Exercise sheet 2 in DNA sequence alignments, phylogenetic trees and PCMs

Krzysztof Bartoszek
732A51 Bioinformatics
Department of Computer and Information Science, Linköping University

12 XII 2024 (R35)

The exercises here are taken from the below textbooks

BE M. Borodovsky, S. Ekisheva., 2006, Problems and Solutions in Biological Sequence Analysis, Cambridge University Press.

EG W. J. Ewens, G. R. Grant., 2005, Statistical Methods in Bioinformatics, 2nd ed. Springer.

F F. C. Klebaner., 2005 Introduction to Stochastic Calculus with Applications, Imperial College Press

L A. M. Lesk., 2014, Introduction to Bioinformatics, Oxford University Press.

MM J. Momand, A. McCurdy., 2017, Concepts in Bioinformatics and Genomics, Oxford University Press.

Sequence alignments

Exercise 1 (L: Exercise 5.1)
What is the Hamming distance between the words $DECLENSION$ and $RECREATION$.

Exercise 2 (L: Exercise 5.3)
The Levenshtein distance between the strings $AGTCC$ and $CGCTCA$ is 3, consistent with the following alignment

$$\begin{array}{cccccc} A & G & - & T & C & C \\ C & G & C & T & C & A \end{array}$$

Provide a sequence of three edit operations that convert $AGTCC$ to $CGCTCA$.

Exercise 3 (BE: Problem 2.9)
Calculate the Dynamic programming matrix and an optimal global alignment for the DNA sequences $GAATTC$ and $GATTA$, scoring $+2$ for a match, 1 for a mismatch, and with a linear gap penalty of $d = 2$.

Exercise 4 (MM: Problem 5.5)
Use the Needleman–Wunsch algorithm to align the DNA sequence pairs ($ATAGC$; $ATATGA$) and ($ATATGC$; $ATATGA$), scoring +1 for a match, 0 for a mismatch and 0 for a gap.

Exercise 5 (BE: Problem 2.10)
Calculate the score of the example alignment

$$\begin{array}{cccccccc} V & L & S & P & A & D & - & K \\ H & L & - & - & A & E & S & K \end{array}$$

using the BLOSUM50 matrix gap opening penalty $d = 12$ and gap extension penalty $e = 2$. You will find the matrix [BE] at the end of this file.

Exercise 6 (BE: Problem 2.18)
The statistical test is applied to establish relatedness of locally aligned protein sequences X and Y of lengths $n = 100$ and $m = 300$. It is assumed that $K = 0.1$, $\lambda = 0.7$.
Define the cut–off value S^α corresponding to the significance level α (the false negative rate) of the test equal to 0.05.
Given observed scores $S_1 = 15$, $S_2 = S_3 = 12$, $S_4 = 11$, $S_5 = 10$ of the highest–scoring segment pairs, use the tests with the cut-off value defined as previously to test if the two protein sequences are random with respect to each other.

In the lecture we provided the test statistics and p–value formula when comparing a sequence against a database. When comparing two sequence against each other, we will use

$$E = Knme^{-\lambda S}, \qquad \text{p} - \text{value} \approx 1 - e^{-E}.$$

Again here E approximates the expected number of high–scoring segment pairs (HSPs).

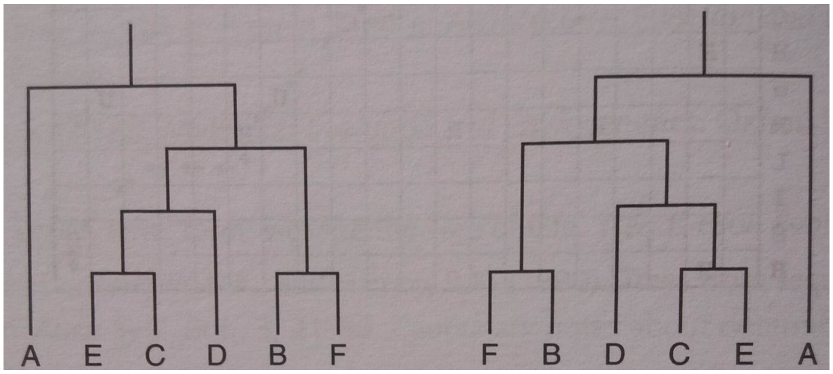
Exercise 7 (EG: Problem 6.1)
Explain why

$$\sum_{k=0}^{\min\{m,n\}} \binom{m}{k}\binom{n}{k} = \binom{m+n}{n} = \binom{m+n}{m}$$

