CENG 465 – Introduction to Bioinformatics
Spring 2006-2007

Assignment #2 (Programming Assignment)

Due Date: May 01, 2007

Programming Assignment About Phylogenetic Trees

In this assignment, you are going to implement the neighbor joining algorithm to construct a phylogenetic tree of a family of protein sequences, the globins. Figure 1 taken from an American Scientist article shows an example phylogenetic tree of a set of globin proteins. You are going to construct a similar tree using the neighbor joining algorithm for the same proteins given in the figure. This assignment is more than just a programming assignment; because it includes an extensive amount of data collection as well. The main steps of your assignment are listed as follows:

1. Get the sequence data for the globin proteins listed in Figure 1. You may use the protein search page at the NCBI web site:
   a. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search&DB=protein

   or get the sequences from the SCOP database

   b. http://scop.mrc-lmb.cam.ac.uk/scop/data/scop.b.b.b.b.d.html

   It will not be straightforward to find the protein sequences you are looking for. Look for keywords that may describe the protein you are searching. Also look for external links that may lead you to the sequence information. There are other databases that may contain sequence information like Swiss-Prot or PDB.

2. Pairwise align all the pairs of acquired sequences using Smith-Waterman dynamic programming algorithm. At the end of this step, you will have a scoring matrix including all the globing sequences. Use the BLOSUM62 matrix for scoring the alignments. (You can find BLOSUM62 matrix easily on the Internet.)
3. Convert the scoring matrix between globin sequences to a *distance matrix* using any function of your choice. Basically, your function should invert the similarity measure to a distance measure; smaller distances will imply more similar sequences.
4. Use the Neighbor Joining algorithm on the distance matrix to create a phylogenetic tree of the globin sequences.

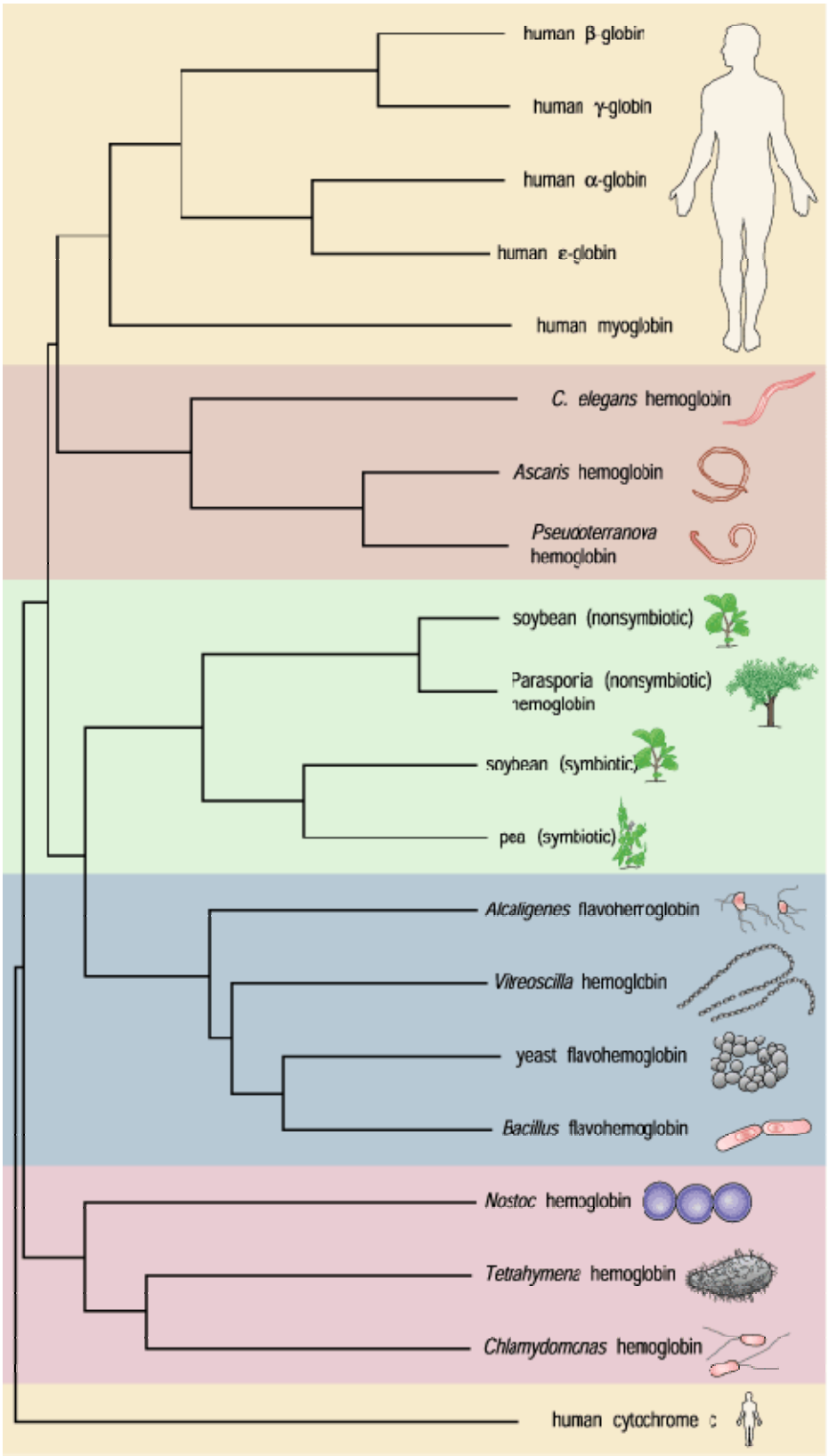

Figure 1. The Evolution of Hemoglobin, *American Scientist*, Volume: 87 Number: 2 Page: 126, 1999.

Deliverables:

- The source code of your program(s). You may use any programming language of your choice.
- A figure showing the pyhlogenetic tree output found by your program. The tree can be drawn manually.
- Answer the following questions:
    - How does the tree generated by your program compare to the phylogenetic tree given in Figure 1? Describe in a paragraph.
    - Interpret the edge lengths of the phylogenetic tree produced by your program.
    - What was the most time consuming part of the assignment, data collection or programming?

Submission:

Send the deliverables as a zip bundle or as a tarball to umuero@ceng.metu.edu.tr by the due date.