f(x) \log \frac{f(x)}{g(x)} dx & f,\, g \text{ continuous} \end{cases}$$

$$D\left(f|g\right) \geq 0$$

This is proved using the **IT−lemma**:

$$\ln(x) \leq (x - 1), \quad x > 0.$$

14 Markov Chains

14.1 Markov Property

A sequence of random variables $\{X_n\}_{n=0}^{\infty}$ with the state space S is called a **Markov chain,(MC)**, if for all $n \geq 1$ and $j_0, j_1, \ldots, j_n \in S = \{1, \ldots, J\}$,

$$P\left(X_n = j_n | X_{n-1} = j_{n-1}, \ldots, X_1 = j_1, X_0 = j_0\right) = P\left(X_n = j_n | X_{n-1} = j_{n-1}\right).$$

$$p_{ij} = P\left(X_n = j | X_{n-1} = i\right)$$

$$P = \left(p_{ij}\right)_{i=1,j=1}^{J,J}$$

$$P\left(X_0 = j_0, X_1 = j_1, \ldots, X_n = j_n\right) = p_{X_0}\left(j_0\right) \prod_{l=1}^{n} p_{j_{l-1} j_l}.$$

14.2 Chapman-Kolmorogorov Equations

For all $m, n \geq 1$ and $i, j \in S$,

$$p_{i|j}(m+n) = \sum_{k=1}^{J} p_{i|k}(m) \cdot p_{k|j}(n).$$

14.3 More Properties, Equilibrium Distribution

$$\pi(n) = \left(p\left(X_n = 1\right), \ldots, p\left(X_n = J\right)\right)$$

$$\pi(n) = \pi(n-1)P = \pi(0)P^n.$$

Every invariant (equilibrium) distibution satisfies the equation

$$\pi = \pi P$$

with the restrictions

$$\sum_{j=1}^{J} \pi_j = 1, \pi_j \geq 0.$$

15 Hidden Markov Models (HMM)

15.1 Definitions

(I) At time $n = 0$ the state X_0 is specified by the initial probability distribution $\pi_j(0) = P\left(X_0 = j\right)$ with

$$\pi\left(0\right) = \left(\pi_1(0), \ldots, \pi_J(0)\right).$$

$\pi_j(n) = P(X_n = j)$

$$\pi(n) = (\pi_1(n), \ldots, \pi_J(n)).$$

(II) **Observable Random Process** A random process $\{Y_n\}_{n=0}^{\infty}$ with a finite state space $\mathcal{O} = \{o_1, o_2, \ldots o_K\}$, where K can be $\neq J$. The processes $\{Y_n\}_{n=0}^{\infty}$ and $\{X_n\}_{n=0}^{\infty}$ are for any fixed n related by the conditional probability distributions

$$b_j(k) = P(Y_n = o_k | X_n = j).$$

We set

$$B = \{b_j(k)\}_{j=1,k=1}^{J,K}$$

and call this the *emission probability matrix*. This is another stochastic matrix in the sense that

$$b_j(k) \geq 0, \sum_{k=1}^{K} b_j(k) = 1.$$

(III) **Conditional independence** For any sequence of states $j_0 j_1 \ldots j_n$ the probability of the sequence $o_0 o_1 \ldots o_n$ is

$$P(Y_0 = o_0, \ldots, Y_n = o_n \mid X_0 = j_0, \ldots, X_n = j_n, B) =$$

$$\prod_{l=0}^{n} b_{j_l}(l).$$

15.2 A Formalism

An HMM is designated by

$$\lambda = (A, B, \pi(0)).$$

UNDER THE HMM ASSUMPTIONS THE STRING $\mathbf{o} = o_0 \ldots o_n$ HAS THE PROBABILITY

$$P(\mathbf{o}) = P(Y_0 = o_0, \ldots, Y_n = o_n; \lambda) =$$

$$\sum_{j_o=1}^{J} \ldots \sum_{j_n=1}^{J} P(Y_0 = o_0 \ldots, Y_n = o_n, X_0 = j_0, \ldots, X_n = j_n; \lambda)$$

$$\sum_{j_o=1}^{J} \ldots \sum_{j_n=1}^{J} P(Y_0 = o_0 \ldots, Y_n = o_n, X_0 = j_0, \ldots, X_n = j_n; \lambda)$$

where

$$P(Y_0 = o_0, \ldots, Y_n = o_n, X_0 = j_0, \ldots, X_n = j_n; \lambda) =$$

$$\pi_{j_0}(0) \cdot \prod_{l=0}^{n} b_{j_l}(l) \prod_{l=1}^{n} a_{j_{l-1}|j_l}.$$

16 Random Walks

Let for $n = 1, \ldots,$

$$S_i = \begin{cases} 1 & \text{with probability } p \\ -1 & \text{with probability } q = 1 - p. \end{cases}$$

Hence S_i are (transformed) I.I.D. Bernoulli R.V's. We take $W_0 = h$. Let

$$W_n = h + \sum_{i=1}^{n} S_i.$$

This is a random motion of a particle that inhabits one of the integer points of the real line. We record the motion of the particle as the sequence $\{(n, W_n)|n =, 1, \ldots, \}$, if these are joined by solid lines between neighbors, the graph is called the *path* of the particle.
We call this the **simple random walk**. The **classical simple random walk** is defined by $p = q = 1/2$.

One interpretation of random walk in terms of bioinformatics is as an accumulated similarity score for two local segments of sequences:

$$W_n(= \overline{s}\left(\mathbf{x}[j:l], \mathbf{y}[j:l]\right)) = \sum_{i=j}^{l} s\left(x_i, y_i\right).$$

where

$$s(x, y) = \begin{cases} +1 & \text{if } x = y \\ -1 & x \neq y. \end{cases}$$

16.1 Absorption probabilities: the m.g.f. method

Let a and b be two integers, $a < b$. We take these points as barriers of absorption in the sense that when the simple random walk hits either of these points it will terminate there. This is a Markov chain, whose state space is the set of integers $= \{a, \ldots, b\}$ and the transition probability matrix of which is of the form

$$\begin{pmatrix} 1 & 0 & 0 & 0 & \ldots & 0 \\ q & 0 & p & 0 & \ldots & 0 \\ 0 & q & 0 & p & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & q & 0 & p \\ 0 & 0 & 0 & \ldots & 0 & 1 \end{pmatrix}.$$

We will be interested in two problems

(i) finding the probability that the walk stops at b rather than at a.

(ii) finding the mean number of steps until the walk terminates.

We use the moment generating functions for both of these problems. The m.g.f. of any single step S is obviously

$$m(\theta) = qe^{-\theta} + pe^{\theta}.$$

We recall a theorem from page 35 in the textbook to show that there exists θ^* such that

$$m(\theta^*) = 1.$$

In fact this is a quadratic equation solved by

$$\theta^* = \log\left(\frac{q}{p}\right).$$

Let

$$N = \min\{n \geq 0 \mid W_n = a \quad \text{or} \quad W_n = b\}.$$

Then N is the random number of steps until the simple random walk terminates and is an example of a **Markov time**. The m.g.f. of the total displacement

$$T_N = \sum_{i=1}^{N} S_i$$

after N steps is, since single steps are I.I.D., by a previous result

$$m(\theta)^N = \left(qe^{-\theta} + pe^{\theta}\right)^N$$

Hence

$$m(\theta^*)^N = \left(qe^{-\theta^*} + pe^{\theta^*}\right)^N = 1.$$

We get also

$$w_h = P(W_N = b) = P(T_N = b - h),$$

and

$$u_h = P(W_N = a) = P(T_N = a - h) = 1 - w_h.$$

The moment generating function of T_N is thus

$$\begin{aligned} m_{T_N}(\theta) &= w_h \cdot e^{(b-h)\theta} + u_h e^{(a-h)\theta} \\ &= w_h \cdot e^{(b-h)\theta} + (1 - w_h)e^{(a-h)\theta}. \end{aligned} \tag{16.1}$$

Then also

$$m_{T_N}(\theta^*) = m(\theta^*)^N = 1,$$

which gives

$$w_h \cdot e^{(b-h)\theta^*} + (1 - w_h)e^{(a-h)\theta^*} = 1$$

$\Leftrightarrow$

$$w_h \cdot \left(e^{b\theta^*} - e^{a\theta^*}\right) = e^{h\theta^*} - e^{a\theta^*}$$

$\Leftrightarrow$

$$\underline{w_h = \frac{e^{h\theta^*} - e^{a\theta^*}}{(e^{b\theta^*} - e^{a\theta^*})}}.$$

Hence we have found the answer to (i) above.

Proposition 16.1. The probability of absorption at b is

$$w_h = \frac{e^{h\theta^*} - e^{a\theta^*}}{(e^{b\theta^*} - e^{a\theta^*})}. \tag{16.2}$$

□

$$E[T_N] = E[N] \cdot E[S].$$

Hence

$$E[N] = \frac{E[T_N]}{E[S]}.$$

The expected time to absorption is thus the ratio of the expected displacement at absorption to the expected step size. But we know that

$$E[T_N] = w_h(b-h) + u_h(a-h),$$

and

$$E(S) = p - q.$$

Hence we have found the solution to (ii) above.

Proposition 16.2. The mean time to absorption is

$$E[N] = \frac{w_h(b-h) + u_h(a-h)}{p-q}. \tag{16.3}$$

□

16.2 An Asymptotic Case

Let us suppose $h = 0$, $a = -1$ and let $b = y$. We are going to let $y \to +\infty$. By this we want to compute

(iii) the distribution of the maximum value of the walk before it hits $a = -1$.

(iv) the mean number of steps until the walk eventually terminates at $a = -1$.

We are going to assume that we are dealing with the case in the theorem from page 35 in the textbook showing that there exists θ^* such that

$$m(\theta^*) = 1$$

with $\theta^* > 0$: Then we have

$$0 < \theta^* = \log\left(\frac{q}{p}\right). \tag{16.4}$$

$$\Leftrightarrow$$

$$1 < \frac{q}{p} \Leftrightarrow p < q \Leftrightarrow 0 < q - p,$$

which means that the random walk has a negative drift.

We discuss (iii), i.e., the maximum value of the walk ever reaches before hitting -1. In (16.2) we have

$$w_h = \frac{e^{h\theta^*} - e^{a\theta^*}}{(e^{b\theta^*} - e^{a\theta^*})},$$

which with $h = 0$, $a = -1$ and $b = y$ yields

$$w_0 = \frac{1 - e^{-\theta^*}}{(e^{y\theta^*} - e^{-\theta^*})}.$$

Hence we can see that for large y, since $\theta^* > 0$ is assumed in (16.4),

$$w_0 \propto \left(1 - e^{-\theta^*}\right) e^{-y\theta^*}.$$

Hence, if Y is the maximum of the walk,

$$\text{Prob}(Y \geq y) \sim \left(1 - e^{-\theta^*}\right) e^{-y\theta^*}$$

as $y \to \infty$, this is a geometric-like probability, where we put

$$C \stackrel{\text{def}}{=} \left(1 - e^{-\theta^*}\right). \tag{16.5}$$

For (iv) we let

$$A \stackrel{def}{=} E[N]. \tag{16.6}$$

Then with $a = -1$, $h = 0$ and $b = y$

$$A = \frac{w_h(b-h) + u_h(a-h)}{p - q}$$

becomes

$$A = \frac{w_0 y - u_0}{p - q} = \frac{u_0 - w_0 y}{q - p}.$$

In the expression above $w_0 y \to 0$, as $y \to \infty$ and $u_0 = 1 - w_0 \to 1$, since $w_0 \to 0$ Hence

$$A = \frac{1}{q - p}$$

as $y \to \infty$.

A and C will be the quantities we shall concentrate upon in the sequel in more general random walks, since these are needed in BLAST calculations.

17 Markov chains in Continuous Time

Let $X = \{X(t) \mid t \geq 0\}$ be a family of random variables taking values in a discrete, countable, alphabet or state space $\mathcal{X}$. The variable t is called time. We denote the generic elements of $\mathcal{X}$ by $j, i, \ldots,$. The special case we have in mind is the finite state space $\mathcal{X} = \{A, T, C, G\}$.

17.1 The Markov property

The process $X = \{X(t) \mid t \geq 0\}$ is called a continuous-time Markov chain if it satisfies the following definition.

Definition 17.1. *$X = \{X(t) \mid t \geq 0\}$ satisfies the* **Markov property**, *if*

$$P\left(X(t_n) = j | X(t_1), X(t_2), \ldots, X(t_{n-1})\right) = P\left(X(t_n) = j | X(t_{n-1})\right) \tag{17.1}$$

for $j \in \mathcal{X}$, $i \in \mathcal{X}$, and any sequence $t_1 < t_2 < \ldots < t_{n-1} < t_n$ of times.

