Chapter 12

Solutions to Exercises

Grady Weyenberg and Ruriko Yoshida

Statistics department, University of Kentucky

Please note that a reader should use a treeviewer software such as FigTree http://tree.bio.ed.ac.uk/software/figtree/ to view a tree formatted in Newick format below.

Exercise 12.1. Upon completing this exercise, readers should have a base `R` installation which is ready to be used in further examples and exercises.

(a) Obtain and install the R system from http://cran.r-project.org/ or a local mirror. The basic installation includes little more than a basic console interface to the `R` interpreter, so you may also wish to download the RStudio IDE from https://www.rstudio.com/ for an improved GUI.

(b) Open `R` and enter the following command into the console.

```
"Hello, World!"
```

The interpreter should respond with the following.

```
[1] "Hello, World!"
```

(c) Install the `ape` add-on package by entering

```
install.packages("ape")
```

The system might ask for you to select a download location and will then proceed to download the add-on. When it is complete, you must activate the package by doing the following.

```
library(ape)
```

You should also take a moment to install and load the "`phangorn`" package, which provides more features used in subsequent examples.

(d) The program below reconstructs and plots a tree from a built-in example dataset.

```
data(woodmouse)
dm <- dist.dna(woodmouse)
trw <- nj(dm)
plot(trw)
```

(e) *Bootstrapping* is a nonparametric technique for exploring variance. A new dataset is generated by uniformly sampling (with replacement) from the observed data. The code chunk below generates a bootstrap resample from the woodmouse data and builds a new tree from the result.

```
i <- sample.int(ncol(woodmouse), replace = TRUE)
wm2 <- woodmouse[, i]
plot(nj(dist.dna(wm2)))
```

Execute the code a few times (a new dataset is generated each time) and observe the changes in the reconstructed tree. Are there any features of the tree which seem to be relatively constant or, conversely, particularly variable?

Algebraic and Discrete Mathematical Methods for Modern Biology. http://dx.doi.org/10.1016/B978-0-12-801213-0.00018-6

