

BIO 480/680 Lab Exercise 12
Protein Sequence Alignment
Due date: 09 May 2025

For this exercise, you will be studying the amino acid sequence for a protein that is very important for the process of photosynthesis in plants – ribulose biphosphate carboxylase. This protein is made up of multiple subunits and so we will be considering only the large subunit. The amino acid sequences for this subunit have been identified for a number of plant species and there are some differences in the sequences among plants. For comparing plant species with publicly available sequence data for this protein, you will be provided with the sequences found in the following six plants:

the conifer *Abies magnifica* (Abies.aa),
a moss *Chenia leptophylla* (Chenia.aa),
a flowering plant, soybean *Glycine max* (Glycine.aa),
an early flowering plant *Magnolia grandiflora* (Magnolia.aa),
another flowering plant, pea *Pisum sativum* (Pisum.aa),
and another moss *Physcomitella patens* (Physcomitrella.aa).

For analysis of the amino acid sequence data for this protein, you will be working with one of the basic methods for aligning sequences of proteins, as discussed in lecture.

Sequence alignment using FASTA

The FASTA programs used in this section are available from Bill Pearson at the University of Virginia (<https://fasta.bioch.virginia.edu/wrpearson/fasta/>), look under the folder executables/. For Windows, you can collect the zip file with Win32 in the name. This will get you a large set of files. But we only need one part of this program for our exercise (the FASTA program file fasta36.exe). As a simpler alternative, see the course web page under “Sample Programs”, where you can download this program plus the sequence data files. Apple Mac OSX versions can be downloaded from Bill Pearson’s web page at: <https://fasta.bioch.virginia.edu/wrpearson/fasta/>. Look under the directory “executables” for the macosx versions.

The web page <https://fasta.bioch.virginia.edu/wrpearson/fasta/> also makes available a “fasta_guide.pdf” if you are interested in more detail.

Running the fasta36 program

The fasta program is not a program with a graphical user interface. Like the programs we have been creating for most of our exercises, it is a simple program that has to be run from a command line. Therefore you generally cannot run this program by just double-clicking on the program name – it will start but just finish immediately without leaving you any results to view.

Here's an easy way for running this program that involves opening a command line window in the directory where the program is located. Let's say you have downloaded the program and the necessary data files (amino acid sequences) to a folder on your computer. You have a folder open showing all those files.

Position your mouse pointer over a blank space in the folder view. Hold down the shift key and right-click. You should be able to select the menu item "Open command window here". That will open up a command line window where you can type in the program name to run it (see below). If the choice that appears says "Open Powershell window here", you can pick that choice and make it work as shown below, but it might be easier to revert back to the earlier command window by selecting the Powershell choice and typing "cmd" when the Powershell window opens.

Run the fasta36 program and compare each species sequence with every other species as follows:

```
fasta36 first_file second_file
```

If you want to save the results to a data file, here's an example comparing the conifer with one of the mosses and saving the results to a data file called "output.txt":

```
fasta36 -O output.txt Abies.aa Chenia.aa
```

Note that the option for the output file is the upper-case letter "O" and it goes before the input file names. Of course it might make sense to use the plant names for the file name instead of the one chosen here.

For your report, you will need to present data as a table showing the comparison results for each sequence being compared (*Abies* versus *Chenia*, *Abies* versus *Glycine*, *Abies* versus *Magnolia*, *Abies* versus *Pisum*, *Abies* versus *Physcomitrella*, and so forth...). There are six species being compared, so your table should have 6 rows and 6 columns. Include in the table the percent identity and the Smith-Waterman score for each comparison.

For your conclusions, think about the sequence comparisons between these species. Do the similarities make sense from your knowledge of the evolutionary relationships between these plants species?