□

The evolution of continuous-time Markov chains can be described in very much the same terms as those used for Markov Chains.

The general situation is as follows. For Markov Chains we wrote the n-step transition probabilities in matrix form and expressed them in terms of the one-step matrix $\mathbf{P}$. In continuous time there is no analogue for $\mathbf{P}$, since there is no implicit unit length of time. Some differential calculus enables us to see that there is a matrix $\mathbf{Q}$, called the **generator** of the continuous-time chain, which takes over the role of $\mathbf{P}$.

17.2 The transition probability

Definition 17.2. *The* **time-homogeneous transition probability** *is denoted by $P_{ij}(t)$ and is defined as*

$$P_{ij}(t) = P\left(X(t) = j | X(0) = i\right) \tag{17.2}$$

or

$$P_{ij}(t - s) = P\left(X(t) = j | X(s) = i\right) \tag{17.3}$$

for $j \in \mathcal{X}$.

□

This is most readily presented in a matrix form.

$$\mathbf{P}(t) = \{P_{ij}(t)\}_{i \in \mathcal{X}, j \in \mathcal{X}}.$$

17.3 Chapman - Kolmogorov equations

Proposition 17.1. *The family $\{\mathbf{P}(t) | t \geq 0\}$ satisfies*

(a) $\mathbf{P}(0) = \mathbf{I}$ (= the identity matrix).

(b) $\mathbf{P}(t)$ *is a stochastic matrix.*

(c) the Chapman - Kolmogorov equations

$$\mathbf{P}(t+s) = \mathbf{P}(t)\mathbf{P}(s). \tag{17.4}$$

□

Assumption 17.1. *We shall now assume that the transition probabilities* $P_{ij}(t)$ *are continuous functions of* t*. We shall also assume that*

$$\mathbf{P}(t) \to \mathbf{I}, \quad \text{as } t \downarrow 0. \tag{17.5}$$

□

This is to say that

$$P_{ij}(t) \to 0, i \neq j, \quad P_{ii}(t) \to 1, i = j, \quad \text{as } t \downarrow 0.$$

17.4 The generator

We make another assumption.

Assumption 17.2. *We assume that*

$$P_{ij}(h) = q_{ij}h + o(h), \quad i \neq j, \tag{17.1}$$

and

$$P_{ii}(h) = 1 + q_{ii}h + o(h), \tag{17.2}$$

where $o(h)$ *('small ordo') is a function such that* $o(h)/h \to 0$*,as* $h \to 0$*.*

□

The numbers q_{ij} are known as the *instantaneous transition rates* or *intensities* of the continuous-time Markov chain. We are here assuming that the probability of two or more transitions in an interval $t, t+h$ is small. This can in fact be proved in a more rigorous treatment. Note that we are also implicitly thinking that the transition rates are not infinite $(-\infty)$.

From (17.1) and (17.2) we get

$$1 = \sum_{j \in \mathcal{X}} P_{ij}(h) = h \sum_{j \in \mathcal{X}, j \neq i} q_{ij} + 1 + q_{ii}h + o(h), \tag{17.3}$$

and this implies

$$h \sum_{j \in \mathcal{X}, j \neq i} q_{ij} = -q_{ii}h + o(h)$$

or, by dividing by h and letting h go to zero,

$$\sum_{j \in \mathcal{X}, j \neq i} q_{ij} = -q_{ii} \tag{17.4}$$

or

$$\sum_{j\in\mathcal{X}} q_{ij} = 0 \tag{17.5}$$

We introduce the symbol q_i by

$$q_i \stackrel{\text{def}}{=} \sum_{j\in\mathcal{X}, j\neq i} q_{ij}. \tag{17.6}$$

The assumption (17.2) gives thus

$$\lim_{h\downarrow 0} \frac{P_{ii}(h) - 1}{h} = q_{ii} = -q_i, \tag{17.7}$$

and the assumption (17.2)

$$\lim_{h\downarrow 0} \frac{P_{ij}(h)}{h} = q_{ij}, i \neq j. \tag{17.8}$$

We introduce the square matrix

$$\mathbf{Q} = (q_{ij})_{i,j\in\mathcal{X}}\,. \tag{17.9}$$

In matrix form (17.7) and (17.8) are

$$\lim_{h\downarrow 0} \frac{\mathbf{P}(h) - \mathbf{I}}{h} = \mathbf{Q}. \tag{17.10}$$

The matrix $\mathbf{Q}$ is called the *generator*.

The generator for a Markov chain on a state space with four elements is in general

$$\mathbf{Q} = \begin{pmatrix} -q_1 & q_{12} & q_{13} & q_{14} \\ q_{21} & -q_2 & q_{23} & q_{24} \\ q_{31} & q_{32} & -q_3 & q_{34} \\ q_{41} & q_{42} & q_{43} & -q_4 \end{pmatrix}.$$

17.4.1 An Example: the Poisson process

Let $X = \{X(t) \mid t \geq 0\}$ be a process that has the set of non-negative integers as the state space. One way of defining X as a Poisson process is to assume the following ((1)−(3)).

(1) The increments of the process are independent or

$$P\left(X(t) - X(s), X(u) - X(v)\right) = P\left(X(t) - X(s)\right) \cdot P\left(X(u) - X(v)\right)$$

for $v < u \leq s < t$.

(2) $X(0) = 0$.

(3) $X(t) - X(s) \in \text{Po}(\lambda(t-s))$.

By these assumptions X is a continuous-time Markov chain. This is found by

$$P\left(X(t_n) = j_n | X(t_1) = j_1, X(t_2) = j_2, \ldots, X(t_{n-1}) = j_{n-1}\right) =$$

$$= P\left(X(t_n) - X(t_{n-1}) = j_n - j_{n-1} | X(t_1) = j_1, X(t_2) = j_2, \ldots, X(t_{n-1}) = j_{n-1}\right)$$

$$= P\left(X(t_n) - X(t_{n-1}) = j_n - j_{n-1}\right)$$

by assumptions (1) and (2), and this equals

$$= P\left(X(t_n) - X(t_{n-1}) = j_n - j_{n-1} | X(t_{n-1}) - X(0) = j_{n-1}\right)$$

$$= P\left(X(t_n) = j_n | X(t_{n-1}) = j_{n-1}\right),$$

again by assumption (1). By assumption (3)

$$P_{ij}(t) = P\left(X(t+s) = j | X(s) = i\right) = P\left(X(t+s) - X(s) = j - i\right)$$

$$= e^{-\lambda t} \frac{(\lambda t)^{j-i}}{(j-i)!}.$$

Then from (17.7) we get

$$\lim_{h \downarrow 0} \frac{P_{ii}(h) - 1}{h} = \lim_{h \downarrow 0} \frac{e^{-\lambda h} - 1}{h} = -\lambda,$$

and from (17.8)

$$\lim_{h \downarrow 0} \frac{P_{ij}(h)}{h} = \lim_{h \downarrow 0} \frac{e^{-\lambda h} \frac{(\lambda h)^{j-i}}{(j-i)!}}{h}$$

$$= \begin{cases} \lambda, & j = i+1 \\ 0 & \text{otherwise.} \end{cases}$$

We have the generator

$$\mathbf{Q} = \begin{pmatrix} -\lambda & \lambda & 0 & 0 & \ldots \\ 0 & -\lambda & \lambda & 0 & \ldots \\ 0 & 0 & -\lambda & \lambda & \ddots \\ \vdots & \vdots & \vdots & \ddots & \ddots \end{pmatrix}.$$

A Poisson proces is thus a continuous time Markov chain with $q_i = \lambda$, so that from (17.6) and (17.4) we have

$$P_{ii}(h) = 1 - \lambda h + o(h),$$

or

$$1 - P_{ii}(h) = \lambda h + o(h),$$

so that the total intensity of leaving i is λ. But, by (3), λ is the intensity of one jump upwards for the proces X.

17.4.2 A Binary Process

Let $Y \in \text{Be}(1/2)$ and $N = \{N(t)|t \geq 0\}$ be a Poisson process like in the previous example, and let Y be independent of N. Set

$$X(t) = (-1)^{Y+N(t)}.$$

Then $X = \{X(t) \mid t \geq 0\}$ is a continuous-time Markov chain, as esentially follows by the same argument as used in the preceding example. The transition matrix is

$$\mathbf{P}(t) = \begin{pmatrix} \frac{1}{2}\left(1+e^{-2\lambda t}\right) & \frac{1}{2}\left(1-e^{-2\lambda t}\right) \\ \frac{1}{2}\left(1-e^{-2\lambda t}\right) & \frac{1}{2}\left(1+e^{-2\lambda t}\right) \end{pmatrix}.$$

The sample paths of X are sequences of -1 and 1, each digit prevailing a random (exponentially distributed time).

17.5 Forward and backward equations

The meaning of the notion of a generator can be explained as follows. Suppose that $X(0) = i$, and by conditioning $X(t+h)$ on $X(t)$ we get by the Chapman-Kolmogorov equations

$$P_{ij}(t+h) = P\left(X(t+h) = j | X(0) = i\right) =$$

$$= \sum_{k \in \mathcal{X}} P_{ik}(t) P_{kj}(h)$$

$$= P_{ij}(t)\left(1 + q_{jj}h + o(h)\right) + \sum_{k \in \mathcal{X}, k \neq j} P_{ik}(t)\left(q_{kj}h + o(h)\right)$$

from (17.1) and (17.2). Then we get

$$= P_{ij}(t) + h \sum_{k \in \mathcal{X}} P_{ik}(t) q_{kj} + o(h).$$

Thus we get that

$$\frac{P_{ij}(t+h) - P_{ij}(t)}{h} = \sum_{k \in \mathcal{X}} P_{ik}(t) q_{kj} + o(h)/h.$$

Hence we have, letting $h \to 0$, and letting $P_{ij}^{'}(t)$ denote the first derivative with respect to t,

$$P_{ij}^{'}(t) = \sum_{k \in \mathcal{X}} P_{ik}(t) q_{kj} = (\mathbf{P}(t)\mathbf{Q})_{ij}.$$

Thus we have derived the following proposition.

Proposition 17.2.

$$P_{ij}^{'}(t) = \sum_{k \in \mathcal{X}} P_{ik}(t) q_{kj}, \tag{17.1}$$

or the matrix **forward equation**

$$\mathbf{P}^{'}(t) = \mathbf{P}(t)\mathbf{Q}. \tag{17.2}$$

□

By a similar way, we can prove

Proposition 17.3.

$$P_{ij}^{'}(t) = \sum_{k \in \mathcal{X}} q_{ik} P_{kj}(t), \tag{17.3}$$

or the matrix **backward equation**

$$\mathbf{P}^{'}(t) = \mathbf{Q}\mathbf{P}(t). \tag{17.4}$$

□

Thus we have the system of differential equations

$$\mathbf{P}^{'}(t) = \mathbf{P}(t)\mathbf{Q}$$

with the initial conditions

$$P(0) = \mathbf{I}.$$

In courses on differential equations the solution is often written using the exponential of a matrix

$$\mathbf{P}(t) = e^{\mathbf{Q}t} = \sum_{l=0}^{\infty} \frac{t^l}{l!} \mathbf{Q}^l. \tag{17.5}$$

17.6 Absolute probabilities

Definition 17.3. *Let $X = \{X(t) \mid t \geq 0\}$ be continuous-time Markov chain. The probability*

$$p_i(t) = P\left(X(t) = i\right)$$

is called the absolute probability for the chain to be in state $i \in \mathcal{X}$ at time t. The vector $\mathbf{p}(t)$ is a row vector whose components are $p_i(t)$. In particular $\mathbf{p}(0)$ is called the initial vector or the initial distribution.