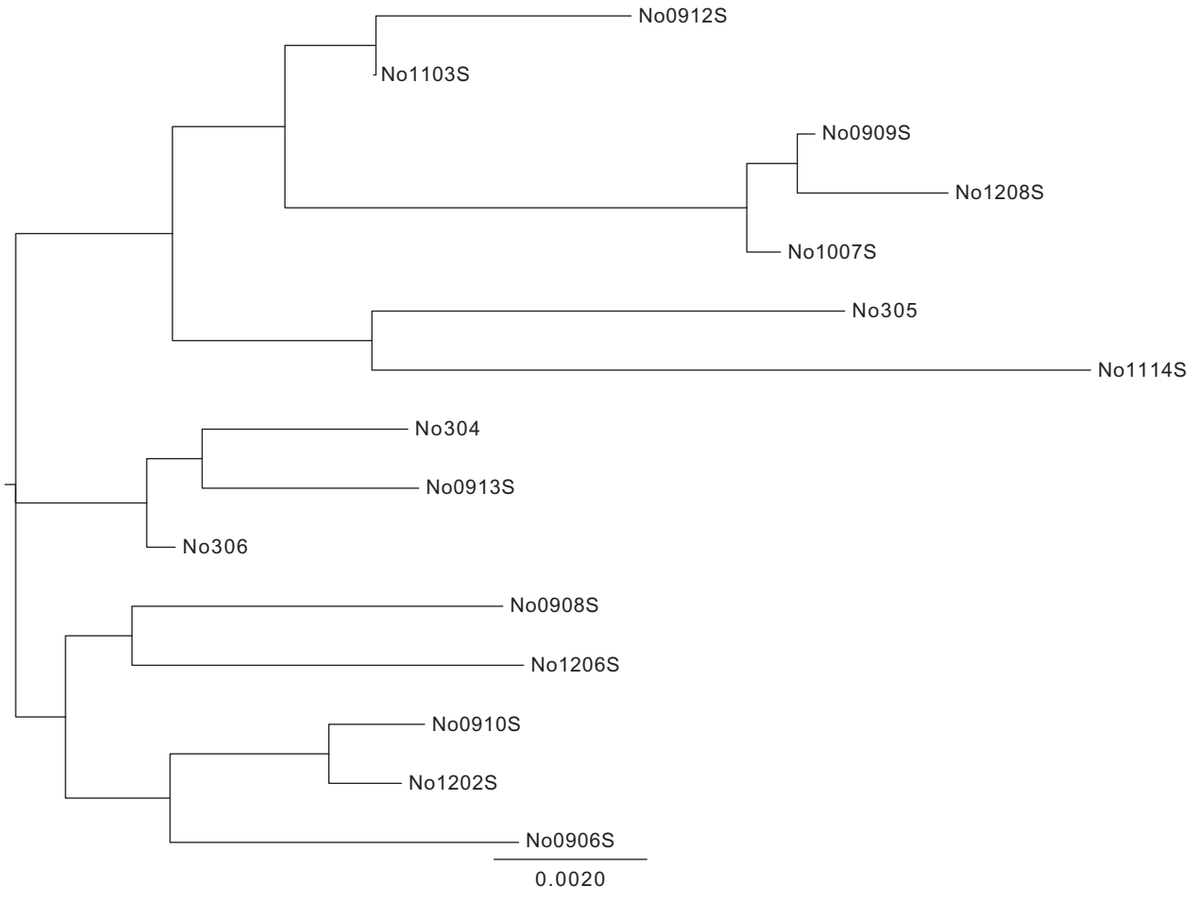

FIGURE S12.1 The first tree for Exercise 12.1.

Solution. See Figures S12.1 and S12.2. □

Exercise 12.2. In this exercise, you are asked to conduct the following procedure.

1. Download file: lolC introns trm.fasta at http://cophylogeny.net/courses/S12/lolC.
2. Open the file in a text editor. Select all and copy. In a browser, go to http://www.phylogeny.fr/.
3. Click on Phylogeny analysis, and select "One Click."
4. Go to the bottom of the page, and click Create Workflow.
5. Paste text (from lolC introns trm.fasta) into the text box.
6. Scroll to the bottom, and click Submit.
7. After several seconds, a page will appear with Tree Rendering results.
8. You will see several options and buttons on this page. Among those options, try those under Tree style. Also click on the Reroot (outgroup) button, then go to the tree and double-click on some branches. Describe what you observed. Feel free to try other options on this page.
9. Click the Alignment tab. Click (under Outputs), Alignment in FASTA format.

How does "alignment in FASTA format" differ from the input FASTA?

Solution. No question to answer here, just a list of things to do to set up for subsequent exercises. Each sequence has the same lengths by adding "-" to them. □

Exercise 12.3. *Newick format* is a way to represent a tree using parentheses and commas. One can find a nice explanation of Newick format at http://evolution.genetics.washington.edu/phylip/newicktree.html. This format was developed by the English mathematician Arthur Cayley.

