Grundlagen der Bioinformatik Assignment 4: Hierarchical Clustering SS 2016

Yvonne Lichtblau

Vorstellung Lösungen Übung 3

Overview – Assignment 3 (20P)

(1) Local Alignment (10P)
Vorstellung durch zwei Gruppen

(2) Global Alignment (5P)
Vorstellung durch eine Person

(3) Aligning real sequences (5P)
Vorstellung durch eine Person

Assignment 4
Hierarchical Clustering

Overview – Assignment 4 (20P)

(1) Global alignment (5P)

(2) Finding sequences (3P)

(3) Hierarchical Clustering (12P)

(1) Global Alignment (5P)

- Get back to your program for local alignment

- Modify the program to:

  - Calculate the global alignment
  - Work with amino acid sequences
  - Use BLOSUM62 as cost matrix (NCBI, EMBOSS, ...)
    - Cost matrix must be loaded and not hardcoded

```
   A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V  B  Z  X  *
A  4 -1 -2 -2  0 -1 -1  0 -2 -1 -1 -1 -1 -2 -1  1  0 -3 -2  0 -2 -1  0 -4
R -1  5  0 -2 -3  1  0 -2  0 -3 -2  2 -1 -3 -2 -1 -1 -3 -2 -3 -1  0 -1 -4
N -2  0  6  1 -3  0  0  0  1 -3 -3  0 -2 -3 -2  1  0 -4 -2 -3  3  0 -1 -4
D -2 -2  1  6 -3  0  2 -1 -1 -3 -4 -1 -3 -3 -1  0 -1 -4 -3 -3  4  1 -1 -4
C  0 -3 -3 -3  9 -3 -4 -3 -3 -1 -1 -3 -1 -2 -3 -1 -1 -2 -2 -1 -3 -3 -2 -4
Q -1  1  0  0 -3  5  2 -2  0 -3 -2  1  0 -3 -1  0 -1 -2 -1 -2  0  3 -1 -4
(...)
```

(2.1) Find Sequence (1P)

- Phenylketonuria (PKU) is a frequent hereditary disease
  - Can be well treated if found early
  - Life long and strict low-phenylalanine diet
  - Otherwise severe effects on brain development

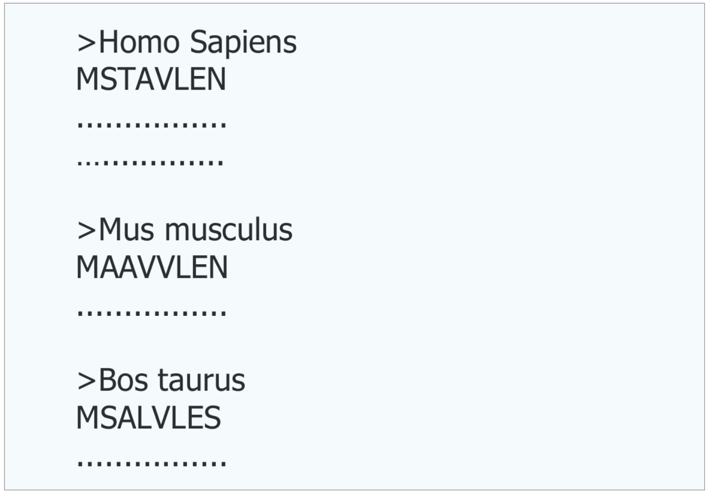
- Find the disease causing protein in OMIM database
  - What is the name of the disease causing protein?

- Retrieve Sequence of (human) protein from UniProt
  - What is the UniProt-ID?
  - How many amino acids is the protein long?