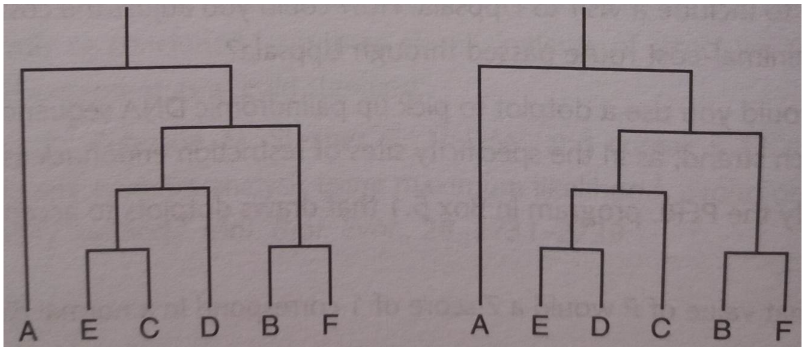
Think of having to select n objects from a set with $s = m + n$ objects, of which m are of one kind and n are of another kind.

Phylogenetic trees

Exercise 8 (L: Exercise 5.20)
(a) Is the following pair of trees identical in topology?



(b) Is the following pair of trees identical in topology?



Exercise 9 (L: Exercise 5.14)
A convenient notation for trees uses nested parentheses to indicate the clusters.
(a) Expand the following into a rooted tree $((A(BC))D)$.
(b) Write the parenthesis notation for the trees of shown in Exercise 8.

Exercise 10 (MM: Problem 8.7)
Construct an UPGMA tree for the following distance matrix. Show values of branch lengths on the tree.

	A	B	C	D
A	0	9	7	5
B	9	0	8	10
C	7	8	0	8
D	5	10	8	0

Exercise 11 (BE: Problem 8.1)
Show that the Jukes-Cantor nucleotide substitution matrix introduced by Jukes and Cantor (1969) is multiplicative: $\mathbf{P}_{JC}(t)\mathbf{P}_{JC}(s) = \mathbf{P}_{JC}(t+s)$ for all values of s and t.

Stochastic calculus and PCMs

Exercise 12
Assume $s < t$. Calculate $\mathrm{Cov}\left[X(s), X(t)\right]$ for X an
(a) Brownian motion,
(b) Ornstein–Uhlenbeck process.

Exercise 13 (K: Exercise 3.4)
Assume that $B(t)$ is a Brownian motion. Show that
(a) $X(t) = -B(t)$
(b) $X(t) = cB(t/c^2) \quad c \neq 0$
are also Brownian motions.

Exercise 14 (K: Exercise 3.5)
Let $B(t)$ and $W(t)$ be two independent Brownian motions. Show that $X(t) = (B(t)+W(t))/\sqrt{2}$ is also a Brownian motion. Find the correlation between $B(t)$ and $W(t)$.

Exercise 15 (K: Exercise 3.8)
Let $B(t)$ be a Brownian motion and $0 \leq s < t$. Show that the conditional distribution of $B(s)$ given $B(t) = b$ is normal and give its mean and variance.

Exercise 16
Let the process $X(t)$ be defined by the SDE

$$\mathrm{d}X(t) = \mu(t, X(t))\mathrm{d}t + \sigma(t, X(t))\mathrm{d}B(t)$$

and consider a new process $Y(t) = f(X(t))$, where f is twice differentiable. Then, the Itô formula says

$$\mathrm{d}Y(t) = (f'(X(t))\mu(t, X(t)) + \frac{1}{2}f''(X(t))\sigma^2(t, X(t))\mathrm{d}t + f'(X(t))\sigma(t, X(t))\mathrm{d}B(t).$$

Itô's formula is the stochastic calculus version of the usual change of variables formula.