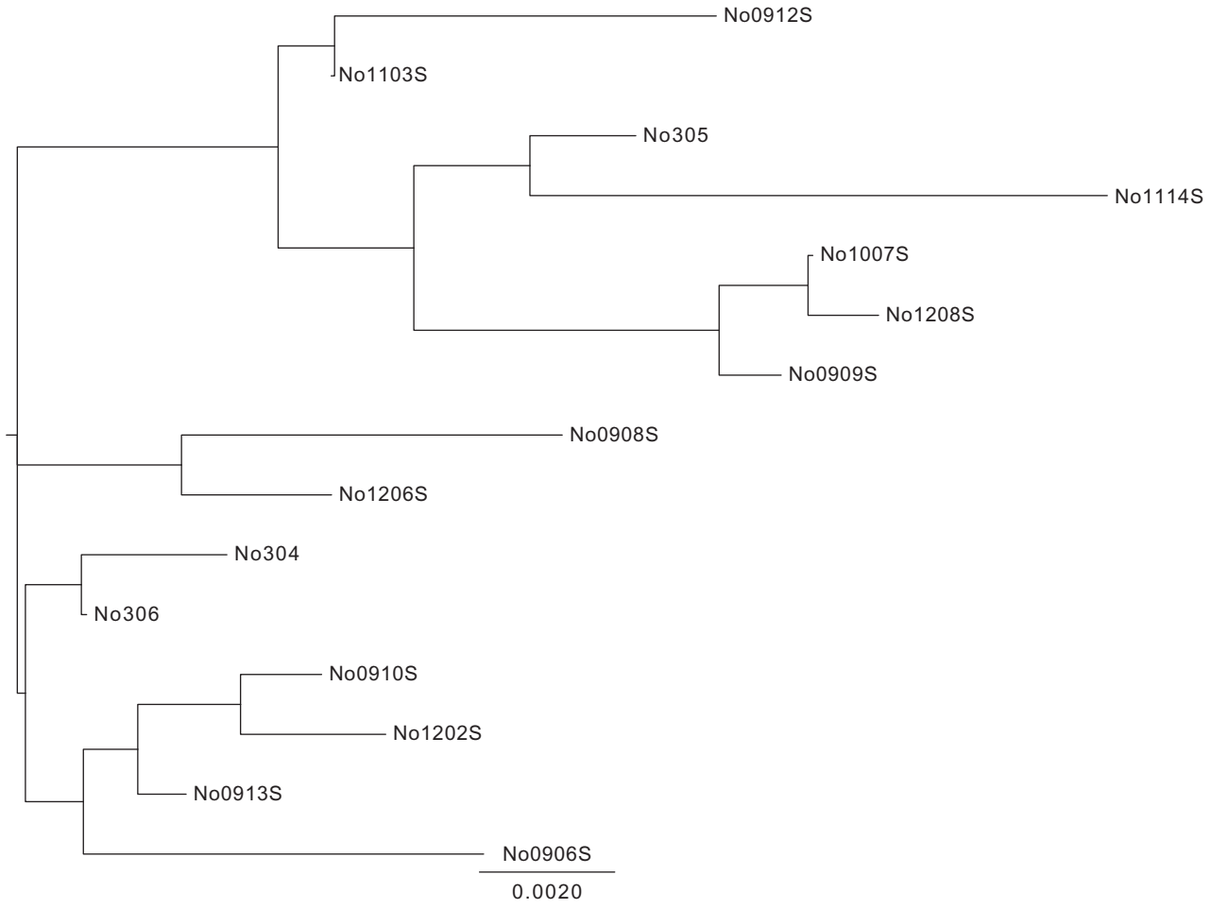

FIGURE S12.2 The second tree for Exercise 12.1.

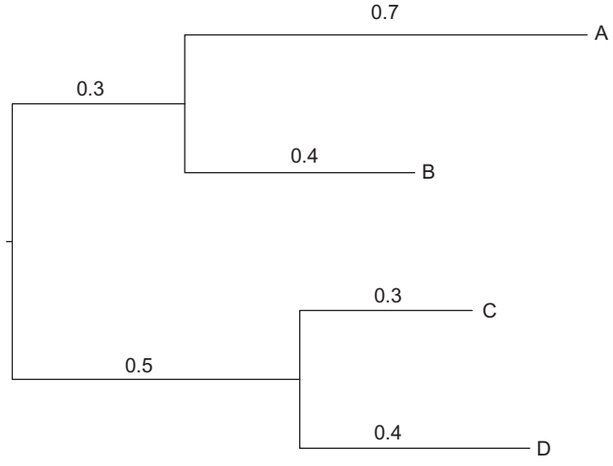

FIGURE 12.2 Tree for the Newick format in the example in Exercise 12.3.

