Übungsblatt 3

Phylogenetik

Aufgabe 3.1

Teilaufgabe 3.1.1

A	GATAG
B	GATAA
C	ATGGA
D	GATAG

a) p-Distanz

$$\boldsymbol{p = \frac{D}{L}}$$

$$p(AB) = \frac{1}{5} = 0.2$$

$$p(AC) = \frac{5}{5} = 1$$

$$p(AD) = 0$$

$$p(BC) = \frac{4}{5} = 0.8$$

$$p(BD) = \frac{1}{5} = 0.2$$

$$p(CD) = \frac{5}{5} = 1$$

Distanzmatrix

	A	**B**	**C**	**D**
A	-	0.2	1	0
B	0.2	-	0.8	0.2
C	1	0.8	-	1

b) Poisson-korrigierte Distanz

$$d_p = -\ln(1-p)$$

$d_{p(AB)} = -\ln(1-0.2) = 0.223$

$d_{p(AC)} = -\ln(1-1)$ → keine Poisson-Korrektur, da die Sequenzen A und C keine Überreinstimmungen aufweisen

$d_{p(AD)} = -\ln(1-0) = 0$

$d_{p(BC)} = -\ln(1-0.8) = 1.6$

$d_{p(BD)} = -\ln(1-0.2) = 0.223$

$d_{p(CD)} = -\ln(1-1)$ → keine Poisson-Korrektur, da die Sequenzen A und C keine Überreinstimmungen aufweisen

Die Poisson-Korrektur gibt größere Distanzen als die „einfache“ p-Distanz. Die p-Distanz unterschätzt die wahre evolutionäre Distanz, da sie die wiederholte Mutation an der gleichen Position nicht berücksichtigt.

c) Jukes–Cantor Distanz

$$d_{JC} = -\frac{3}{4}\ln\left(1-\frac{4}{3}p\right) = 0$$

$d_{AB} = -\frac{3}{4}\ln\left(1-\frac{4}{3}0{,}2\right) = 0.236$
$d_{AC} = -\frac{3}{4}\ln\left(1-\frac{4}{3}1\right) =$ undefined*
$d_{AD} = -\frac{3}{4}\ln\left(1-\frac{4}{3}0\right) = 0$
$d_{BC} = -\frac{3}{4}\ln\left(1-\frac{4}{3}0.8\right) =$ undefined*
$d_{BD} = -\frac{3}{4}\ln\left(1-\frac{4}{3}0.2\right) = 0.236$
$d_{CD} = -\frac{3}{4}\ln\left(1-\frac{4}{3}1\right) =$ undefined*

* Jukes & Cantor Distanz nicht gültig für hohe *p*

Ähnlich wie die Poisson-Distanz gibt die Jukes & Cantor Distanz die Distanz zwischen den Sequenzen, die näher an der Realität sind als die „einfache“ p-Distanz.

Teilaufgabe 3.1.2

Prozentuale Übereinstimmung: $I = 100\,x\frac{M}{L}$

M = Anzahl der Übereinstimmungen

L = Länge der Sequenz

Nicht-prozentuale Übereinstimmung (sodass sie bei den weiteren Berechnungen mit der p-Distanz Sinn macht): $I^* = \frac{I}{100} = \frac{M}{L}$

p-Distanz:
$p = \frac{D}{L}$

$D =$ Anzahl der Nicht-Übereinstimmungen

L = Länge der Sequenz

Gegeben:

$$I^* = \frac{M}{L}$$

M + D = L

$$d_p = -\ln(1-p) = -\ln\left(1-\frac{D}{L}\right)$$

Zu beweisen, dass

$$1 - \frac{D}{L} = \frac{M}{L}$$
$$L/L - \frac{D}{L} = \frac{M}{L}$$

⇨ $d_p = -\ln(I^*)$

Aufgabe 3.2

Teilaufgabe 3.2.1

	A	**B**	**C**	**D**	**E**
A	0	5	9	9	8
B	5	0	10	10	9
C	9	10	0	8	7
D	9	10	8	0	3
E	8	9	7	3	0

-> es handelt sich um eine valide Distanzmatrix für evolutionäre Distanzen, da die Identität des Ununterscheidbaren und die Symmetrie stimmen