□

The law of total probability gives

$$p_i(t) = P\left(X(t) = i\right) = \sum_{k \in \mathcal{X}} P\left(X(t) = i | X(0) = k\right) P\left(X(0) = k\right) =$$

$$= \sum_{k \in \mathcal{X}} P_{ki}(t) p_k(0),$$

which we write in matrix form as

$$\mathbf{p}(t) = \mathbf{p}(0)\mathbf{P}(t). \tag{17.1}$$

This gives $\mathbf{p}^{'}(t) = \mathbf{p}(0)\mathbf{P}^{'}(t)$. If we multiply the forward equation (17.2) by $\mathbf{p}(0)$ from the left, we get

$$\mathbf{p}^{'}(t) = \mathbf{p}(0)\mathbf{P}^{'}(t) = \mathbf{p}(0)\mathbf{P}(t)\mathbf{Q} = \mathbf{p}(t)\mathbf{Q}. \tag{17.2}$$

If the state space is finite, this is a correct computation, in the case of countable state spaces there are things to be checked. When the equation (17.2) is expressed elementwise, we get

$$p_j^{'}(t) = \sum_{i \in \mathcal{X}} p_i(t) q_{ij} = p_j(t) q_{jj} + \sum_{i \in \mathcal{X}, i \neq j} p_i(t) q_{ij}$$

$$= -q_j p_j(t) + \sum_{i \in \mathcal{X}, i \neq j} p_i(t) q_{ij}.$$

This can be seen as a flow of probabilities. The probability $p_j(t)$ gets an increment that corresponds to the probability that the process is in state i at time t, which is $p_i(t)$ multiplied by the instantaneous transition rate from i to j, q_{ij}. This is summed over all states $i \neq j$. On the other hand $p_j(t)$ is depleted with the probability that the chain is already in the state j multiplied by the instantaneous transition rate to leave the state, i.e., q_j. Inflow minus outfow equals the rate of change $p_j^{'}(t)$.

17.7 Stationary distribution

17.7.1 Definition & the global balance equations

Definition 17.4. *The vector $\pi = (\pi_i)_{i \in \mathcal{X}}$ is a* stationary distribution *of the chain if*

$$\pi = \pi \mathbf{P}(t) \quad \textit{for all } t \geq 0$$

and $\sum_{i \in \mathcal{X}} \pi_i = 1$ and $\pi_i \geq 0$.

□

Thus, (17.1) yields that if $\mathbf{p}(0) = \pi$, then the absolute probabilities are

$$\mathbf{p}(t) = \pi$$

for all $t \geq 0$.

Proposition 17.4.

$$\pi = \pi \mathbf{P}(t) \Leftrightarrow \pi \mathbf{Q} = \mathbf{0}. \tag{17.1}$$

Here $\mathbf{0}$ is matrix of zeros.

Proof:

$$\pi \mathbf{Q} = \mathbf{0}$$

$$\Leftrightarrow$$

$$\pi \mathbf{Q}^n = \mathbf{0} \quad \text{for all } n \geq 1$$

$$\Leftrightarrow$$

$$\sum_{n=1}^{\infty} \frac{t^n}{n!} \pi \mathbf{Q}^n = \mathbf{0} \quad \text{for all } t \geq 0$$

$$\Leftrightarrow$$

$$\pi \sum_{n=0}^{\infty} \frac{t^n}{n!} \mathbf{Q}^n = \pi \quad \text{for all } t \geq 0$$

$$\Leftrightarrow$$

$$\pi \mathbf{P}(t) = \pi \quad \text{for all } t \geq 0 \ ,$$

as was claimed. □

If we write $\pi \mathbf{Q} = \mathbf{0}$ elementwise we get

$$\sum_{i \in \mathcal{X}} \pi_i q_{ij} = 0, \quad \text{for every } j,$$

and this we write as

$$\sum_{i \in \mathcal{X}, i \neq j} \pi_i q_{ij} = -\pi_j q_{jj} = \pi_j q_j. \tag{17.2}$$

The left hand side is interpreted as the flow into the state j, since q_{ij} is the instantaneous transition rate from i to j and this is weighted by π_i, which is the probability that the chain is in the state i. These are summed over all states $i \neq j$. In the same way the right hand side is flow out from the state j, since $q_j = \sum_{k \neq j} q_{jk}$ is the total instantaneous transition rate out from that state. A stationary state is reasonably described by inflow being equal to outflow. The system of equations $\pi \mathbf{Q} = \mathbf{0}$ is called the *global balance equations*. The global balance equations or (17.2) will be used in several evolutionary biological contexts in the next lecture.

17.8 A special generator

We shall next compute the solution to $\pi \mathbf{Q} = \mathbf{0}$ and $e^{\mathbf{Q}t}$, when the generator is of the form

$$\mathbf{Q} = \begin{pmatrix} -(u-u_1) & u_2 & u_3 & u_4 \\ u_1 & -(u-u_2) & u_3 & u_4 \\ u_1 & u_2 & -(u-u_3) & u_4 \\ u_1 & u_2 & u_3 & -(u-u_4) \end{pmatrix}, \tag{17.1}$$

where

$$u = u_1 + u_2 + u_3 + u_4. \tag{17.2}$$

17.9 Rate of change

Proposition 17.5. *If X is a continuous time Markov chain with the generator $\mathbf{Q}$ in (17.1), then*

$$\pi \mathbf{Q} = \mathbf{0}$$

has the solution

$$\pi = \left(\frac{u_1}{u}, \frac{u_2}{u}, \frac{u_3}{u}, \frac{u_4}{u} \right) \tag{17.3}$$

The proof is left to the reader. □

Proposition 17.6. *If X is a continuous time Markov chain with the generator $\mathbf{Q}$ in (17.1), then the rate of change is*

$$R = \lim_{h \downarrow 0} \frac{P\left(X(t+h) \neq X(t)\right)}{h} = u\left(1 - \sum_{i \in \mathcal{X}} \pi_i^2\right). \tag{17.4}$$

where u is given in (17.2).

17.9.1 The exponential of a generator

Now we find

$$\mathbf{P}(t) = e^{\mathbf{Q}t},$$

when the generator is given in (17.1). For this a couple of smart observations are needed. We introduce the matrix

$$\mathbf{A} = \begin{pmatrix} \pi_1 & \pi_2 & \pi_3 & \pi_4 \\ \pi_1 & \pi_2 & \pi_3 & \pi_4 \\ \pi_1 & \pi_2 & \pi_3 & \pi_4 \\ \pi_1 & \pi_2 & \pi_3 & \pi_4 \end{pmatrix}.$$

One notes that

$$\mathbf{Q} = -u\left(\mathbf{I} - \mathbf{A}\right). \tag{17.5}$$

The interesting thing about $\mathbf{A}$ is, that it is idempotent, i.e.,

$$\mathbf{A}^n = \mathbf{A}, \quad \text{for } n \geq 1.$$

This is easily verified by a computation

$$\mathbf{A}^2 = \begin{pmatrix} \pi_1(\pi_1+\pi_2+\pi_3+\pi_4) & \ldots\ldots & \pi_4(\pi_1+\pi_2+\pi_3+\pi_4) \\ \pi_1(\pi_1+\pi_2+\pi_3+\pi_4) & \ldots\ldots & \pi_4(\pi_1+\pi_2+\pi_3+\pi_4) \\ \pi_1(\pi_1+\pi_2+\pi_3+\pi_4) & \ldots\ldots & \pi_4(\pi_1+\pi_2+\pi_3+\pi_4) \\ \pi_1(\pi_1+\pi_2+\pi_3+\pi_4) & \ldots\ldots & \pi_4(\pi_1+\pi_2+\pi_3+\pi_4) \end{pmatrix}$$

$$= \mathbf{A}.$$

Thus $\mathbf{A}^n = \mathbf{A}$ for all $n \geq 1$.

Next we recall that

$$e^{-ut\mathbf{I}} = \sum_{l=0}^{\infty} \frac{(-ut)^l}{l!}\mathbf{I}^l = \mathbf{I}\sum_{l=0}^{\infty} \frac{(-ut)^l}{l!} = e^{-ut}\mathbf{I}.$$

Then we have

$$e^{\mathbf{Q}t} = e^{-ut(\mathbf{I}-\mathbf{A})} = e^{-ut}\mathbf{I}e^{ut\mathbf{A}}$$

$$= e^{-ut}\mathbf{I}\sum_{l=0}^{\infty}\frac{(ut)^l}{l!}\mathbf{A}^l = e^{-ut}\mathbf{I}\left[\mathbf{I}+\sum_{l=1}^{\infty}\frac{(ut)^l}{l!}\mathbf{A}^l\right]$$

$$= e^{-ut}\mathbf{I}\left[\mathbf{I}+\mathbf{A}\sum_{l=1}^{\infty}\frac{(ut)^l}{l!}\right] = e^{-ut}\mathbf{I}\left[\mathbf{I}+\mathbf{A}(e^{-ut}-1)\right]$$

$$= e^{-ut}\mathbf{I}+\mathbf{A}\left(1-e^{-ut}\right).$$

To summarize

$$\mathbf{P}(t) = e^{\mathbf{Q}t} = e^{-ut}\mathbf{I}+\mathbf{A}\left(1-e^{-ut}\right), \quad \text{for } \mathbf{Q} \text{ in (17.1).} \tag{17.6}$$

If we write this elementwise, we get

$$P_{ij}(t) = e^{-ut}\delta_{i,j} + \left(1-e^{-ut}\right)\pi_j, \tag{17.7}$$

where $\delta_{i,j}$ is the Kronecker delta defined by

$$\delta_{i,j} = \begin{cases} 1 & i=j \\ 0 & i\neq j. \end{cases} \tag{17.8}$$

17.10 Separation of Species

We are still assuming $\mathbf{Q}$ in (17.1), and make the additional assumption of *reversibility*. We assume namely first that

$$\pi_i P_{ij}(t) = \pi_j P_{ji}(t) \quad \text{for all } t,\, i,j\ \mathcal{X}. \tag{17.9}$$

This implies by (17.8) even that

$$\pi_i q_{ij} = \pi_j q_{ji}.$$

We see immediately that this is satisfied for $\mathbf{Q}$ in (17.1).

Let us now suppose that we have two continuous-time Markov chains X and Y, with the same generator $\mathbf{Q}$ in (17.1), assuming reversibility, and such that

$$X(0) = Y(0).$$

Then we have

Proposition 17.7. *Assume two continuous-time Markov chains X and Y, with the same generator $\mathbf{Q}$ in (17.1), assuming reversibility, and such that*

$$X(0) = Y(0) \in \pi,$$

but evolving independently thereafter. Then

$$P\left(X(t)=i, Y(t)=j\right) = \pi_i P_{ij}(2t) =$$

$$= \begin{cases} \pi_i\left(1-e^{-2ut}\right)\pi_j & i\neq j, \\ \pi_i e^{-2ut} + \pi_i\left(1-e^{-2ut}\right)\pi_j & i=j. \end{cases} \tag{17.10}$$

17.10.1 Fraction of divergence

Next we compute

$$P\left(X(t) \neq Y(t)\right).$$

From the preceding proposition we get

$$P\left(X(t) \neq Y(t)\right) = \sum_{i \neq j} P\left(X(t) = i, Y(t) = j\right)$$

$$= \sum_{i \neq j} \pi_i P_{ij}(2t) = \sum_{i \neq j} \pi_i \left(1 - e^{-2ut}\right) \pi_j,$$

in view of (17.10). The right hand side equals

$$= \left(1 - e^{-2ut}\right) \sum_{i \neq j} \pi_i \pi_j.$$

Here the sum $\sum_{i \neq j} \pi_i \pi_j$ is actually a double sum

$$\sum_{i \neq j} \pi_i \pi_j = \sum_{i \in \mathcal{X}} \pi_i \sum_{j \in \mathcal{X}, j \neq i} \pi_j =$$

$$= \sum_{i \in \mathcal{X}} \pi_i (\pi_1 + \pi_2 + \ldots \pi_{i-1} + \pi_{i+1} + \ldots) =$$

$$= \sum_{i \in \mathcal{X}} \pi_i (1 - \pi_i) = \sum_{i \in \mathcal{X}} \pi_i - \sum_{i \in \mathcal{X}} \pi_i^2 = 1 - \sum_{i \in \mathcal{X}} \pi_i^2.$$

Now we recall the rate of change from (17.4), and get

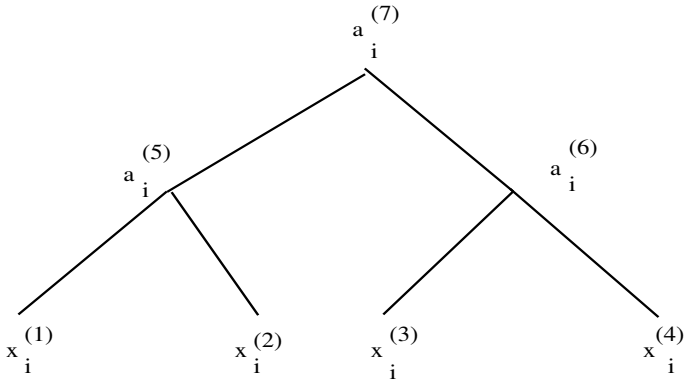
$$P\left(X(t) \neq Y(t)\right) = \frac{R}{u} \left(1 - e^{-2ut}\right).$$

18 Phylogenetic Trees

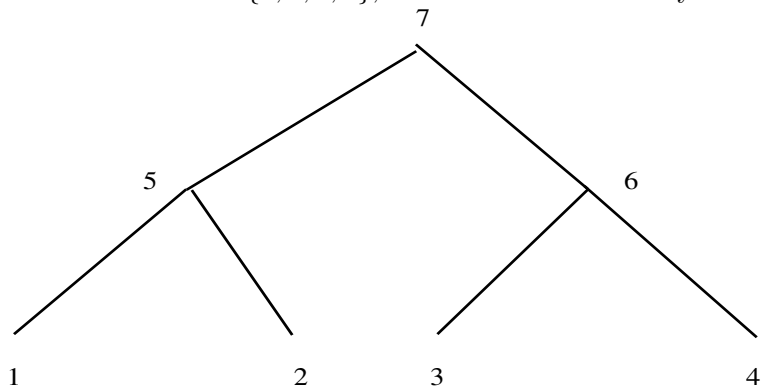
18.1 Likelihood of a Tree with Sequence Data at the Leaves (I)

A rooted binary phylogenetic tree $\mathcal{T}$

$x_i^{(l)}$ is the base at site i for the extant species l, $a_i^{(k)}$ is the base at site i for the ancestral species k.



