
Assignment 3

Hierarchical Clustering

1. Global alignment (5P)
 2. Find homologous 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
 - Retrieve from NCBI
 - Retrieve from EMBOSS
 - ...
- Cost matrix must be loaded and not hardcoded!

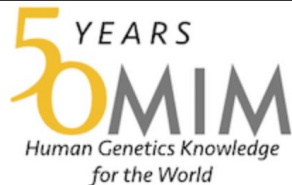
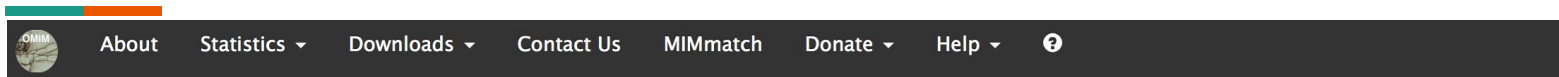
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0	-2	-1	0	-4
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3	-1	0	-1	-4
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3	3	0	-1	-4
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3	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
1. Find the disease causing protein in OMIM database
 2. Retrieve Sequence of (human) protein from UniProt

OMIM database



OMIM[®]

Online Mendelian Inheritance in Man[®]

An Online Catalog of Human Genes and Genetic Disorders

Updated May 29, 2018

Advanced Search : [OMIM](#), [Clinical Synopses](#), [Gene Map](#)

Need help? : [Example Searches](#), [OMIM Search Help](#)

Mirror site : mirror.omim.org

2.2 Find homologous sequences (2P)

Retrieve homologous protein sequences using NCBI's BLASTP -> blast the UniProt sequence

- 1.1. Write out finding for BLAST
 1. Homo sapiens
 2. Mus musculus
 3. Bos taurus
 4. Rattus norvegicus
 5. Gallus gallus
 6. Xenopus tropicalis
 7. Drosophila melanogaster
 8. Danio rerio

The screenshot shows the UniProt BLAST interface. At the top, there's a UniProt logo and a search bar. Below the search bar, there are tabs for BLAST, Align, Retrieve/ID mapping, and Peptide search. A yellow banner indicates that from June 20, 2018, all traffic will be automatically redirected to HT. The main heading is "BLAST" with a subheading "How to use this tool". Below this, there's a description of BLAST. On the right, there's a color scale from 100 (red) to 80 (green). The "Filter by" section on the left lists various filters: Reviewed (4) Swiss-Prot, Unreviewed (246) TrEMBL, With 3D structure (2), and Proteomes (156). Below this, there's a "Popular organisms" section with links to Human (5) and Mouse (2). The "Overview" section on the right shows a table of results. The table has two columns: "Entry" and "Protein names". The first entry is P00439, which is Phenylalanine-4-hydroxylase (Homo sapiens). The second entry is A0A024RBG4, which is Phenylalanine hydroxylase, isoform CRA_a (Homo sapiens). The third entry is H2Q6R0, which is PAH isoform 8 (Pan troglodytes). The fourth entry is Q8TEY0, which is Phenylalanine hydroxylase (Homo sapiens).

UniProt

UniProtKB

BLAST Align Retrieve/ID mapping Peptide search

From June 20, 2018 all traffic will be automatically redirected to HT

BLAST

How to use this tool

The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences, which can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

100 80

Filter by¹

- Reviewed (4) Swiss-Prot
- Unreviewed (246) TrEMBL
- With 3D structure (2)
- Proteomes (156)

Popular organisms

- Human (5)
- Mouse (2)

Edit and resubmit Order by: Score

Overview

Show all 250

Entry	Protein names
P00439	Phenylalanine-4-hydroxylase (Homo sapiens)
A0A024RBG4	Phenylalanine hydroxylase, isoform CRA_a (Homo sapiens)
H2Q6R0	PAH isoform 8 (Pan troglodytes)
Q8TEY0	Phenylalanine hydroxylase (Homo sapiens)

Example, you can use different blasters

What we want



Provide information on:

