Exercises to Introduction to Bioinformatics Assignment 3: Hierarchical Clustering

Philippe Thomas

Alignment with a Substitution Matrix (5P)

- Get back to your program for global alignment written for the last assignment
- Modify the program such that (5P)
  - Modify the program such that it works with amino acid sequences, not with DNA
  - It takes a second command line parameter, the file name of a substitution matrix
    - Use the BLOSUM80 matrix (download from the web)
  - Modify the algorithm such that it computes a similarity using the matrix (and not constants as before)

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

Hierarchical Clustering (8+3P)

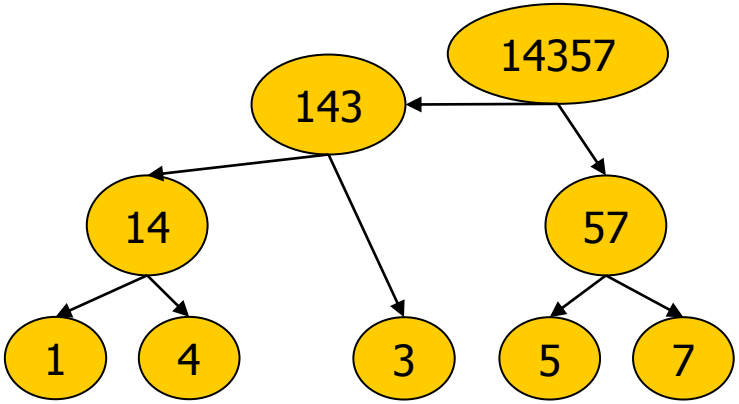
- Implement the algorithm for hierarchical clustering as explained in the last lecture (8P)
  - Program reads a FASTA file + scoring matrix
  - Compute similarity matrix on all pairs of sequences from the file
    - Print all pairwise scores
  - Build a tree using hierarchical clustering
    - Of course, you need to find the maximum in the matrix (similarity)
  - Output the tree as text as follows
    - If first sequences 1 and 4 are merged to a sequence 14, than 5 and 7 to 57, then the virtual sequence 14 is merged to 3 etc, the output should look like this: (1,4), (5,7), (14,3) etc.

Hierarchical Clustering (8+3P)

- Draw the tree such that novel clusters are sorted from left-to-right manually or automatically (e.g. graphviz) (3P)

e.g. in graphviz

```
digraph G {
    14 -> 1;
    14 -> 2;
    143 -> 14;
    143 -> 3;
    14357 -> 57;
    14357 -> 143;
    57 -> 5;
    5 -> 7;
}
```



ClustalW (4)

- Use the ClustalW website to compute the guide tree for the eight sequences (2P) (http://www.ebi.ac.uk/Tools/msa/clustalw2/)
  - Make a screenshot
- Compare the two trees. Are there any differences? Explain why (2P).

Submission

- Submit all requested data as plain text by Wednesday, 7.06.2012, 23.59
  - JAR to compute the alignment + tree (source+class files)
  - Output:
    - All pairwise alignment scores
    - All merging steps [e.g. (1,4), (5,7), (14,3)]
  - Both trees (PDF)
  - The explanation for the differences, if there are any
  - Approximate working time !
- Send by mail to Stefan Kröger!
  - Kroeger <at> informatik.hu-berlin.de
- Questions should be addressed to Stefan and Philippe