The leaves are $\mathbf{l} = \{1, 2, 3, 4\}$, the root r is labelled by $r = 7$ $(= 2 \cdot 4 - 1)$.



The joint probability of the bases (ancestral and extant) at site i in the tree is computed as

$$P\left(x_i^{(1)}, x_i^{(2)}, x^{(3)}, x_i^{(4)}, a_i^{(5)}, a_i^{(6)}, a_i^{(7)} | \mathcal{T}\right) =$$

$$\pi_{a_i^{(7)}} \cdot P\left(a_i^{(5)} | a_i^{(7)}\right) \cdot P\left(x_i^{(1)} | a_i^{(5)}\right) \cdot P\left(x_i^{(2)} | a_i^{(5)}\right) P\left(a_i^{(6)} | a_i^{(7)}\right) \cdot P\left(x_i^{(3)} | a_i^{(6)}\right) \cdot P\left(x_i^{(4)} | a_i^{(6)}\right).$$

π is the equilibrium probability.

Note the assumption inherent in this: the probabilities are of the form

$$P(\text{ child } \mid \text{ parent})$$

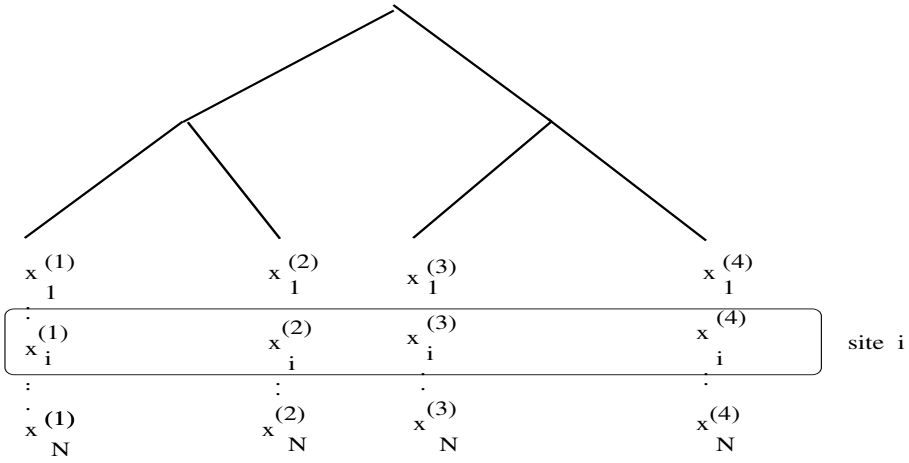
one for each branch of the tree, all are multiplied together. This implies a Markov property on the binary tree under consideration.

We eliminate the unknown ancestral sequences $\{\mathbf{a}^{(5)}, \mathbf{a}^{(6)}, \mathbf{a}^{(7)}\}$, by marginalization. The subscript $_{\text{av}}$ refers to the resulting probability of a site pattern:

$$P_{\text{av}}\left(x_i^{(1)}, x_i^{(2)}, x^{(3)}, x_i^{(4)} \mid \mathcal{T}\right)$$

$$= \sum_{a_i^{(7)} \in \mathcal{X}} \sum_{a_i^{(6)} \in \mathcal{X}} \sum_{a_i^{(5)} \in \mathcal{X}} \pi_{a_i^{(7)}} \cdot \Phi$$

$$\Phi = P\left(a_i^{(5)} | a_i^{(7)}\right) \cdot P\left(x_i^{(1)} | a_i^{(5)}\right) \cdot P\left(x_i^{(2)} | a_i^{(5)}\right) P\left(a_i^{(6)} | a_i^{(7)}\right) \cdot P\left(x_i^{(3)} | a_i^{(6)}\right) \cdot P\left(x_i^{(4)} | a_i^{(6)}\right).$$



Figur 1: Site patterns

18.2 Likelihood of a Tree with Sequence Data at the Leaves (II)

18.2.1 Notations

Let again $\mathcal{T}$ be a binary rooted phylogenetic tree. We let $\mathcal{T}$ designate even its topology. The number of leaves (=number of species) is designated by $L =\mid \mathbf{l} \mid$.

For each of the nodes we have a string $\mathbf{x}^{(k)}$ of N letters from an alphabet $\mathcal{X} = \{x_1, \ldots, x_K\}$, or

$$\mathbf{x}^{(k)} = \left(x_1^{(k)} \ldots x_N^{(k)}\right)$$

for $k = 1, \ldots, L$. The index i corresponds to a site in the genome.

In the internal nodes v and at the root the strings are denoted by

$$\mathbf{a}^{(v)} = \left(a_1^{(v)} \ldots a_N^{(v)}\right),$$

with $a_i^{(k)} \in \mathcal{X}$.

The index i corresponds to a site in the genome. The situation is depicted in Figure 1. The vector

$$x_i^{(1)} x_i^{(2)} \ldots x_i^{(L)}$$

is often called a *site pattern*.

The sequences at the internal nodes are not a part of the data available for calculating the likelihood of the tree, which can only depend on the $\mathbf{x}^{(k)}$.

Assumption 18.1. For any site the site patterns $x_i^{(0)}, \ldots, x_i^{(L)}$ are outcomes of random variables that are independent of any other site.

□

We wish to compute the joint probability of the sequences $\mathbf{x}^{(k)}, k = 1, \ldots, L$ under a fixed tree topology.

Let us denote by

$$P\left(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)} | \mathcal{T}_r\right) \tag{18.1}$$

the joint probability of the data at the leaves of the tree $\mathcal{T}_r$, where r ranges from 1 to $\Psi_r(L)$ = the number of rooted trees with L leaves.

The assumption 18.1 implies that

$$P\left(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)} | \mathcal{T}_r\right) = \prod_{i=1}^{N} P\left(x_i^{(1)} x_i^{(2)} \ldots x_i^{(L)} | \mathcal{T}_r\right). \tag{18.2}$$

The probability of a generic site pattern

$$P\left(x_i^{(1)} x_i^{(2)} \ldots x_i^{(L)} | \mathcal{T}_r\right).$$

19 Exercises

1. **Basic properties of conditional independence** Show that

   a)
   $$X \perp Y | Z \Longleftrightarrow (X, Z) \perp (Y, Z) | Z.$$

   b) If $U = u(X)$ is a function of X and $X \perp Y | Z$ then

   (i)
   $$U \perp Y | Z.$$

   (ii)
   $$X \perp Y | (Z, U).$$

   c) If $X \perp Y | Z$ and $X \perp W | (Y, Z)$, then
   $$X \perp (W, Y) | Z.$$

   (Dawid 1980)

2. **A Markov chain** A sequence of random variables, $\{X_n\}_{n=0}^{\infty}$ assumes values in a finite state space $S = \{1, 2, \ldots, J\}$ with J states. We set
   $$Y_t = (X_{t+1}, X_{t+2} \ldots)$$
   as the future of the chain and
   $$Z_t = (\ldots, X_{t-2}, X_{t-1})$$
   for the past of the chain. Then the Markov property can be stated as
   $$Y_t \perp Z_t | X_t.$$

In particular

$$(X_t, X_{t+1}) \perp Z_{t-1} | X_{t-1}.$$

Show using this and the results above that

$$X_t \perp (Z_{t-1}, Y_{t+1}) \,|\, (X_{t-1}, X_{t+1}) \,.$$

This is the *nearest-neighbor property* of a Markov chain: given the states at all times other than t, the conditional distribution of X_t is in fact determined by the states at times $t-1$ and $t+1$ only.
Hint: Convince yourself of the following

$$X_t \perp Z_{t-1} | \,(X_{t-1}, X_{t+1}) \,,$$

also

$$(Z_{t-1}, X_t) \perp Y_{t+1} | X_{t+1},$$

so that (why ?)

$$X_t \perp Y_{t+1} | \,(X_{t-1}, X_{t+1}, Z_{t-1})$$

and then use the properties of conditional independence established in the preceding exercise.(Dawid 1980)

3. **Fisher's Device:** Let $Y_1, Y_2, \ldots, Y_k$ be independent Poisson distributed random variables. Let $Y_j \in Po(\Psi_j)$, $\Psi_j > 0$, $j = 1, 2, \ldots, k$. Let

$$N = Y_1 + Y_2 + \ldots + Y_k.$$

Show that

$$P\left(Y_1 = n_1, \ldots, Y_k = n_k | N = n\right) = \frac{n!}{n_1! \cdots n_k!} \theta_1^{n_1} \cdots \theta_k^{n_k}$$

where

$$\theta_j = \frac{\Psi_j}{\sum_{j=1}^k \Psi_j}.$$

4. **Trinomial Probabilities:** The trinomial distribution is

$$P\left(n_1, n_2, n_3\right) = \frac{n!}{n_1! n_2! n_3!} \theta_1^{n_1} \theta_2^{n_2} \theta_3^{n_3}.$$

Show that

$$P\left(n_2, n_3 | n_1\right) = \binom{n_2 + n_3}{n_2} \pi_2^{n_2} \left(1 - \pi_2\right)^{n_3},$$

where

$$\pi_2 = \frac{\theta_2}{\theta_2 + \theta_3}.$$

Hint: Note that the claimed distribution is a binomial distribution.

5. Consider the alphabet $\mathcal{X}=\{1,2,3,\ldots\}$ consisting of the positive integers, i.e. $x_j = j, j = 1,2,3,\ldots$. Explain why the function $f_X(j)$ defined for each $j \geq 1$ as

$$f_X(j) = \frac{1}{j(j+1)} = \frac{1}{j} - \frac{1}{j+1}$$

can be regarded as a probability distribution for some random variable X? If X is a random variable with this distribution, does X have an expectation $E[X]$?

6. Let $X \in U(0,100)$. Given that $X = x$, an integer $Y \in U(0,x)$. Compute the expectation $E[Y]$ and the variance $Var[Y]$.

7. **Mullen's ratio theorem** Let X is a random variable assuming values in a finite discrete alphabet of real numbers. The values of X are always positive. Show that

$$E\left(\frac{1}{X}\right) \geq \frac{E(1)}{E(X)}.$$

(K. Mullen, *The American Statistician*, Vol. 21, 1967, pp. 30-31.)

8. Two strings $\mathbf{x}$ and $\mathbf{y}$ with $|\mathbf{x}| = |\mathbf{y}| = n$ of symbols from an alphabet $\mathcal{X} = \{x_1, \ldots, x_J\}$ are considered. Hence e.g.

$$\mathbf{x} = (x_{l_1} x_{l_2} \ldots x_{l_i} \ldots x_{l_m}) ; x_{l_i} \in \mathcal{X}, i = 1, \ldots, m.$$

We define $S(\mathbf{x},\mathbf{y})$ be the number of *matches*, or of positions, where x_{l_i} and y_{l_i} are equal, $1 \leq i \leq n$.