(2.2) Find Sequence Homologues (2P)

- Retrieve homologous protein sequences using NCBI’s BLASTP
  - Use non-redundant sequences for BLAST
  - *Homo sapiens, Mus musculus, Bos taurus, Rattus norvegicus, Gallus gallus, Xenopus tropicalis, Drosophila melanogaster, Danio rerio*
  - State the used accession numbers for all 8 sequences

- Store sequences in a single FASTA file (e.g. sequences.fasta):



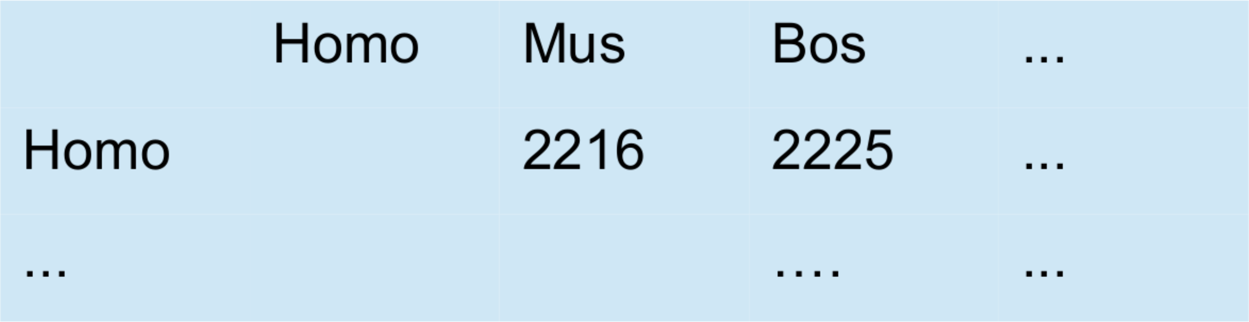
```
>Homo Sapiens
MSTAVLEN
................
................

>Mus musculus
MAAVVLEN
................

>Bos taurus
MSALVLES
................
```

sequences.fasta

(3.1) Hierarchical Clustering (7P)

- Implement the algorithm for hierarchical clustering

  - Program reads a single FASTA file + scoring matrix
  - Compute similarity matrix on all pairs of sequences from the file
  - Print all pairwise scores in tabularized manner



Homo	Mus	Bos	...
Homo	2216	2225	...
...		….	...

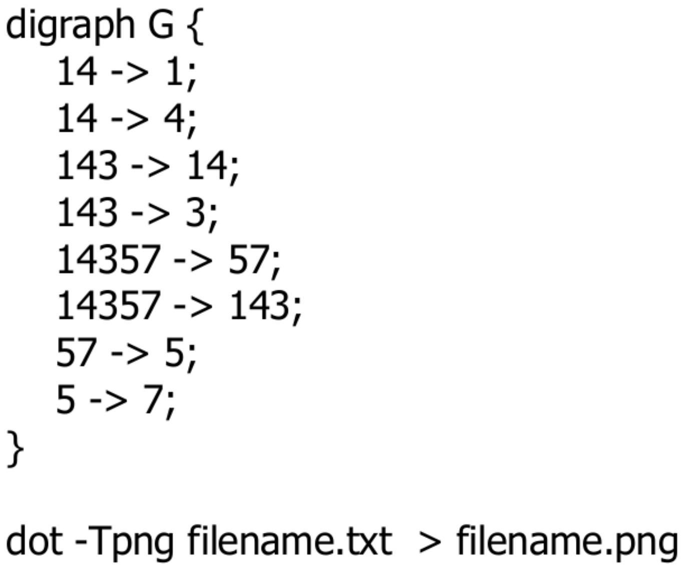
(3.1) Hierarchical Clustering (7P)

- Build a guide tree using hierarchical clustering
- Of course, you need to find the maximum in the similarity matrix
- Output the tree as text as follows (sequences numbered by order on slide 8/in FASTA file):
    - Assume sequence 1 and 4 are merged to '14', then 5 and 7 to 57, then the virtual sequence 14 is merged with 3 etc, the output of your programm should look like this:
      (1,4), (5,7), (14,3) etc.
- Programmaufruf:

```
java -jar assignment4_GRXY.jar sequences.fasta blosum.txt
```

