Humboldt-Universität Zu Berlin
Institute for Informatics and Theoretical Biology (ITB)

Silke Trißl   **Email: trissl@informatik.hu-berlin.de**

Lectures on Bioinformatics: Assignment 4

More information about the exercises can be found at
http://zope.informatik.hu-
berlin.de/forschung/gebiete/wbi/teaching/archive/sose09/vl_biophysik

Exercise 1: Modify your program from last assignment such that it can perform the following tasks:

a) Comparing protein sequences (instead of only DNA)

b) Reading a substitution table from a file (additionally to the sequences to be compared) and using this table during the alignment.

Exercise 2: **BLOSUM Matrices**

Download the matrices BLOSUM50, BLOSUM80 and BLOSUM100 from the BLOCKS FTP server (ftp://ftp.ncbi.nih.gov/repository/blocks/unix/blosum/BLOSUM). Format these matrices in such way that your program can read them. Calculate the similarity between two protein sequences using:

a) An identity matrix (InsDelRep = -1, Match = +1)

b) With all three BLOSUM matrices.

c) Compare the results from both types of matrices qualitatively.

Protein sequences are available on the web site.

Exercise 3: **Database Search and Sequence Alignment tool BLASTX**

PHENYLKETONURIA is a relatively frequent hereditary disease, which must/can be treated on newborn children.

a) Find out in the OMIM database which protein is responsible for this illness. On which chromosome is the corresponding gene located?

b) Search the DNA sequence of this gene in the NCBI Genbank database. Describe its length, name and accession number.

c) Search similar proteins using BLASTX

      i.   Give the accession number of homologous DNA sequences for the following organisms (in each case the first occurrence for the species in the hit list)

            - Homo Sapiens
            - Mus Musculus
            - Bos Taurus
            - Rattus Novegicus
            - Gallus gallus
            - Xenopus tropicalis
            - Drosophila Melanogaster
            - Danio rerio

d) Write a program which compares the similarity between all pairs of those sequences (global alignment, INS/DEL/REP=-1, Match=+1). Which are the closest pair, which species have the longest distance from each other?

      Note: Results of all exercises and the code should be zipped into one file including a README file describing the commands for running your code. Your code must be able to perform all task asked above.