Let $P(S(\mathbf{x},\mathbf{y}) = k)$ be the probability that the number of matches of $\mathbf{x}$ and $\mathbf{y}$ equals k, assuming that the two strings are chosen by some random rule. Assume in fact that $\mathbf{x}$ is selected by independent sampling with replacement from $\mathcal{X}$ with the probabilities

$$p_j^{(1)} = \frac{n_j^{(1)}}{n}, j = 1, \ldots, J$$

and $\mathbf{y}$ is selected (independently of $\mathbf{x}$) by independent sampling with replacement from $\mathcal{X}$ with the probabilities

$$p_j^{(2)} = \frac{n_j^{(2)}}{n}, j = 1, \ldots, J.$$

Prove that

$$P(S(\mathbf{x},\mathbf{y}) = k) = \binom{n}{k} p^k (1-p)^{n-k},$$

where

$$p = \sum_{j=1}^{J} \frac{n_j^{(1)} \cdot n_j^{(2)}}{n^2}.$$

9. **Null Model for Pairwise Alignments**

Definition 19.1. A *(global) alignment* of two sequences $\mathbf{x}$ and $\mathbf{y}$ is obtained by first inserting chosen spaces ('-'), either into or at ends of $\mathbf{x}$ and $\mathbf{y}$, and then placing the two resulting sequences one above the other so that every symbol or space in every in either sequence is opposite to a unique symbol or a unique space in the other sequence. We denote the aligned sequences by $\mathbf{x}^*$ and $\mathbf{y}^*$.

□

Example 19.1. Let $\mathbf{x} = ATAAGC$, $\mathbf{y} = AAAAACG$. To obtain an alignment in the sense defined, one of the many possible ones, we may write

$$\begin{matrix} \mathbf{x}^* \\ \mathbf{y}^* \end{matrix} = \begin{matrix} -\,A\,T\,A\,A\,G\,C\,- \\ A\,A\,A\,A\,A\,-\,C\,G \end{matrix} \tag{19.1}$$

□

This means in the first place that two aligned sequences have same length. By the qualifier *global* we mean that all of the symbols in $\mathbf{x}$ and $\mathbf{y}$ are in the alignment. In a *local* alignment we seek to find substrings that are well matched to each other.

Two strings $\mathbf{x}$ and $\mathbf{y}$ with $|\mathbf{x}| = n$ and $|\mathbf{y}| = m$, $n \leq m$, of symbols from an alphabet $\mathcal{X} = \{x_1, \ldots, x_J\}$ are to be aligned. Let us suppose that the sequences are selected as in the preceding exercise with the probabilities for $\mathbf{x}$ being

$$p_j^{(1)} = \frac{n_j^{(1)}}{n}, j = 1, \ldots, J$$

and $\mathbf{y}$ is selected (independently of $\mathbf{x}$) and by independent sampling with replacement from $\mathcal{X}$ with the probabilities

$$p_j^{(2)} = \frac{n_j^{(2)}}{m}, j = 1, \ldots, J.$$

Let us suppose that g_1 and g_2 spaces $'-'$ are inserted to $\mathbf{x}$ and $\mathbf{y}$, respectively, to get the pairwise alignment. Let us define $P\left(S\left(\mathbf{x}^*, \mathbf{y}^*\right) = k\right)$ as the probability that the number of matches (e.g. $\begin{smallmatrix} A \\ A \end{smallmatrix}$ above) in $\mathbf{x}^*$ and $\mathbf{y}^*$ equals k. Show that

$$P\left(S\left(\mathbf{x}^*, \mathbf{y}^*\right) = k\right) = \binom{m - g_1}{k} p^k \left(1 - p\right)^{m - g_1 - k},$$

where

$$p = \frac{\sum_{j=1}^{J} n_j^{(1)} \cdot n_j^{(2)}}{n \cdot m}.$$

10. **Variation distance.** Let X and Y be two random variables with the non-negative integers as alphabet. Define $d\left(X, Y\right)$ as

$$d\left(X, Y\right) := \frac{1}{2} \sum_{m=0}^{\infty} \left| P\left(X = m\right) - P\left(Y = m\right) \right|. \tag{19.2}$$

Show that this is a *distance function* or a *metric* between the random variables X and Y. A metric is a function such that

1. $d(X,Y) \geq 0$, with equality $d(X,Y) = 0$ iff $X = Y$ (with probability one, i.e. $P(X \neq Y) = 0$).
2. $d(X,Y) = d(Y,X)$. (symmetry")
3. $d(X,Z) \leq d(X,Y) + d(Y,Z)$. (triangle inequality")

11. **Contraction.** Let X, Y and Z be three random variables with the non-negative integers as alphabet. Z is independent of (X,Y). Show that the metric in (19.2) has the property

$$d(X+Z, Y+Z) \leq d(X,Y). \tag{19.3}$$

Hint: Use the identity

$$P(X+Z \in A) = \sum_{m=0}^{\infty} P(X \in A-m, Z=m),$$

where A is any subset (event) of the set of non-negative integers, and $A-m$ is the event A with m excluded.

12. **Poisson Approximation.**

a) Let $X \in Be(p)$, $0 < p < 1$ and $Y \in Po(p)$. Show using the metric in (19.2) that

$$d(X,Y) \leq \frac{1}{2}p^2. \tag{19.4}$$

Hint: Compute first that $d(X,Y) = p \cdot (1 - e^{-p})$.

b) Let $X_1, \ldots, X_n$ be independent and identically distributed $U_i \in Be(p)$, so that $X = X_1 + \ldots + X_n \in Bin(n,p)$. Let $Y \in Po(n\lambda)$. Show using the metric in (19.2) that

$$d(X,Y) \leq \frac{1}{2}n \cdot p^2.$$

Hint: Use an inductive extension of (19.3).

The argument used for this exercise is taken from (Serfling 1978). Poisson approximation can be generalized to deal with sums of dependent Bernoulli distributed variables. The bound on $d(X,Y)$ in (19.4) can be improved, see (Serfling 1978). Applications of Poisson approximation to technical problems in bioinformatics are found in (Waterman 1995, chapter 12.3).

13. **Fragment Accuracy.** Can the sum in (5.5)

$$P(\text{error}) = \frac{1}{1-e^{-\lambda}} \sum_{d=1}^{\infty} \frac{\lambda^d}{d!} e^{-\lambda} P(\text{error}|d)$$

be given a closed form expression?

14. **Number of Fragments.**

a) Prove the result in (5.7).

b) Show that the mean and variance of N, the number of fragments of length greater than x, are given by

$$E[N] = h(1), Var[N] = h(2) + h(1) - h(1)^2,$$

where

$$h(r) = \frac{[P - r(x-m)]}{m^r} [P - rx]^{r-1} e^{-r/m}.$$

In fact $h(r)$ is the *descending factorial moment* (Blom et.al. 1994, p. 24)

$$h(r) = E[N(N-1)\cdots(N-r+1)].$$

(Bishop et.al. 1983).

15. On the Probability Generating Function (p.g.f.) of the Luria-Delbrück Distribution

– Problems **1**. – **3**.

1. Consider a stochastic variable X with values in $\mathcal{X} = \{0, 1, 2, 3, \ldots\}$ and with the probability mass function

$$\begin{aligned} f_X(0) &= P(X = 0) = 0, \\ f_X(k) &= P(X = k) = \frac{1}{k(k+1)}, \quad k = 1, 2, \ldots. \end{aligned} \tag{19.5}$$

(a) Verify that this is in fact a probability mass function. *Aid:* It may be helpful to use the identity

$$\frac{1}{k(k+1)} = \frac{1}{k} - \frac{1}{k+1}.$$

(b) Show that the probability generating function (p.g.f.) $G(t) = E\left[t^X\right] = \sum_{k=0}^{\infty} f_X(k) t^k$ of the distribution in (19.5) is

$$G(t) = 1 + \frac{(1-t)\ln(1-t)}{t}. \tag{19.6}$$

Aid: You may perhaps find it useful to consider the series expansion (from **BETA**)

$$\sum_{k=1}^{\infty} \frac{t^k}{k} = -\ln(1-t), \quad -1 \le t < 1.$$

2. Let N be a stochastic variable with values in $\{0, 1, 2, 3, \ldots\}$. Let

$$f_N(k) = P(N = k), \quad k = 0, 1, 2, \ldots$$

Let $P(t)$ be the the p.g.f. of f_N.

Let $X_1, X_2, \ldots, X_n, \ldots$ be independent and identically distributed random variables with values in $\{0, 1, 2, 3, \ldots\}$ and with the probability mass function

$$f_X(k) = P(X = k), \quad k = 0, 1, 2, \ldots$$

Note that this is an arbitrary probability mass on $\{0, 1, 2, 3, \ldots\}$ and need not be the distribution in (19.5). In words, every X_i has the distribution f_X. Let $Q(t)$ be the p.g.f. of f_X. N is independent of $X_1, X_2, \ldots, X_n, \ldots$.

Consider **a sum of a random number of random variables**

$$S_N = X_1 + X_2 + \ldots + X_N, \quad S_0 = 0. \tag{19.7}$$

The p.g.f. of S_N is

$$G_{S_N}(t) = E\left[t^{S_N}\right].$$

Find an explicit expression for $G_{S_N}(t)$ using the method of Double expectation.

Answer:

$$G_{S_N}(t) = P(Q(t)). \tag{19.8}$$

Aid: The answer and one solution to this exercise is found in Ewens & Grant pp. 78−79 (2001 edition) and p. 90 (2005 edition), in the subsection with the title *'Random n'*. You are, however, expected to use the method in the lecture notes, which is a bit different from the one found in the textbook.

3. Let now $N \in \text{Po}(m)$, $m > 0$, and let $X_1, X_2, \ldots, X_n, \ldots$ be independent and identically distributed random variables with the distribution in (19.5), and let N be independent of $X_1, X_2, \ldots, X_n, \ldots$. We set

$$S_N = X_1 + X_2 + \ldots + X_N, \quad S_0 = 0. \tag{19.9}$$

(a) Find the p.g.f. of S_N in (19.9).
Aid: Use exercises **1**. and **2**. above. You will obviously need the p.g.f. of a Poisson random variable, which is found in Problem 1.11 in Ewens & Grant pp. 78−79 (2001 edition) and in Problem 1.17 (2005 edition).

(b) Show that

$$P\left(S_N = 0\right) = e^m$$

using the p.g.f. established in (a).

In the preceding you have worked out the details of a simple (and formal) mathematical derivation of the p.g.f. of the Luria-Delbrück distribution. There are other derivations that rely on more detailed modelling of mutations.

Those interested in the experiments and analysis underlying the distribution and performed by Salvador E. Luria and Max Delbrück[1] can check Qi Zheng (2010) The Luria-Delbrück Distribution, *CHANCE*, 23:2, 15-18.

16. A probability model for sequences generated as independent samples from the alphabet

$$\{\text{A}, \text{C}, \text{T}, \text{G}\}$$

is

$$P(\text{C}) = P(\text{G}) = 0.275, P(\text{A}) = P(\text{T}) = 0.225.$$

Find the *information content of the distribution.*

17. A sequence of 6 independent letters is is drawn from the alphabet

$$\{\text{A}, \text{C}, \text{T}, \text{G}\}$$

under the probability model

$$P(\text{C}) = P(\text{G}) = 0.275, P(\text{A}) = P(\text{T}) = 0.225.$$

a) What is the probability of getting exactly the sequence
GTTACA ?

[1]The 1969 Nobel Prize in Medicine was awarded to Luria and Delbrück jointly with Alfred D. Hershey for *'their discoveries concerning the replication mechanism and the genetic structure of viruses'*, see
`http://nobelprize.org/nobel_prizes/medicine/laureates/1969/`

b) What is the probability of getting a sequence with twice T, twice A, once G and once C ?

c) What is the conditional probability of getting T twice, A twice given that we have gotten G once and C once ?

18. Let Z_i, $i = 1, 2, \ldots$, be independent random variables with

$$Z_i = \begin{cases} 1 & \text{with probability } 1/4 \\ 0 & \text{with probability } 3/4. \end{cases}$$

Let

$$Y_i = Z_i - \frac{1}{9}(1 - Z_i)$$

The interpretation is that Y_i is the score at the ith position of a pairwise global alignment. The score is 1 for match, $-\frac{1}{9}$ for mismatch. Z_i, $i = 1, 2, \ldots$, is model for background noise for random DNA sequences.

Find the moment generating function of the score over a finite segment

$$Y_1 + Y_2 + \ldots + Y_n.$$

18. In a paper in the journal **CABIOS**, Vol.5, (1989) R. Staden introduced and used probability generating functions, e.g., for calculating the probabilities of scores of DNA words with respect to Position Specific Scoring Matrix (PSSM) $\mathcal{W}$.

The matrix $\mathcal{W}$ has as entries the scores (e.g. frequencies) w_{ij} of nucleotide i at position j of aligned (binding) sites.

