

Homework #1B (R, Databases, Alignment)

Submit in the inbox outside Clark S260 before 5:00 PM.

This second problem set covers R, databases, and basic alignment. Note that you will need to view the electronic version of this pdf to see all the URLs.

1. *Reading and Literature*

The following materials will be useful to look at if you need a refresher or introduction to R. They are hyperlinked so you will need to look at the online pdf.

- (a) [Official R tutorial](#) (skip to Appendix A)
- (b) [Basic R tutorial](#) (sections 2,3,7)
- (c) [R Reference Card](#)
- (d) Data Analysis and Graphics Using R: Chapters 1, 2, 14 (skim chapters as necessary)

The following materials will be useful to browse for the topics covered this week

- (a) Databases
 - i. Start with the [NCBI minicourses](#). Note that not all of these are online.
 - ii. [BMI 214 Introductory Lecture](#) (start at page 63)
 - iii. [NAR 2007 Database issue](#)
 - iv. [NAR 2006 Web Server issue](#)
- (b) Sequence Alignment
 - i. CS 262 Lectures on Sequence Alignment: [Lecture 2](#), [Lecture 3](#), [Lecture 4](#), [Lecture 13](#) (slides 16–43)
 - ii. BMI 214 Lectures on Sequence Alignment: [Pairwise Alignment](#), [Multiple Alignment](#)
 - iii. Batzoglou review: [The Many Faces of Sequence Alignment](#)
 - iv. Edgar review: [Multiple Sequence Alignment](#)

2. *NCBI introduction*

I would like you to complete the following tutorials. The [Biobar](#) may come in very useful.

- (a) Read the [Entrez Tutorial](#) and [Entrez Manual](#). Do the questions on [MLH1 \(colon cancer\)](#) and [PER2 \(circadian\)](#). You do not need to hand in answers to these questions as they are provided; the point is to increase your familiarity with the Entrez query system. After doing this tutorial, find the location of the SNP which causes sickle-cell anemia. Print a screenshot of the webpage and include it with your HW.

- (b) Do the interactive [BLAST Tutorial](#). Then, using the skills you developed in both the previous question and the BLAST tutorial, find the proteins which are sequence similar to CC3035, a cell cycle regulatory protein in *Caulobacter crescentus*. This will be a web page on NCBI with either BLAST results or BLINK results (pre-cached BLAST). Print a screenshot and include it with your HW.

3. R and Sequence Alignment

This problem is an introduction to programming in R.

- (a) Download the BLOSUM62 matrix from NCBI: <http://www.ncbi.nlm.nih.gov/Class/FieldGuide/BLOSUM62.txt>. You can read this directly into a matrix in R with the following command (typed all on one line). Note that this cleans up the headers after download to make the row and column headers symmetrical.

```
b62 <- read.table(`http://www.ncbi.nlm.nih.gov/Class/FieldGuide/BLOSUM62.txt`,
header=TRUE,row.names=1)
colnames(b62)[24] <- "*"
```

- (b) Write a program to do pairwise sequence alignment via dynamic programming for a specified similarity matrix S and linear gap penalty g as covered in Lecture 4. See the sequence alignment references in the first problem if you want more details/figures on alignment. Your program should have the invocation: `myalign(x,y,S,g)`, where the first two arguments are character strings. The third and fourth arguments are optional: the similarity matrix S and the gap penalty. Here is an example of a function header in R which defines defaults for these optional arguments:

```
myalign <- function(x,y,S=b62,g=-4) {
  print("Default arguments can be programmatically initialized; see this example")
  b62 <- read.table("http://www.ncbi.nlm.nih.gov/Class/FieldGuide/BLOSUM62.txt",
                    header=TRUE,row.names=1)
  colnames(b62)[24] <- "*"
  print(S)
}
```

Your code should return an object with the following elements:

- i. $2 \times L$ matrix of characters with '-' for gaps (the optimal alignment matrix)
- ii. The $(M + 1) \times (N + 1)$ matrix of optimal scores. The first row and column correspond to $F(0, j)$ and $F(i, 0)$.
- iii. The $(M + 1) \times (N + 1)$ matrix of traceback pointers. There are many ways to represent this, choose a good one. Note that there will not be any pointers along the left and top rows.

Apply it to align the protein sequences with GIs 90111240 and 16131163 from the `COG00009J.faa` file which you built in the previous homework. Use the BLOSUM62 matrix for similarity scoring, and $g = -4$ for the gap penalty. Include a print out of your code and of a sample run which prints out the alignment.

- (c) Visually compare your results to NCBI's [bl2seq](#) program with the appropriate matrix. Note that you will need to choose blastp in the drop down menu.
- (d) Find the nucleotide gene sequences which correspond to the proteins you just aligned. Generate an alignment for these sequences with your code. Use the BLASTN matrix for similarity scoring (+5 for match, -4 for mismatch) and a -2 linear gap penalty. Compare the alignments visually and remark on whether they are similar or dissimilar.