- Name of the disease causing protein
 - Its UniProt-ID
 - Its amount of amino acids
- State the used accession numbers for all eight sequences from blasting the Uniprot sequence
 - Store sequences in a single FASTA file (e.g. sequences.fasta):

```
>Homo Sapiens  
MSTAVLEN  
.....  
.....  
  
>Mus musculus  
MAAVLEN  
.....  
  
>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)

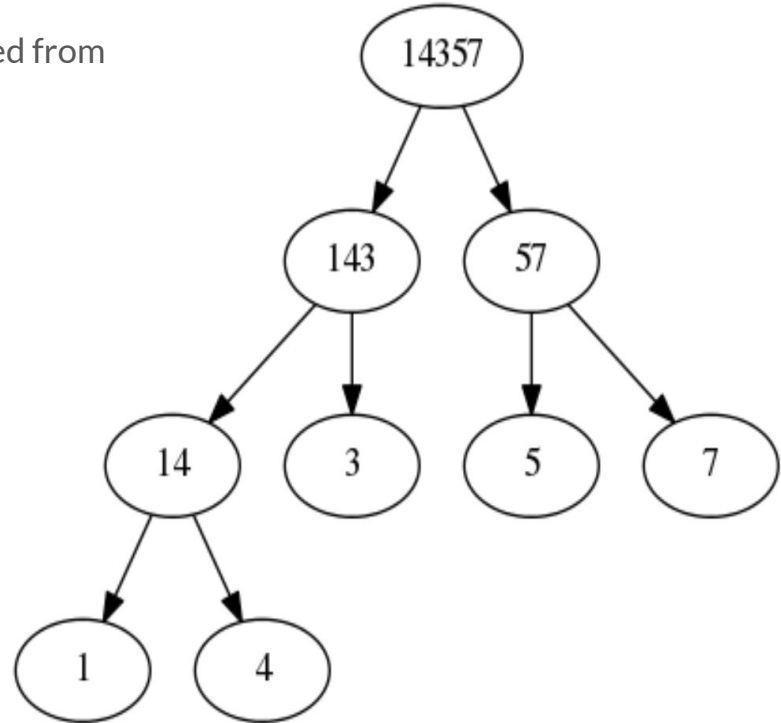


- Build a guide tree using hierarchical clustering
- Find maximum in the similarity matrix
- Program call
 - `java -jar GdBioinf_[Assignment_nr]_Gruppe_[Gruppen_nr].jar sequences.fasta blosum.txt`
- Output the tree
 - BE AWARE OF SEQUENCE NUMBERING, see below
- Assume sequence 1 and 4 are merged to '14', then 5 and 7 to 57, then the virtual sequence 14 is merged with 3 etc, the output of your program should look like this: (1,4), (5,7), (14,3) etc.
 1. Homo sapiens
 2. Mus musculus
 3. Bos taurus
 4. Rattus norvegicus
 5. Gallus gallus
 6. Xenopus tropicalis
 7. Drosophila melanogaster
 8. Danio rerio

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

What we want

Submit

- Global alignment algorithm
- Algorithm for clustering & distance matrix calculation
- Similarity matrix of all eight sequences with similarity score
- State the used accession numbers for all eight sequences from blasting the Uniprot sequence

```
>Homo Sapiens  
MSTAVLEN
```

```
.....  
.....
```

```
>Mus musculus  
MAAVVLEN
```

```
.....
```

```
>Bos taurus  
MSALVLES
```

```
.....
```

sequences.fasta

- Store sequences in a single FASTA file (e.g. sequences.fasta)
- Cluster-structure (see slide before)

Deadline



- Deadline for submission 13.06.2018 at 12:00 p.m.
- [SUBMIT HERE](https://box.hu-berlin.de/u/d/f94821b8a39b41259e9e/)
 - <https://box.hu-berlin.de/u/d/f94821b8a39b41259e9e/>
- Code as .jar/.py.R
- Remember correct source code nomenclature
 - GdBioinf_[Assignment_nr]_Gruppe_[Gruppen_nr].jar