We write

$$\{1, 2, 3, 4\} = \{A, T, C, G\}$$

and

$$\mathcal{W}: \begin{array}{ccccccc} A & w_{11} & \ldots & w_{1j} & \ldots & w_{1n} \\ T & w_{21} & \ldots & w_{2j} & \ldots & w_{2n} \\ C & w_{31} & \ldots & w_{3j} & \ldots & w_{3n} \\ G & w_{41} & \ldots & w_{4j} & \ldots & w_{4n} \end{array}$$

The generating function $G_j(t)$ for column j in $\mathcal{W}$ is given by

$$G_j(t) = \sum_{i=1}^{4} p_i t^{w_{ij}}$$

where p_i is the relative frequency (probability) of nucleotide i. The probability generating function $F(t)$ w.r.t. PSSM $\mathcal{W}$ is given by

$$F(t) = \prod_{j=1}^{n} G_j(t).$$

Hence the columns regarded as independent random units.

(a) What is the interpertation of the coefficient of t^k in $F(t)$?

(b) Consider $n = 5$ and a nucleotide count matrix

$$\mathcal{W}: \begin{array}{ccccccc} A & 9 & 1 & 1 & 10 & 7 \\ T & 0 & 0 & 7 & 0 & 0 \\ C & 1 & 9 & 1 & 0 & 0 \\ G & 0 & 0 & 1 & 0 & 3 \end{array}$$

Find the probability of getting PSSM score $= 0$ from columns 1 and 2 using $F(t)$ (or the appropriate factors).

19. **Bayes factor**

We wish to compare two different model families M_i, $i = 0, 1$ proposed for a given sequence $\mathbf{x}$. Under the model family M_i the sequence is related to the parameters θ_i by a distribution $f_i(\mathbf{x} \mid \theta_i)$ and the prior densities for the parameters are $\phi_i(\theta_i)$ for $i = 0, 1$, respectively.

We compare the two families by computing the ratio of the posterior probabilities $P(M_i \mid \mathbf{x})$ of the two model families to the ratio of prior probabilities.

We introduce also

$$q_i(\mathbf{x}) = \int f_i(\mathbf{x} \mid \theta_i) \cdot \phi_i(\theta_i)\, d\theta_i, i = 0, 1.$$

Then Bayes$'$ factor, B_f is thus defined as

$$B_f = \frac{\text{posterior odds ratio}}{\text{prior odds ratio}}.$$

and Bayes$'$ rule gives

$$= \frac{\frac{P(M_0|\mathbf{x})}{P(M_1|\mathbf{x})}}{\frac{P(M_0)}{P(M_1)}} = \frac{q_0(\mathbf{x})}{q_1(\mathbf{x})},$$

where $P(M_i)$ is the prior for M_i for $i = 0, 1$.

Consider now

M_0 : Poisson distribution with unknown parameter

M_1 : Geometric distribution with unknown parameter

– Poisson Distribution

$$f_0(x \mid \theta_0) = e^{-\theta_0} \frac{\theta_0^x}{x!}, x = 0, 1, 2, \ldots,$$

$$x_i|\theta_0 \in f_0(x|\theta_0), \text{ I.I.D. },$$

$$\mathbf{x} = (x_1 x_2 \ldots x_n)$$

$$f_0(\mathbf{x} \mid \theta_0) = e^{-n\theta_0} \frac{\theta_0^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!}$$

– Geometric Distribution

$$f_1(x \mid \theta_1) = \theta_1 \cdot (1-\theta_1)^x, x = 0, 1, 2, \ldots,$$
$$x_i|\theta_1 \in f_1(x|\theta_1), \text{ I.I.D. },$$
$$\mathbf{x} = (x_1 x_2 \ldots x_n)$$
$$f_1(\mathbf{x} \mid \theta_1) = \theta_1^n \cdot (1-\theta_1)^{\sum_{i=1}^n x_i}$$

The prior densities are

– Gamma distribution

$$\theta_0 \in \mathcal{G}\mathrm{a}(k, \lambda),$$

see Ewens & Grant.

– Beta distribution

$$\theta_1 \in \mathcal{B}\mathrm{e}(\alpha_1, \beta_1)$$

a) Establish that the Bayes factor is

$$B_f = \frac{\frac{\lambda^k}{\Gamma(k)\prod_{i=1}^n x_i!} \cdot \frac{\Gamma(k+\sum_{i=1}^n x_i)}{(n+\lambda)^{k+\sum_{i=1}^n x_i}}}{\frac{\Gamma(\alpha_1+\beta_1)}{\Gamma(\alpha_1)\Gamma(\beta_1)} \cdot \frac{\Gamma(n+\alpha_1)\Gamma(\sum_{i=1}^n x_i+\beta_1)}{\Gamma(n+\sum_{i=1}^n x_i+\alpha_1+\beta_1)}}.$$

b) Assume that

$$k = \lambda = 2,$$

and

$$\alpha_1 = \beta_1 = 1.$$

Which distribution, known under another name, is this ? Assume that

$$x = (1, 1, 2, 0, 0, 0, 0, 1, 1, 1, 2, 3).$$

Which model family is now favoured by B_f? You may use the expression in a) even if you have not obtained it.

19. Let $\{X_n\}_{n\geq 0}$ be a finite Markov chain with stationary transition probabilities $p_{i|j}$. It has been shown that then

$$P(X_0 = j_0, X_1 = j_1, \ldots, X_n = j_n) = p_{X_0}(j_0) \prod_{l=1}^n p_{j_{l-1}|j_l}.$$

Check that the right-hand side of this equality is in fact a probability distribution.

20 **A Model of Recombination Processes Incorporating Interference**

Let us consider a genome and suppose that between any two markers there is an array of discrete positions at which crossovers (of strands of homologus chromosomes) can occur. Think of the positions as ordered from left to right. The following Markov model addresses a locking mechanism as vehicle of interference.

A position is said to be in state 0 if it is locked or unlocked with no crossover there, and in state 1 if a crossover occurs there. When a position is *locked*, then a crossover occurs at the next position with probability p, or no crossover occurs with probability $1-p$. In absence of crossover, then the next position is *unlocked*.

When a crossover occurs, then with probability λ the next position is locked, and with probability $1-\lambda$ either a crossover results (with probability p) or no crossover results (with probability $1-p$). The recombination process is then modeled as a two-state Markov chain on the state space $S = \{0, 1\}$ governed by the transition probability matrix

$$\mathbf{P} = \begin{pmatrix} 1-p & p \\ \lambda + (1-\lambda)(1-p) & (1-\lambda)p \end{pmatrix}$$

a) Find the invariant distribution of this chain. (This is of importance for calculating the recombination frequency between two loci with L intervening positions.)

b) Take $\lambda = 0.7$, and $p = 0.2$. If the distribution $\phi(0)$ at the leftmost position is $(0.8, 0.2)$, what is the probability of crossover at position $n = 2$?

21. Let a Markov chain with states $\mathcal{S} = \{0, 1\}$ have the transition probability matrix

$$\mathbf{P} = \begin{pmatrix} 0.75 & 0.25 \\ 0.25 & 0.75 \end{pmatrix}$$

Take the initial distribution

$$\phi(0) = (1, 0).$$

Show (using induction) that

$$\phi(n) = \left(\frac{1}{2}\left(1 + 2^{-n}\right), \frac{1}{2}\left(1 - 2^{-n}\right) \right)$$

Show that the invariant distribution is the limit of $\phi(n)$, as $n \to \infty$.

22. A simplified version of the Wright model of population genetics is as follows. X_n is the number of individuals bearing the genetic configuration A in a population with N individuals and is a Markov chain $\{X_n\}_{n \geq 0}$ with the state space $S = \{0, 1, \ldots, N\}$ and with the transition probabilities

$$p_{i|j} = P\left(X_{n+1} = j | X_n = i\right) = \begin{pmatrix} N \\ j \end{pmatrix} \left(\frac{i}{N}\right)^j \left(1 - \frac{i}{N}\right)^{N-j}$$

In other words, $X_{n+1}|X_n = i \in \text{Bin}\left(N, \frac{i}{N}\right)$.

(a) Is this Markov chain irreducible ? Hint: Check $p_{0|0}$ and $p_{N|N}$.

(b) Compute

$$E\left[X_{n+1}|X_0 = i_0, X_1 = i_1, \ldots, X_{n-1} = i_{n-1}, X_n = i\right].$$

23. Let $\{X_n\}_{n\geq 0}$ be a finite Markov chain with stationary transition probabilities $p_{i|j}$. Let $\mathbf{x} = j_0 j_1 \ldots j_n$. It has been shown that the probability of this sequence w.r.t to a Makrov model is

$$\begin{aligned} P\left(\mathbf{x} \mid \text{Markov}\right) &= P\left(X_0 = j_0, X_1 = j_1, \ldots, X_n = j_n\right) \\ &= p_{X_0}\left(j_0\right) \prod_{l=1}^{n} p_{j_{l-1}|j_l}. \end{aligned} \tag{19.10}$$

Check that the right-hand side of this equality is in fact a probability mass function.

24. Consider the transition probability matrix for a Markov chain with four states (like nucleotide bases)

$$\mathbf{P} = \begin{pmatrix} p_{1|1} & p_{1|2} & p_{1|2} & p_{1|4} \\ p_{2|1} & p_{2|2} & p_{2|3} & p_{2|4} \\ p_{3|1} & p_{3|2} & p_{3|3} & p_{3|4} \\ p_{4|1} & p_{|2} & p_{4|3} & p_{4|4} \end{pmatrix}$$

We use now (19.10) (with four states) as the likelihood function for $\mathbf{P}$, or

$$L\left(\mathbf{P}\right) = \prod_{l=1}^{n} p_{j_{l-1}|j_l} \propto P\left(\mathbf{x} \mid \text{Markov}\right).$$

In other words $L\left(|\mathbf{P}\right)$ is regarded as a function of $\mathbf{P}$ (or of the probabilities in $\mathbf{P}$) and we throw away $p_{X_0}\left(j_0\right)$.

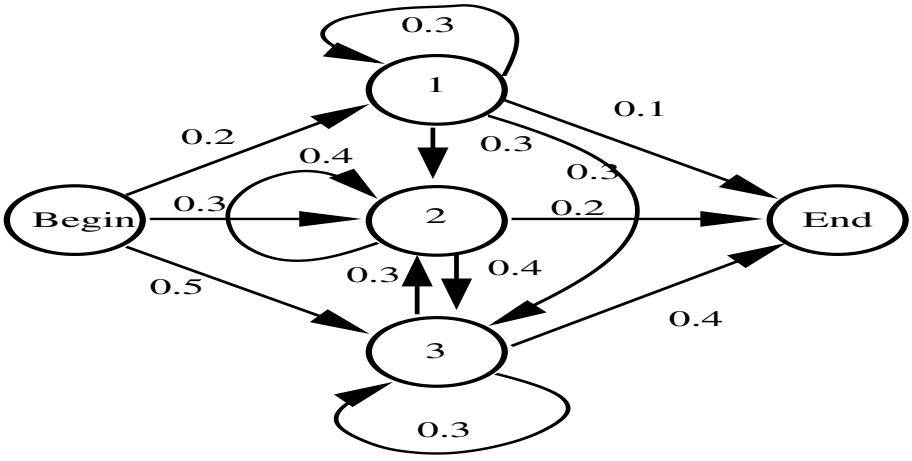
Show that the maximum likelihood estimate $\widehat{p}_{i|j}$ of $p_{i|j}$ is

$$\widehat{p}_{i|j} = \frac{n_{i|j}}{n_i}, \text{ for } i = 1, 2, 3, 4 \text{ and } j = 1, 2, 3, 4.$$

Here $n_{i|j}$ is the number of times the sequence $\mathbf{x} = j_0 j_1 \ldots j_n$ contains the pair of bases (i, j) (in this order), i.e., the number of transitions from i to j and n_i is the number of times the base i occurs in the sequence. *Hint*: Note that this separates to an independent constrained maximization problem for each row.

25. Consider the HMM with three hidden states $\mathcal{S} = \{1, 2, 3\}$, the transition probability matrix A given by the graph in in Figur 2, and with the emission alphabet $\mathcal{O} = \{a, b\}$, and with the emission probability matrix B given by

$$\begin{aligned} b_1(a) = 0.5, b_2(b) = 0.5, \\ b_2(a) = 0.1, b_2(b) = 0.9, \\ b_3(a) = 0.9, b_3(b) = 0.1. \end{aligned}$$



Figur 2: HMM state transition graph and transition probabilities for exercises 1. and 2.