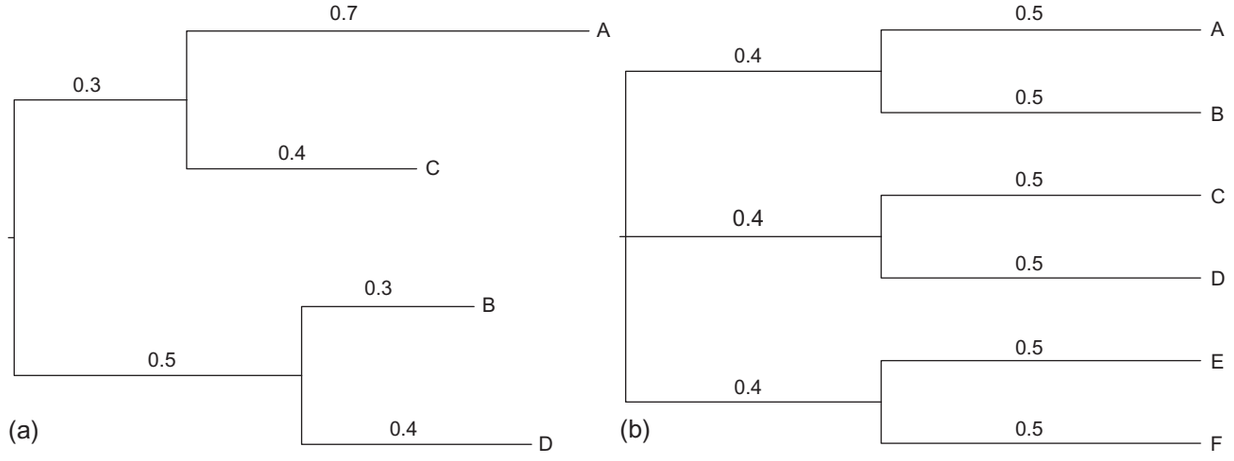

FIGURE 12.3 Trees for Exercise 12.3.

For example the following tree in Newick format:

```
((A:0.7,B:0.4):0.3,(C:0.3, D:0.4):0.5);
```

corresponds to the tree depicted in Figure 12.2.

Write the trees in Figure 12.3 in the Newick format.

Verify with Newick tree viewer at http://www.trex.uqam.ca/index.php?action=newick.

Solution.

(a) `((A:0.7,C:0.4):0.3,(B:0.3, D:0.4):0.5)`

(b)
```
((A:0.5,B:0.5):0.4,(C:0.5, D:0.5):0.4,
(E:0.5,F:0.5):0.4)
```

□

Exercise 12.4. Show that the Jukes-Cantor (JC) rate matrix satisfies the detailed balance condition.

Solution. Since Q is a symmetric matrix, and π is uniform, this immediately follows from plugging values into the detailed balance equation. □

Exercise 12.5. The matrix exponential of the JC rate matrix is given by the formula

$$P_{ij}(d) = \begin{cases} 0.25 + 0.75\mathrm{e}^{-4d/3} & \text{if } i = j, \\ 0.25 - 0.25\mathrm{e}^{-4d/3} & \text{if } i \neq j. \end{cases}$$

Show that Eq. (12.3) solves the likelihood optimization problem for the JC model.

Solution. This is a straightforward convex optimization problem, solvable by equating derivates to 0. The first issue is to note that since there are only two distinct values in P, most of the terms in the log-likelihood can be combined, and we find

$$l(d) \propto p(\log(1 - \exp(-4d/3)) - \log 4) + (1 - p)(\log(1 + 3\exp(-4d/3)) - \log 4)$$

where p is the proportion of sites that differ in the two sequences.

Take the derivative of this expression and set it equal to 0, then solve for d to get the closed-form solution given in the text. A complete solution should also take the second derivative and use that to argue that a maximum has indeed been found. □

Exercise 12.6. Extend the example code from the HKY model to implement the TN93 model. Use your extension to find the TN93 distance between HBA1 and HBA2. Compare your answer with that from `dist.dna`.

