

Exercises – Phylogenetics

Universität Bielefeld, SS 2019

Dr. Roland Wittler, M. Sc. Tizian Schulz

<https://gi.cebitec.uni-bielefeld.de/Teaching/2019summer/Phylogenetik>

Exercise Sheet 8 — 06.06.2019

Due: 13.06.2019

Task 1 Revision Probability Theory.

(5 points)

Let (Ω, Pr) be the probability space of a RNA base triplet with sample space $\Omega = \{\text{AAA}, \text{AAC}, \dots, \text{UUU}\}$ and uniform distribution Pr . Further, let $(\Omega_2, \text{Pr}_2) = (\Omega, \text{Pr}) \times (\Omega, \text{Pr})$ 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 (Ω_2, Pr_2) :

- There is an A in triplet 1.
- The total number of As in both triplets is exactly 3.
- Triplet 1 forms a stop codon.
- Triplet 1 forms a stop codon, and the last two bases of triplet 1 together with the first base of triplet 2 form a codon that codes for lysine.
- Both triplets are the same.

Task 2 Programming exercise: Split decomposition.

(4 points)

Implement a function that calculates the *isolation index* for the *split decomposition*: For a given distance matrix d on an object set O , calculate $\alpha_{J,K}$ for a split J,K of O , i.e., $J \cap K = \{\}$ and $J \cup K = O$:

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

For 4 bonus points implement also

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

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

- a function that calculates the matrix d^1 : $d_{ij}^1 = \sum_{\text{splits } J,K} \alpha_{J,K} \delta_{J,K}(i, j)$. (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.

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

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

Send your program by email and describe how to use the program. Make it as easy as possible to calculate all the α and ρ in Task 3. A further matrix and some solutions are provided on the back of this sheet to test your implementation.

Please turn over! Bitte wenden!

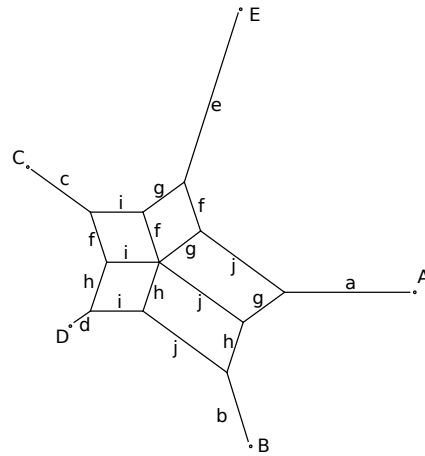
Task 3 Splitstree.

(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:

	A	B	C	D	E
A :	0	9	13	12	13
B :		0	12	7	15
C :			0	6	10
D :				0	12
E :					0



- (a) Calculate the length of the edges *a* to *j* of the given network.

(Hint: $a = \alpha_{\{A\},\{B,C,D,E\}}$, $g = \alpha_{\{A,E\},\{C,D,B\}}$, etc.)

- (b) The *Neighbor Joining* algorithm on the same distance matrix gives the tree

$$\left(\left(\left(E : \frac{55}{8}, C : \frac{25}{8} \right) : \frac{11}{8}, D : \frac{21}{8} \right) : \frac{19}{8}, \left(A : \frac{31}{6}, B : \frac{23}{6} \right) \right);$$

Compare this tree with the result of 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 solutions to test your implementation for Task 2 (optional!):

	A	B	C	D	E		J	K	$\alpha_{J,K}$
A :	0	6	8	5	10		{A, B, C, E}	{D}	0.5
B :		0	5	8	10		{A, B, D, E}	{C}	0.5
C :			0	4	8		{A, B, E}	{C, D}	0.5
D :				0	7		{A, C, D, E}	{B}	1.5
E :					0		{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

$$\sum_{\text{taxa } i,j} d_{ij}^1 = 64 \text{ (or 128 if you sum over both } (i,j) \text{ and } (j,i)).$$

$$\sum_{\text{taxa } i,j} d_{ij} = 71 \text{ (or 142, respectively.)}$$

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