Let the observed sequence be $\mathbf{o} = bab$. Find the Viterbi path $j^* = j_0 j_1 j_2$. Show your calculations.
Answer:

$$\max_{j_0 j_1 j_2} P(\mathbf{o}, j_0, j_1, j_2) = 0.0052488$$

$$j^* = 232$$

26. Consider the HMM in exercise 26. Find $P(aab)$ using the forward algorithm. Show your calculations.
Answer:

$$P(aab) = 0.0132864.$$

27. Let $\{X_n | n = 0, 1, 2, \ldots\}$ be time homogeneous Markov chain with a finite state space $\mathcal{S}$. Show, in pedantic detail, that X_{n+1} and X_{n-1} are conditionally independent given X_n, or

$$P(X_{n+1} = j_{n+1}, X_{n-1} = j_{n-1} | X_n = i) =$$

$$P(X_{n+1} j_{n+1} | X_n = i) \cdot P(X_{n-1} = j_{n-1} | X_n = i)$$

for all triplets j_{n+1}, i, j_{n-1} from $\mathcal{S}^3$.

if X has $\mathcal{B}in(n, p)$ distribution with $p \leq 0.1$, then X is approximatively $\mathcal{P}oi(np)$.

28. A **palindrome** is a word, phrase, number, or other sequence of characters which reads the same backward or forward. A restriction enzyme is cutting DNA at a palindromic site 6 nt long.

Determine the probability that a circular chromosome, a double-stranded DNA molecule of length $L = 84000$ nt, will be cut by the restriction enzyme into exactly twenty fragments.

We assume the DNA die with independent tosses as the description of the DNA sequence. Approximation of a binomial distribution by a Poisson distribution is required.

29. A circular double-stranded DNA of $L = 3400$ nt long was cut by a restriction enzyme. A subsequent gel electrophoresis separation indicated the presence of five DNA pieces. It turned out that the absent-minded technician could not recall exact type of restriction enzyme that was used. Still, he knew that the chemical was picked up from a box containing an equal number of 4-base cutters and 6-base cutters (restriction enzymes that cut specific 4 nt long sites and specific 6 nt long sites, respectively).

    What is the posterior probability that 4-nucleotide cutter was used?

    We assume the DNA die with independent tosses as the description of the DNA sequence. Approximation of the number of restriction sites by a Poisson distribution is permitted.

30. C. Darwin has formulated the following problem[2] on what might be called rare deviations:

    > *Let it be assumed that, in a large population, a particular affection occurs ... in one out of a million, so that the á priori chance that an individual taken at random will be so affected is only one in a million. Let the population consist of sixty millions, composed, we will assume, of ten million families, each containing six members.*

    Darwin assumes without any doubt that a family has two parents and four children.

    The Question (by Darwin)

    > *What are the odds that there will not be even a single family in which at least one parent and two children will be affected?*

    Let X = the number families thus affected, i.e., the number of families with in which at least one parent and two children are affected. The total number of families is $n = 10^7$.

    a) What is the probability in a certain family to have at least one parent and two children affected? Assume independent draws from the overall population.

    b) What is the distribution of X?

    c) What is now $P(X = 0)$? Approximation of distribution expected.

    d) What is $P(X > 0)$? The approximation $P(X > 0) \approx P(X = 1)$ is applicable.

    e) Find the desired odds. Darwin continues (loc.cit)

    > *Professor Stokes* [3] *has calculated for me that the odds will be no less than 8333 millions to 1.*

[2] *The variation of animals and plants under domestication.* London: John Murray. 1st ed, 1st issue. Volume 2., Ch XII, p. 5, c.f. `http://darwin-online.org.uk/converted/published/1868_Variation_F877/1868_Variation_F877.2.html`

[3] (maybe this is George Gabriel Stokes of the Stokes' formula fame ?)

Compare now your answer with the result by Stokes.

f) Darwin states that such families do exist in England (he observes $X = x > 0$). What conclusion does he/do you draw from this ?

31. **Lander-Watermans statistisc foor shotgun sequencing**

Method
Frederick Sanger was one of the laureates of the Nobel Prize in Chemistry in 1980 for developing a method to sequence short regions of DNA. It was the most widely used sequencing method for approximately 40 years since its invention.

- If the sequence is larger than 500-1000 consecutive nucleotides, the rest of it will not be read (?). There is no current technology to simply read the whole genome sequence from one end to the other. The human genome is 3 billion nucleotides long. Sequencing it using the Sanger method requires breaking it into little pieces, sequencing the pieces separately, and fitting them back together.
- Break DNA at random into many smaller pieces, and randomly select a large number of these pieces to be sequenced. Approximately the first 500 nucleotides are read from one end of the pieces. These small sequenced regions are called **reads**. We do not know their location in the genome or their strand. Combine overlapping reads into **contigs**. Sequence alignment is used to detect the overlaps.
- Additional information (scaffolds) is used to place the contigs into the proper order and direction on chromosomes.

Notations and statistical assumptions

- G =genome length in nucleotides $\approx$ 3 billion in human
- L = read length in nucleotides (assume 500)
- N = number of reads sequenced
- NL = number of nucleotides in all sequenced reads
- $a = \frac{N}{G}L$ is the **coverage** (=average number of times each nucleotide in the whole genome is sequenced)
- In each chromosome, a read of length L could start anywhere except the last $L-1$ positions.
  - In a genome of length G with c chromosomes, there are $G-c\cdot(L-1)$ possible starting positions. For human, $c\cdot(L-1) = 23(499) = 11477 << G$ so we will approximate that there are G possible starting positions. (That is, we will ignore the end effects.) The probability that one of the N reads starts at any specific nucleotide is N/G.
  - Assume reads are distributed uniformly through the genome and independently of each other.

Note that the exercises to follow do not require any extensive calculations.

a) Let I be any interval of L consecutive nucleotides. Let X = number of reads starting in I.

a.i) Under the preceding assumptions, why is the distribution of X is binomial? *Hint*: success= a read starts at a nucleotide.

a.ii) What are the parameters of the binomial distribution?

a.iii) What is the probability of no reads in I? What is the probability of at least one read starting in I, i.e., the probability of a contig?

b) Approximate the binomial distribution in **a)** by the appropriate Poisson distribution.

b.i) What are now the answers to **a.ii) - a.iii)**?

b.ii) We need to define gaps:

The nucleotide r is in a **gap**, if no read starts within the interval $[r-L+1, r]$. In **b.i)** we found the probability of a gap.

What is your back-of-the-envelope estimate[4] of the expected number of nucleotides in gaps? What is your estimate of the expected number of nucleotides in contigs?

[4]A back-of-the-envelope calculation is a rough calculation, typically jotted down on any available scrap of paper such as the actual back of an envelope.

b.iii) Assume that 99% of the genome is in contigs and 1 % in gaps. What is the coverage of, e.g., the human genome? Comment on your finding.

c) Each of the contigs has a unique rightmost read. The probability that a read is the rightmost read equals the probability that no other read starts within that read, and has been found in **b.i)**. If you label the rightmost read as success and others as failures. With Y = the number of successes, Y has a binomial distribution.

c.i) What are the parameters of this distribution?

c.ii) What is the expected number of contigs expressed in terms of a, G and L ?

32. The transition probability matrix for a finite Markov chain is

$$\mathbf{P} = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

Find $P(2n)$ and $P(2n-1)$ for all $n \geq 1$, and discuss whether the chain can have an asymptotic distribution.

33. Let a Markov chain with states $\mathcal{S} = \{0, 1\}$ have the transition probability matrix

$$P = \begin{pmatrix} 0.7 & 0.3 \\ 0.4 & 0.7 \end{pmatrix}$$

a) Find the stationary/invariant distribution of this chain.

b) If the initial (start) distribution $\phi(0)$ is $(0.8, 0.2)$, what is the probability of this chain visiting state 1 at time $n = 3$?

34. A Markov chain has the 4-state space $\mathcal{S} = \{(0,0), (0,1), (1,0), (1,1)\}$ and the transition matrix

$$\mathbf{P} = \begin{array}{c|cccc} & (0,0) & (0,1) & (1,0) & (1,1) \\ (0,0) & (1-\alpha)b_{00} & (1-\alpha)b_{01} & \alpha b_{10} & \alpha b_{11} \\ (0,1) & (1-\alpha)b_{00} & (1-\alpha)b_{01} & \alpha b_{10} & \alpha b_{11} \\ (1,0) & \beta b_{00} & \beta b_{01} & (1-\beta)b_{10} & (1-\beta)b_{11} \\ (1,1) & \beta b_{00} & \beta b_{01} & (1-\beta)b_{10} & (1-\beta)b_{11} \end{array} .$$

The state of the chain is the value of a two-dimensional random variable (X_n, Y_n), such that, e.g.,

$$(1-\alpha)b_{01} = P\left(X_n = 0, Y_n = 1 \mid X_{n-1} = 0, Y_{n-1} = 0\right),$$

and analogously for the arrays in the matrix $\mathbf{P}$. Both X_n and Y_n are binary variables. We require, of course, that

$$0 \leq \alpha \leq 1, 0 \leq \beta \leq 1, b_{00} \geq 0, b_{01} \geq 0, b_{10} \geq 0, b_{11} \geq 0$$

and

$$b_{00} + b_{01} = 1, b_{10} + b_{11} = 1,$$

Note that the form of the matrix $\mathbf{P}$ implies that

$$P\left(X_n, Y_n \mid X_{n-1}, Y_{n-1}\right) = P\left(X_n, Y_n \mid X_{n-1}\right) \tag{19.11}$$

(Why ?) Check then that

$$P\left(X_n = 1 \mid X_{n-1} = 0\right) = \alpha, P\left(X_n = 0 \mid X_{n-1} = 1\right) = \beta$$

35. **A converse to the preceding**
Assume that (X_n) is a Markov chain with the transition matrix

$$A = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}$$

and the state space $\{0, 1\}$. Assume that (Y_n) is random process with values in $\{0, 1\}$, and with the (emission) probability matrix

$$B = \begin{pmatrix} b_{00} & b_{01} \\ b_{10} & b_{11} \end{pmatrix},$$

where $b_{ij} = P\left(Y_n = j | X_n = i\right)$.

Let the entries in A and B be combined as in $\mathbf{P}$ in exercise 34. above.

Assume that $(X_n, Y_n)_{n\geq 0}$ is an HMM with parameter $\lambda = (A, B)$.

Show that $(X_n, Y_n)_{n\geq 0}$ is a Markov chain on the state space space $\mathcal{S} = \{(0,0), (0,1), (1,0), (1,1)\}$ with the transition matrix $\mathbf{P}$.

Comment: The statements here hold for all HMMs with finite state spaces. In other words, if we have an HMM, we can always regard it as a two-component MC.

36. **Global Similarity Alignment**

The global similarity alignment is the optimal alignment of $\mathbf{x}$ with $\mathbf{y}$ corresponding to

$$S\left(\mathbf{x}, \mathbf{y}\right) = \max_{\mathbf{x}^*, \mathbf{y}^*} \overline{s}\left(\mathbf{x}^*, \mathbf{y}^*\right),$$

where the maximum is computed over the set of all alignments of $\mathbf{x}$ with $\mathbf{y}$ and where

$$\overline{s}\left(\mathbf{x}^*, \mathbf{y}^*\right) = \sum_{i=1}^{\infty} s\left(x_i^*, y_i^*\right).$$

An example of a similarity function is

$$s(x, y) = \begin{cases} +1 & \text{if } x = y \\ -1 & x \neq y. \end{cases}$$

$$\delta = s(x, -) = s(-, x) > 0,$$

The recursion for global alignment is

$$S_{i,j} = \max\{S_{i-1,j} - \delta, S_{i,j-1} - \delta, S_{i-1,j-1} + s(x_i, y_j)\}$$

with the notation for maximum similarity for prefixes and adding obvious start conditions.

Let $\mathbf{x} = CAGTATCGCA$, $\mathbf{y} = AAGTTAGCAG$ be two sequences. Show that the maximum similarity alignment with $\delta = 1$ is

$$\begin{matrix} \mathbf{x}^* \\ \mathbf{y}^* \end{matrix} = \begin{matrix} C\,A\,G\,T\,A\,T\,C\,G\,C\,A - \\ A\,A\,G\,T - T\,A\,G\,C\,A\,G \end{matrix}$$

You are expected to display the dynamic programming table.

37. Reconsider exercise 34. above.

    a) By loc.cit. we have

    $$P(X_n = i, Y_n = j \mid X_{n-1} = x, Y_{n-1} = y) = b_{ij} P(X_n = i \mid X_{n-1} = x),$$

    where $i \in \{0,1\}$, $j \in \{0,1\}$, $y \in \{0,1\}$, $x \in \{0,1\}$, and $Y_{n-1} = y$ needs not to be included necessarily. Check that

    $$b_{ij} = P(Y_n = j \mid X_n = i).$$

    (*Hint*: It can be a smart idea to first show that

    $$P(Y_n = j \mid X_n = i) = \frac{\sum_y \sum_x P(Y_n = j, X_n = i, Y_{n-1} = y, X_{n-1} = x)}{P(X_n = i)}.)$$

    b) Show now that (X_n, Y_n) is an HMM. (What remains to be shown is that $Y_0, Y_1, \ldots, Y_n$ are conditionally independent r.v.:s given $X_0, X_1, \ldots, X_n$.)