Solution.

```
P.TN93 <- function(d, k, pi) {
Q <- matrix(pi, 4, 4, byrow = TRUE)
Q[1, 2] <- Q[1, 2] * k[1]
```

```
    Q[2, 1] <- Q[2, 1] * k[1]
    Q[3, 4] <- Q[3, 4] * k[2]
    Q[4, 3] <- Q[4, 3] * k[2]
    diag(Q) <- -rowSums(Q) + pi
    beta <- 1/-sum(diag(Q) * pi)
    colnames(Q) <- rownames(Q) <- DNA.alphabet
    expm(d * beta * Q)
}

TN93.lnL <- function(par, nm) - sum(nm * log(P.TN93(d = par[1], k = par[2:3],
pi = pi)))

optim(par = c(d = 0.1, k = c(1.5, 1.5)), fn = TN93.lnL, nm = N12)$par

##       d      k1      k2
## 0.03825 1.40839 1.78680

dist.dna(hemo.aligned[1:2,], model="TN93")

##          HBA1
## HBA2 0.03797
```

□

Exercise 12.7. Suppose we have a distance matrix as follows:

	1	2	3	4	5	6
1	0	6	8	9	12	11
2	6	0	6	7	10	9
3	8	6	0	3	6	5
4	9	7	3	0	5	4
5	12	10	6	5	0	5
6	11	9	5	4	5	0

Reconstruct a tree using the NJ algorithm.

Solution. The Newick format of the output is

```
((1:4,2:2):3, 3:1, ((5:3,6:2):1, 4:1));
```

□

Exercise 12.8. In this exercise, you are asked to reconstruct a tree by the NJ algorithm from the alignment you have computed in Exercise 12.2.

1. Go to the website http://www.trex.uqam.ca/index.php?action=trex&menuD=1&method=2.
2. Save your alignment you have computed in Exercise 12.2 as PHYLIP format.
3. Upload your alignment in PHYLIP format.
4. Set Substitution model as Kimura 2 parameters.
5. Using the NJ algorithm to reconstruct a phylogenetic tree.

Solution. The Newick format of the output is

```
(((E2368lolC:3.1272,(E1125lolC:3.8386,E2772lolC:2.9800)79:0.5878)50:0.3187,
(E899lolC:2.4000,(E4074lolC:1.4124,(E507ylolC:0.0000,(E167ylolC:0.0000,(E55lolC:0.5719,
E19lolC:0.0000)90:0.0000)92:0.5667)58:0.1342)100:1.8455)93:1.2373)95:0.8369,
(E1031lolC:0.2207,(E57lolC:1.3568,E134ylolC:0.2187)96:0.8349)96:0.7382,
(E134xlolC:1.3158,(E507xlolC:0.0137,(E167xlolC:0.0136,
(E915lolC:0.4802,E999lolC:0.0183)17:0.0000)22:0.0000)73:0.2729)100:1.1514)
```

□

Exercise 12.9. Suppose we have a distance matrix as follows:

	1	2	3	4	5	6
1	0	6	8	9	12	11
2	6	0	6	7	10	9
3	8	6	0	3	6	5
4	9	7	3	0	5	4
5	12	10	6	5	0	5
6	11	9	5	4	5	0

Reconstruct a tree using the BME algorithm.

Solution. The Newick format of the output is

```
((1:4,2:2):3, 3:1, ((5:3,6:2):1, 4:1))
```

□

Exercise 12.10. In this exercise you are asked to reconstruct a tree by the BME algorithm from the alignment you have computed in Exercise 12.2.

1. Save your alignment you have computed in Exercise 12.2 as PHYLIP format named as "lolC_align.phy."
2. Open R.
3. Type in R as follows:

```
## Load directory
setwd("YOUR DIRECTORY")

## Load library
library(ape)

## Read in DNA sequences from Phylip
phy.data<-read.dna('lolC_align.phy', format=
   'interleaved')

## Reconstruct NJ tree with raw pairwise
    distances
dist.mat<-dist.dna(phy.data,model='K80')
fastme.tree<-fastme.bal(dist.mat)
write.tree(fastme.tree, file="outtree")
```

where "YOUR DIRECTORY" is the directory where the PHYLIP formatted file is located.