Let now $X(t)$ be an Ornstein–Uhlenbeck process. Find the SDE representation of $X^2(t)$.

Exercise 17
Find the most parsimonious internal node labels for the first tree of Exercise 8. Assume that at the tips the labels are
(a) $A:2$, $E:2$. $C:1$, $D:1$, $E:1$, $F:2$
(b) $A:1$, $E:1$. $C:1$, $D:1$, $E:2$, $F:2$

Exercise 18
Prove the formula for the covariance between traits measured at two tips, $\mathrm{Cov}\left[X_1, X_2\right]$, under the Ornstein–Uhlenbeck model of evolution.

Table 2.1. *The BLOSUM50 substitution matrix*

The log-odds values are scaled and rounded to the nearest integer.

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	**5**	−2	−1	−2	−1	−1	−1	0	−2	−1	−2	−1	−1	−3	−1	1	0	−3	−2	0
R	−2	**7**	−1	−2	−4	1	0	−3	0	−4	−3	3	−2	−3	−3	−1	−1	−3	−1	−3
N	−1	−1	**7**	2	−2	0	0	0	1	−3	−4	0	−2	−4	−2	1	0	−4	−2	−3
D	−2	−2	2	**8**	−4	0	2	−1	−1	−4	−4	−1	−4	−5	−1	0	−1	−5	−3	−4
C	−1	−4	−2	−4	**13**	−3	−3	−3	−3	−2	−2	−3	−2	−2	−4	−1	−1	−5	−3	−1
Q	−1	1	0	0	−3	**7**	2	−2	1	−3	−2	2	0	−4	−1	0	−1	−1	−1	−3
E	−1	0	0	2	−3	2	**6**	−3	0	−4	−3	1	−2	−3	−1	−1	−1	−3	−2	−3
G	0	−3	0	−1	−3	−2	−3	**8**	−2	−4	−4	−2	−3	−4	−2	0	−2	−3	−3	−4
H	−2	0	1	−1	−3	1	0	−2	**10**	−4	−3	0	−1	−1	−2	−1	−2	−3	2	−4
I	−1	−4	−3	−4	−2	−3	−4	−4	−4	**5**	2	−3	2	0	−3	−3	−1	−3	−1	4
L	−2	−3	−4	−4	−2	−2	−3	−4	−3	2	**5**	−3	3	1	−4	−3	−1	−2	−1	1
K	−1	3	0	−1	−3	2	1	−2	0	−3	−3	**6**	−2	−4	−1	0	−1	−3	−2	−3
M	−1	−2	−2	−4	−2	0	−2	−3	−1	2	3	−2	**7**	0	−3	−2	−1	−1	0	1
F	−3	−3	−4	−5	−2	−4	−3	−4	−1	0	1	−4	0	**8**	−4	−3	−2	1	4	−1
P	−1	−3	−2	−1	−4	−1	−1	−2	−2	−3	−4	−1	−3	−4	**10**	−1	−1	−4	−3	−3
S	1	−1	1	0	−1	0	−1	0	−1	−3	−3	0	−2	−3	−1	**5**	2	−4	−2	−2
T	0	−1	0	−1	−1	−1	−1	−2	−2	−1	−1	−1	−1	−2	−1	2	**5**	−3	−2	0
W	−3	−3	−4	−5	−5	−1	−3	−3	−3	−3	−2	−3	−1	1	−4	−4	−3	**15**	2	−3
Y	−2	−1	−2	−3	−3	−1	−2	−3	2	−1	−1	−2	0	4	−3	−2	−2	2	**8**	−1
V	0	−3	−3	−4	−1	−3	−3	−4	−4	4	1	−3	1	−1	−3	−2	0	−3	−1	**5**