38. Define an HMM $\lambda = (A, B, \pi)$ with the following parameters.

    The state space of the hidden chain is $\mathcal{S} = \{S_1, S_2, S_3\}$, and

    $$A = \begin{pmatrix} 0 & 1/2 & 1/2 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix},$$

    $$\pi(0) = (1, 0, 0),$$

    The emission alphabet is $\mathcal{O} = \{1, 2, 3\}$, and

    $$B = \begin{pmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 1/2 \\ 0 & 1/2 & 1/2 \end{pmatrix},$$

    What are the possible state sequences for the following observed sequences $\mathbf{o}$, and what is the corresponding $p(\mathbf{o}|\lambda)$?

a) $\mathbf{o} = 123$.

b) $\mathbf{o} = 131$.

38. **HMM as a multiple alignment**
Consider the three amino acid sequences

$$WRCCTGC, WCCGGCC, WCGCC$$

Suppose that their respective paths through a protein HMM of length 8 are

$$\begin{array}{cccccccccc} m_0 & m_1 & i_1 & m_2 & m_3 & m_4 & m_5 & d_6 & m_7 & m_8 \\ m_0 & m_1 & m_2 & m_3 & m_4 & m_5 & m_6 & m_7 & m_8 & \\ m_0 & m_1 & m_2 & d_3 & d_4 & m_5 & m_6 & m_7 & m_8 & \end{array} .$$

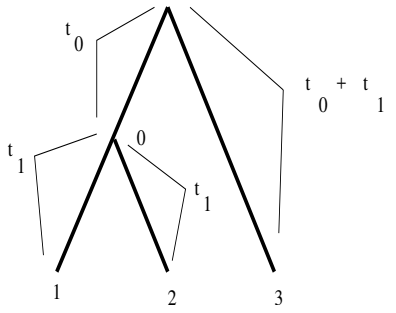
Using the theory in Ewens and Grant to give the alignment of the sequences that these paths determine.

Uppgift 39. We have a Markov chain in continuous time applied as a model for substitution of nucleotides in DNA. We assume independent sites and the same distribution over all sites. The state space S is reduced to two states $S = \{+1, -1\}$, e.g., pyriminide $\mapsto +1$ och purine $\mapsto -1$. The transition probability matrix is

$$\mathbf{P}(t) = \begin{pmatrix} \frac{1}{2}\left(1 + e^{-2\lambda t}\right) & \frac{1}{2}\left(1 - e^{-2\lambda t}\right) \\ \frac{1}{2}\left(1 - e^{-2\lambda t}\right) & \frac{1}{2}\left(1 + e^{-2\lambda t}\right) \end{pmatrix}.$$

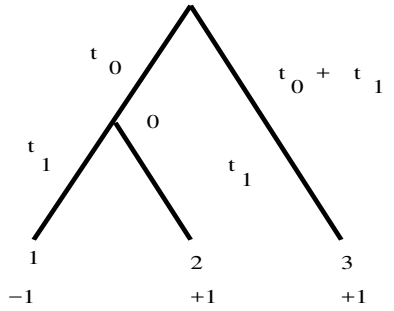
(a) Verify that $\pi = (1/2, 1/2)$ is the equilibrium distribution for this chain and check that the chain is reversible. (Note! You need not necessarily have to derive $\mathbf{Q}$ for this.)

(b) Consider the tree in the Figure. Leaves have been marked by $1, 2, 3$, and the only internal mode is 0.



The expected number of substitutions $= \lambda t$ are weights on the edges and we simplify by $\lambda = 1$. Therefore $t_0 + t_1$ is the weight on the path from the root to node number 3 and analogusly for the other paths.

Compute the probability $p(U)$ fo the sitepattern ' $U = (-1, +1, +1)$ in the Figure by the Markov assumption on the tree, and the model of substitutions above.



Suggestion for solution: $p(U) = \frac{1}{2}\left(\frac{1}{4} - \frac{1}{4}e^{-4t_1}\right)$

40. Jerzy Neyman[5] was perhaps the first statistician to consider molecular. Neyman considered a substitution process with $\mid \mathcal{X} \mid = s$ different states (molecules). Here we express Neyman's model with the matrix semigroup

$$\mathbf{P}_{\mathrm{Ne}}(t) = \sum_{k=0}^{\infty} \mathbf{P}^k P\left(N(t) = k\right) = e^{-\lambda t \mathbf{I}} e^{\lambda t \mathbf{P}}, \tag{19.12}$$

where $N = \{N(t) | t \geq 0\}$ is again a Poisson process, molecular clock, with the intensity $\lambda \geq 1$, $\mathbf{I}$ is the $s \times s$ identity matrix, and the $s \times s$ transition probability matrix $\mathbf{P}$ of the discrete time Markov chain is

$$\mathbf{P} = \begin{pmatrix} 0 & \frac{1}{s-1} & \frac{1}{s-1} & \cdots & \frac{1}{s-1} \\ \frac{1}{s-1} & 0 & \frac{1}{s-1} & \cdots & \frac{1}{s-1} \\ \frac{1}{s-1} & \frac{1}{s-1} & 0 & \cdots & \frac{1}{s-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{s-1} & \frac{1}{s-1} & \frac{1}{s-1} & \cdots & 0 \end{pmatrix}. \tag{19.13}$$

The generator is thus for any $\lambda \geq 1$

$$\mathbf{Q} = \lambda\left(\mathbf{P} - \mathbf{I}\right). \tag{19.14}$$

In Neyman's model the current molecule is substituted with a different molecule chosen from the uniform distribution $\left(\frac{1}{s-1}, \frac{1}{s-1}, \ldots, \frac{1}{s-1}\right)$ at the ticks of the molecular clock $N = \{N(t) | t \geq 0\}$.

Check that $\mathbf{P}_{\mathrm{Ne}}(t)$ is elementwise given by

$$P_{ii}(t) = \frac{1}{s}\left(1 + (s-1)e^{-\frac{s}{s-1}\lambda t}\right), \tag{19.15}$$

and

$$P_{ij}(t) = \frac{1}{s}\left(1 - e^{-\frac{s}{s-1}\lambda t}\right). \tag{19.16}$$

41. Let $Y \in \mathrm{Be}(1/2)$ and $N = \{N(t) | t \geq 0\}$ be a Poisson process with intensity λ, and let Y be independent of N. Set

$$X(t) = (-1)^{Y+N(t)}.$$

[5]Polish-American statistician, 1894−1981,
http://www-groups.dcs.st-and.ac.uk/∼history/Mathematicians/Neyman.html

Then $X = \{X(t) \mid t \geq 0\}$ is a continuous-time Markov chain. The transition matrix is

$$\mathbf{P}(t) = \begin{pmatrix} \frac{1}{2}\left(1 + e^{-2\lambda t}\right) & \frac{1}{2}\left(1 - e^{-2\lambda t}\right) \\ \frac{1}{2}\left(1 - e^{-2\lambda t}\right) & \frac{1}{2}\left(1 + e^{-2\lambda t}\right) \end{pmatrix}.$$

The sample paths of X, the random telegraph, are sequences of -1 and 1, each digit prevailing a random (exponentially distributed time).

a) Find the generator $\mathbf{Q}$ of the process X.

b) Find the equilibrium distribution of the process X.

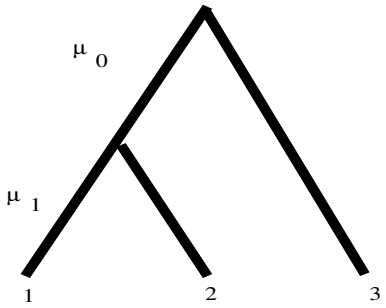
c) Is the process X reversible ? Justify your answer.

1. Find a transition probability matrix P^* for a Markov chain X^* such that

$$X(t) = X^*_{N(t)}$$

where $N(t)$ is a suitable Poisson process.

42. Consider the tree in the Figure



Compute the probability of the site pattern TCA using the Jukes-Cantor model for substitution along the branches.

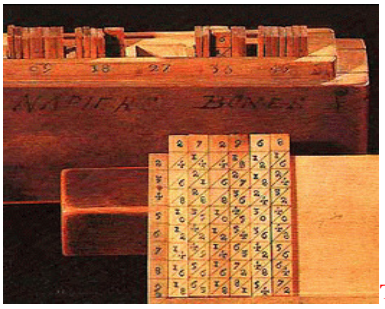
We assume that the branch lengths are such that $\mu_0 = \mu \cdot t_0$, $\mu_1 = \mu \cdot t_1$, ($\mu = \alpha/4$), and that the remaining branch length is $= \mu_0 + \mu_1$.

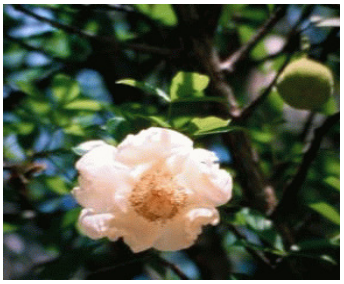
Answer:

$$\frac{1}{16}\left(6 - 6b^2 - 12a^2b^2 + 12a^2b^3\right),$$

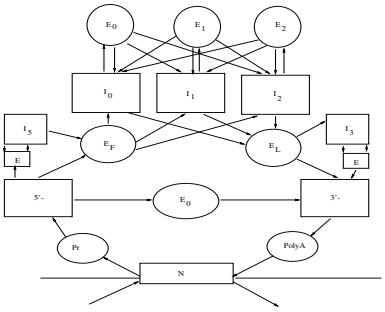
where

$$a = e^{-\frac{4}{3}\mu_0}, b = e^{-\frac{4}{3}\mu_1}.$$





The flower of monkey–bread tree



20 References and Further Reading:

- More extensive accounts of relevant probability calculus at an intermediate level are found in

  G. Blom, L. Holst and D. Sandell (1994): *Problems and Snapshots from the World of Probability*. Springer Verlag.

  A. Gut (2009): *An Intermediate Course in Probability. Second Edition*. Springer Verlag.

  L.L. Helms (1997): *Introduction to Probability Theory With Contemporary Applications*. W.H. Freeman and Company. New York.

  K. Lange (2003): *Mathematical and statistical methods for genetic analysis*, Springer Science & Business Media.

  K. Lange (2010): *Applied probability*, Springer Science & Business Media.

- Treatments of discrete mathematics useful in this course are

  N.L. Briggs (1993): *Discrete Mathematics. Revised Edition*, Oxford Science Publications, Clarendon Press, Oxford.

  R.L. Graham, D.E. Knuth and O. Patashnik (1989): *Concrete Mathematics*. Addison Wesley Publishing Company, Reading Mass., Menlo Park Californina, etc. .

- A reference for calculus is

  A.I. Khuri (1993): *Advanced Calculus with Applications in Statistics*. John Wiley and Sons, Inc. New York.

- Bioinformatics and computational biology

  M. Axelson-Fisk (2010): *Comparative Gene Finding*. Springer, London.

  M. Borodovsky and S. Ekisheva (2006): *Problems and solutions in biological sequence analysis*. Cambridge University Press.

  D. Gusfield (1997): *Algorithms on Strings, Trees and Sequences. Computer Science and Computational Biology*. Cambridge University Press. Cambridge.

  M.S. Waterman (1995): *Introduction to Computational Biology. Maps, sequences and genomes.*

  R.C. Deonier, S. Tavaré and M.S. Waterman (2005): *Computational Genome Analysis*. Springer. Melbourne, Madras.

- Conditional independence

  A.P. Dawid (1980): Conditional independence for statistical operations. *Annals of Statistics*, 8, pp. 598 - 617.

- Poisson approximation

R.J. Serfling (1978): Some elementary results on Poisson approximation in a sequence of Bernoulli trials. *SIAM Review*, 20, pp. 567 - 579.

- Fragment accuracy

G.A. Churchill and M.S. Waterman (1992): The Accuracy of DNA Sequences: Estimating Sequence Quality. *Genomics*, 14, pp. 89 - 98.

- The distribution of the number of fragments

T. Bishop, J.A. Williamson and M.H. Skolnik (1983): A Model for Restriction Fragment Length Distributions. *American Journal of Human Genetics*, 35, pp.795 - 815.

J.F. Gusella (1986): DNA polymorphism and human disease. *Annual Reviews of Biochemistry*, 55, pp. 831-854.

L. Holst (1980): On the Lengths of the Pieces of a Stick Broken at Random. *Journal of Applied Probability*, 17, pp. 623 - 634.