Exercises – Phylogenetics

Universität Bielefeld, WS 2017/2018,
Dr. Roland Wittler, Prof. Dr. Jens Stoye, M. Sc. Tizian Schulz
https://gi.cebitec.uni-bielefeld.de/Teaching/2017winter/Phylogenetik

Exercise Sheet 10 — 14.12.2017

Due: 21.12.2017

Task 1 Revision Probability Theory. (5 points)
Let (Ω, Pr) be the probability space of a DNA base triplet with sample space $\Omega = \{\texttt{AAA}, \texttt{AAC}, \ldots, \texttt{TTT}\}$ and uniform distribution Pr. Further, let $(\Omega_2, \mathrm{Pr}_2) = (\Omega, \mathrm{Pr}) \times (\Omega, \mathrm{Pr})$ be the probability space of two independent consecutive base triplets. We call them *triplet 1* and *triplet 2*, and denote the elementary events as (e_1, e_2).
Write down the probabilities to observe the following events in $(\Omega_2, \mathrm{Pr}_2)$:

(a) There is a `U` in triplet 1.

(b) The total number of `U`s in both triplets is exactly 1.

(c) Triplet 1 codes for glutamine.

(d) Triplet 1 codes for glutamine, and the last two bases of triplet 1 together with the first base of triplet 2 form a codon that codes for lysine.

(e) Both triplets are the same.

Task 2 Programming exercise: Splitstree. (4 points)
Implement a function that calculates the *isolation index* for the *split decomposition*: For two given sets J and K, calculate $\alpha_{J,K}(d)$ with respect to a matrix d:

$$\alpha_{J,K}(d) = \frac{1}{2} \min_{\substack{i,j \in J \\ k,l \in K}} \left(\max\{d_{ij} + d_{kl}, d_{ik} + d_{jl}, d_{il} + d_{jk}\} - d_{ij} - d_{kl}\right)$$

For **4 bonus points** implement also

(a) a function that calculates the *split metric*: (1* point)

$$\delta_{J,K}(i,j) = \begin{cases} 0 & \text{if } i \text{ and } j \text{ both are in } J \text{ or both are in } K \\ 1 & \text{otherwise} \end{cases}$$

(b) a function that calculates the matrix d^1: $d^1(i,j) = \sum_{\text{splits } J,K} \alpha_{J,K}\delta_{J,K}(i,j)$. (2* points)
You do not need to calculate which splits are really valid ($\alpha_{J,K} > 0$). Just sum over all possible splits. Even though this procedure is not efficient, it leads to the correct solution since $\alpha_{J,K} = 0$ for all splits that are invalid.
Hint: For the representation/enumeration of subsets, numbers in binary format can be used. For example the set $\{B, D, E\}$ can be represented by `01011`.

(c) a function that calculates the *splittable percentage* ρ: (1* point)

$$\rho := \left(\sum_{\text{taxa } i,j} d^1_{ij} \Bigg/ \sum_{\text{taxa } i,j} d_{ij} \right) \cdot 100\%$$

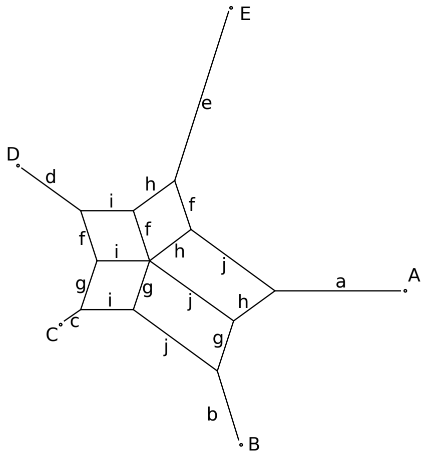
Send a version of your program to your TA[1] and describe how to use the program. Make it as easy as possible to calculate all the α and ρ in Task 3. You'll find another matrix and some solutions on the back of this sheet if you want to test your implementation.

Turn over! Bitte wenden!

[1] tischulz@cebitec.uni...

Task 3 Splitstree. **(4 points)**
Solve this exercise with your implementation. (You can also use the software *Splitstree* (`www.splitstree.org`) — describe the format of your input and your proceeding.)
Consider the following distance matrix and the corresponding splitstree:

$$\begin{array}{r|ccccc} & A & B & C & D & E \\ A: & 0 & 6 & 8 & 9 & 9 \\ B: & & 0 & 5 & 8 & 10 \\ C: & & & 0 & 4 & 8 \\ D: & & & & 0 & 7 \\ E: & & & & & 0 \end{array}$$



(a) Calculate the length of the edges a to j of the given network.
(Hint: $a = \alpha_{\{A\},\{B,C,D,E\}}$, $g = \alpha_{\{A,E\},\{B,C,D\}}$, etc.)

(b) If one uses the *Neighbor Joining* algorithm on the distance matrix, the result is the tree

$$\left(\left(\left(\mathrm{C} : \frac{7}{4}, \mathrm{D} : \frac{9}{4}\right) : \frac{3}{4}, \mathrm{E} : \frac{19}{4}\right) : \frac{7}{4}, \left(\mathrm{A} : \frac{14}{4}, \mathrm{B} : \frac{10}{4}\right)\right);.$$

Compare this tree with the result from Task (a). Hint: Even if you have not solved Task (a), you can nevertheless compare both trees since the edge lengths in the image of the splitstree are scaled to the real edge lengths.

Matrix and some solution to test your implementation for Task 2 (optional!):

$$\begin{array}{r|ccccc} & A & B & C & D & E \\ A: & 0 & 6 & 8 & 5 & 10 \\ B: & & 0 & 5 & 8 & 10 \\ C: & & & 0 & 4 & 8 \\ D: & & & & 0 & 7 \\ E: & & & & & 0 \end{array}$$

J	K	$\alpha_{J,K}$
$\{A,B,C,E\}$	$\{D\}$	0.5
$\{A,B,D,E\}$	$\{C\}$	0.5
$\{A,B,E\}$	$\{C,D\}$	0.5
$\{A,C,D,E\}$	$\{B\}$	1.5
$\{A,C,E\}$	$\{B,D\}$	0.0
$\{A,D,E\}$	$\{B,C\}$	1.5
$\{A,E\}$	$\{B,C,D\}$	0.0
$\{B,C,D,E\}$	$\{A\}$	1.5
$\{B,C,E\}$	$\{A,D\}$	1.0
$\{B,D,E\}$	$\{A,C\}$	0.0
$\{B,E\}$	$\{A,C,D\}$	0.0
$\{C,D,E\}$	$\{A,B\}$	2.0
$\{C,E\}$	$\{A,B,D\}$	0.0
$\{D,E\}$	$\{A,B,C\}$	0.0
$\{E\}$	$\{A,B,C,D\}$	4.5

$$\sum_{\text{taxa } i,j} d^1_{ij} = 128 \text{ (or 256 if you sum over both } (i,j) \text{ and } (j,i).$$

$$\sum_{\text{taxa } i,j} d_{ij} = 142 \text{ (or 284, respectively.)}$$

$$\rho \approx 90,141\%$$