Teilaufgabe 3.2.2

	A	B	C	D	E
A	–	5	9	9	8
B	5	–	10	10	9
C	9	10	–	8	7
D	9	10	8	–	3
E	8	9	7	3	–

$$d_{xy} = \frac{1}{N_x N_y} \cdot \sum_{i \in x, j \in y} d_{ij}$$

kleinste Distanz → DE (3)

⇒ Cluster (DE) ≙ V

	A	B	C	V
A	–	5	9	8,5
B	5	–	10	9,5
C	9		–	7,5
V				–

$d_{AV} = \frac{1}{1.2}(d_{AD} + d_{AE}) = \frac{9+8}{2} = 8,5$

$d_{BV} = \frac{1}{1.2}(d_{BD} + d_{BE}) = \frac{10+9}{2} = 9,5$

$d_{CV} = \frac{1}{1.2}(d_{CD} + d_{CE}) = \frac{8+7}{2} = 7,5$

kleinste Distanz → AB (5)

⇒ Cluster (AB) ≙ W

	C	V	W
C	–	7,5	9,5
V		–	9
W			–

$d_{CW} = \frac{1}{1.2}(d_{CA} + d_{CB}) = \frac{9+10}{2} = 9,5$

$d_{VW} = \frac{1}{2.2}(d_{DA} + d_{DB} + d_{EA} + d_{EB}) = \frac{9+10+8+9}{4} = 9$

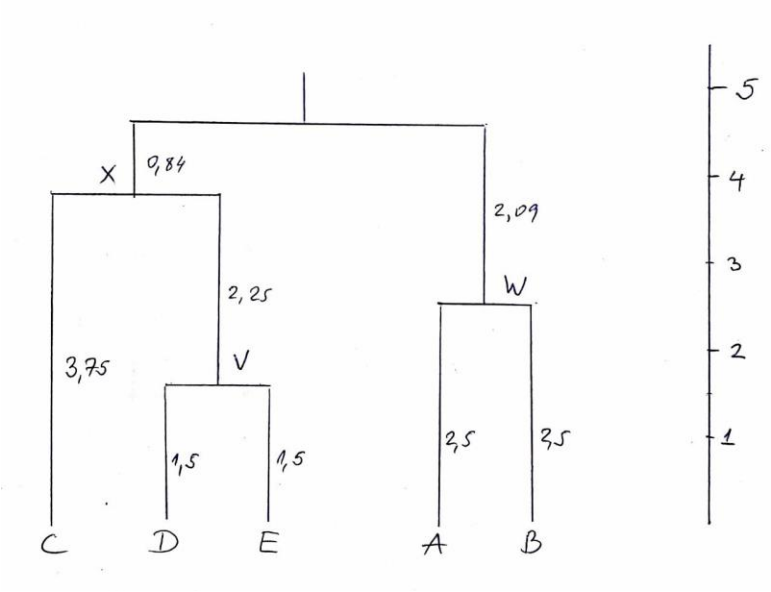
kleinste Distanz → CV (7,5)

⇒ Cluster (CV) ≙ X

	X	W
X	–	9,17
W		–

$d_{XW} = \frac{1}{3.2}(d_{CA} + d_{CB} + d_{DA} + d_{DB} + d_{EA} + d_{EB}) =$

$= \frac{9+10+9+10+8+9}{6} \approx 9,17$



Aufgabe 3.3

> install.packages("phangorn")

WARNING: Rtools is required to build R packages but is not currently installed. Please download and install the appropriate version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/

Installing package into ‘C:/Users/raliz/AppData/Local/R/win-library/4.2’

(as ‘lib’ is unspecified)

also installing the dependencies ‘ape’, ‘fastmatch’, ‘igraph’, ‘quadprog’, ‘Rcpp’

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.2/ape_5.6-2.zip'

Content type 'application/zip' length 3523153 bytes (3.4 MB)