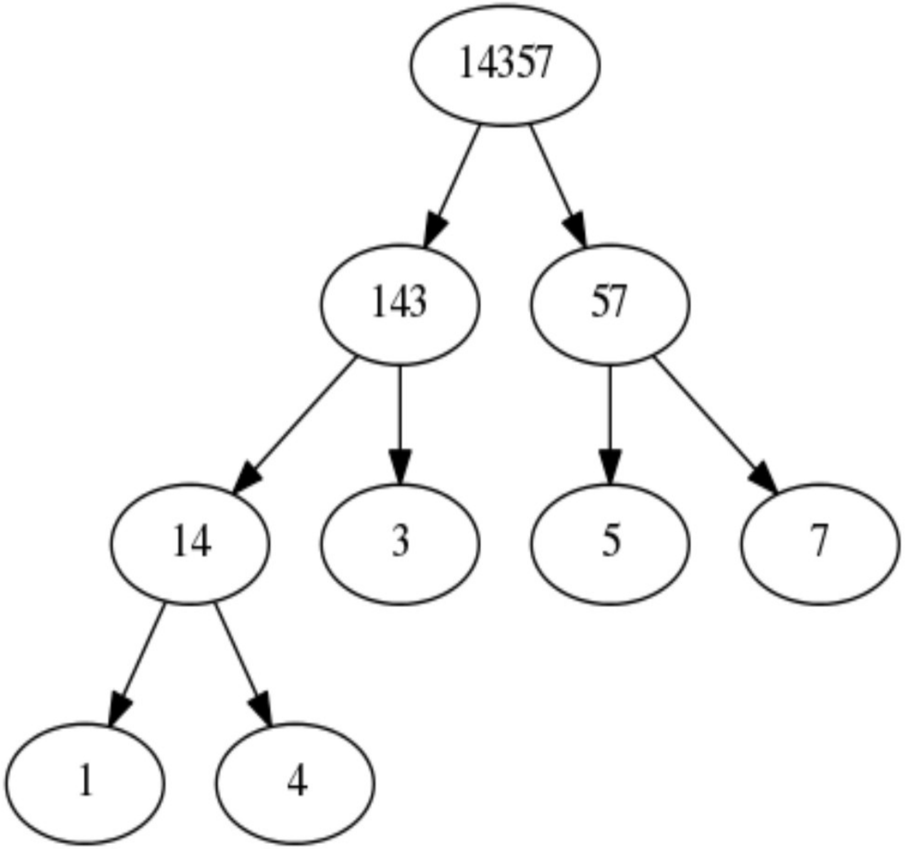
(3.2) Visualization (5P)

- Draw the tree such that novel clusters are added from bottom-to-top and from left-to-right
  - As in the picture below
  - E.g., using graphviz http://www.graphviz.org/



```
digraph G {
   14 -> 1;
   14 -> 4;
   143 -> 14;
   143 -> 3;
   14357 -> 57;
   14357 -> 143;
   57 -> 5;
   5 -> 7;
}
```

```
dot -Tpng filename.txt  > filename.png
```



Abgabe

- Abgabe bis Mittwoch den 22.06.2016 um 23:59 Uhr

- Abgabe per Email an: yvonne.lichtblau@informatik.hu-berlin.de
(gerne auch Fragen zur Übung per Email)
    - PDF mit
        - Task 2.1: Proteinname, UniProt-ID, Sequenz Länge
        - Task 2.2: 8 Accession Nummern
        - Task 3.1: Tabelle mit paarweisen Alignment Scores
        - Task 3.2: Abbildung des Trees

    - FASTA Datei mit den 8 Sequenzen

    - Code als .jar Datei wie beschrieben (Übung 1)
    Ausgabe des Programms wie auf Folien 9+10 beschrieben

    - Sourcecode