Softwarewerkzeuge der Bioinformatik

Prof. Dr. Volkhard Helms
PD Dr. Michael Hutter, Markus Hollander,
Marie Detzler
Winter Semester 2020/2021

Saarland University
Center for Bioinformatics

Exercise Sheet 2

Sequence Analysis: Pairwise Alignments

Learning objective: *The goal is to learn when to use which BLAST–search (ProteinBLAST, NucleotideBLAST, MegaBLAST, PSI–BLAST), and which parameters (E–value, matrix, query database etc.) are useful depending on the search. Additionally, you are going to compute a pairwise sequence alignment with the Needlemann–Wunsch algorithm and answer some theoretical questions.*

Exercise 2.1: Dynamic Alignment

Compute a **global** alignment of the sequences **ACDEFAFGHI** and **KDELAFG** using the **Needlemann–Wunsch** algorithm.

		A	C	D	E	F	A	F	G	H	I
K											
D											
E											
L											
A											
F											
G											

Global alignment:

Exercise 2.2: ProteinBLAST

The lecture slides could be useful for answering the following questions.

a) What is the definition of the *expected threshold* (*E–value*)? Why is an E–value threshold of 10 not particularly useful? What are sensible E–values?

b) How does the *word size* affect run time and accuracy?

c) What is special about the first hit of a BLAST search against a normal database?

d) Run a **ProteinBLAST** search (http://blast.ncbi.nlm.nih.gov/Blast.cgi) for the protein **P00042**

   i. against the **UniProtKB/Swiss–Prot** database with default parameters. Find the 10 proteins with the highest homology to P00042 and display their sequences. What kind of proteins are we dealing with?

   ii. against the **non–redundant** database with an E–value threshold of 0.001. What are the differences to the previous search? For which types of organisms were results found?

Exercise 2.3: MegaBLAST

Select **human** as the genome on the BLAST main page. Search for the mRNA **NM_175054** of the human gene *HIST4H4* with **megaBLAST** in the database **Genome (GRCh38.p13)**.

a) On which chromsome is *HIST4H4* located?

b) Is there a paralogue?

c) Find two or three directly neighbouring genes of *HIST4H4*.

Exercise 2.4: PSI–BLAST

a) Use **ProteinBLAST** to search for many very distantly related homologues of the protein **Q57997** in the **non–redundant** database with an E–value threshold of 0.02.

   What are suitable substitution matrices?

b) Run the same search with **PSI–BLAST** and a threshold of 0.001 for the maximal E–value of the sequences used for constructing the PSSM.

c) What are the differences between the results of a) and the 1. iteration in b)?

d) How do the results of part b) change with further iterations?

Have fun!