4. Using the FASTME algorithm to reconstruct a phylogenetic tree.

Solution. The Newick format of the output is

```
(E1125lolC:0.03875168504,E2772lolC:0.02746262585,((E899lolC:0.02472708856,
((E507ylolC:-0.0002145610484,((E55lolC:0.005362995488,
E19lolC:-5.790593183e-05):5.790593183e-05,E167ylolC:-5.790593183e-05
):0.005533873636):0.0002430240739,
E4074lolC:0.01587004355):0.01963947451):0.01366053536,
(E2368lolC:0.03034345945,((E1031lolC:0.002328382182,
(E57lolC:0.01357739696,E134ylolC:0.002428649119):0.008400386344):0.006062508326,
(E134xlolC:0.01359442236,(((E915lolC:0.00532578498,
E167xlolC:-2.069542378e-05):2.069542378e-05,E507xlolC:-2.069542378e-05):2.069542378e-05,
E999lolC:-2.069542378e-05):0.002427842038):0.01318280255)
:0.005633378671):0.00368810759):0.005368223534)
```

□

Exercise 12.11. In this exercise, you are asked to reconstruct a tree by the MLE algorithm from the alignment you have computed in Exercise 12.2.

1. Save the alignment as PHYLIP format.
2. Go to the website http://www.phylogeny.fr/
3. Upload your alignment in PHYLIP format.
4. Set HKY model (noted as HKY85 model on the website).
5. Write your email address and reconstruct a MLE tree via PHYML by submitting it.

Solution. The Newick format of the output is

```
((((((((((((E507xlolC:0.000001,E915lolC:0.005295)0.000000:0.000001,
E999lolC:0.000001)0.000000:0.000001,E167xlolC:0.000001)0.746000:0.002641,
E134xlolC:0.013410)0.976000:0.013503,((E134ylolC:0.002647,E57lolC:0.013370)
0.923000:0.008037,E1031lolC:0.002644)0.886000:0.005334)0.749000:0.003049,
E2368lolC:0.032255)0.459000:0.008281,(E2772lolC:0.029540,E1125lolC:0.039357)
0.835000:0.007428)0.921000:0.011919,E899lolC:0.022140)0.988000:0.021584,
E4074lolC:0.016057)0.000000:0.000001,E507ylolC:0.000001)0.952000:0.005302,
E167ylolC:0.000001)0.000000:0.000001,E19lolC:0.000001,E55lolC:0.005305)
```

□

Exercise 12.12. In this exercise, we will use ape and phangorn libraries in R.

1. Install ape and phangorn libraries in R.
2. Type the following

```
library(ape)
library(phangorn)
data(chloroplast)
(mTAA <- modelTest(chloroplast, model=c("JTT", "WAG", "LG")))
```

What the output did you get?

Solution. The output from R is

```
> (mTAA <- modelTest(chloroplast, model=c("JTT", "WAG", "LG")))
     Model  df    logLik      AIC      BIC
1      JTT 162 -76842.40 154008.8 155069.2
2    JTT+I 163 -74010.14 148346.3 149413.2
3    JTT+G 163 -72762.74 145851.5 146918.4
4  JTT+G+I 164 -72715.49 145759.0 146832.4
5      WAG 243 -76233.86 152953.7 154544.3
6    WAG+I 244 -73532.54 147553.1 149150.2
7    WAG+G 244 -72410.46 145308.9 146906.0
8  WAG+G+I 245 -72379.40 145248.8 146852.5
9       LG 243 -76018.33 152522.7 154113.2
10    LG+I 244 -73305.82 147099.6 148696.8
11    LG+G 244 -71969.09 144426.2 146023.3
12  LG+G+I 245 -71920.12 144330.2 145933.9
```

□