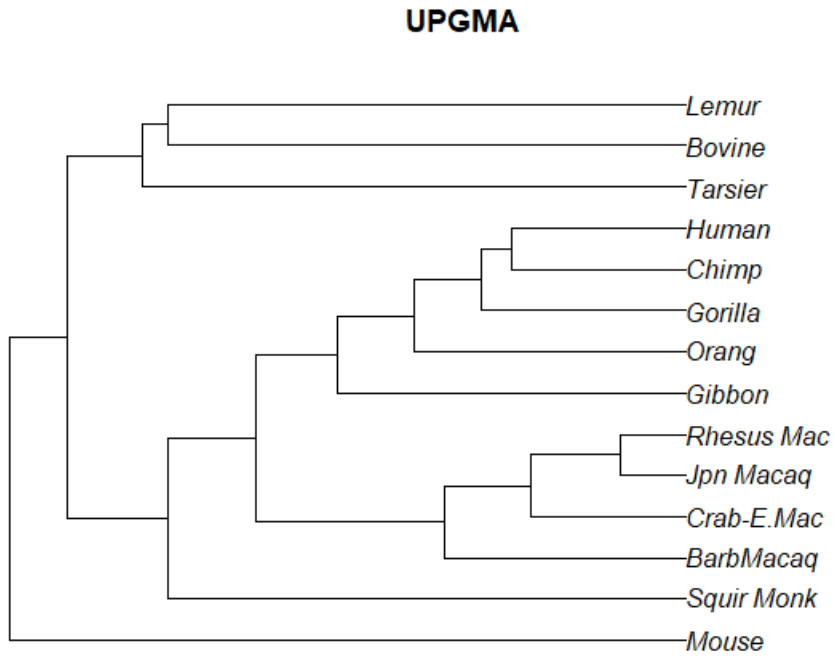
downloaded 3.4 MB

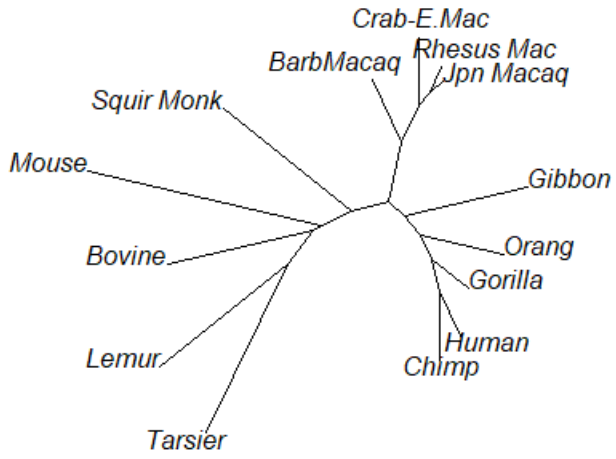
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.2/fastmatch_1.1-3.zip'

Content type 'application/zip' length 39902 bytes (38 KB)

```
downloaded 38 KB

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.2/igraph_1.3.1.zip'
Content type 'application/zip' length 5789827 bytes (5.5 MB)
downloaded 5.5 MB

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.2/quadprog_1.5-8.zip'
Content type 'application/zip' length 36699 bytes (35 KB)
downloaded 35 KB

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.2/Rcpp_1.0.8.3.zip'
Content type 'application/zip' length 2882459 bytes (2.7 MB)
downloaded 2.7 MB

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.2/phangorn_2.8.1.zip'
Content type 'application/zip' length 2967618 bytes (2.8 MB)
downloaded 2.8 MB

package ‘ape’ successfully unpacked and MD5 sums checked
package ‘fastmatch’ successfully unpacked and MD5 sums checked
package ‘igraph’ successfully unpacked and MD5 sums checked
package ‘quadprog’ successfully unpacked and MD5 sums checked
package ‘Rcpp’ successfully unpacked and MD5 sums checked
package ‘phangorn’ successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\raliz\AppData\Local\Temp\Rtmpy2I0VC\downloaded_packages
> library(ape)
> library(phangorn)
```

```
> fdir <- system.file("extdata/trees", package = "phangorn")
> primates <- read.phyDat(file.path(fdir, "primates.dna"),
+                         format = "interleaved")
> dm  <- dist.ml(primates)
> treeUPGMA  <- upgma(dm)
> treeNJ  <- NJ(dm)
> plot(treeUPGMA, main="UPGMA")
```





1. Der Nicht-Primat wäre die Maus, da sie aus der „Wurzel“ des phylogenetischen Baumes kommt, also von der Gruppe der Primaten, die ebenso aus dieses „Wurzel“ kommen, getrennt ist. Beide Einheiten, die Maus und die Primaten, entwickeln sich getrennt voneinander, wie sehr leicht vom UPGMA-Baum zu entnehmen ist.
2. Lemur formt ein Cluster mit Bovine, das Cluster Lemur&Bovine formt ein Cluster mit Tarsier.
3. Der Mensch formt ein Cluster mit dem Schimpanse