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 3

Sequence Analysis: Multiple Sequence Alignment (MSA) and Phylogeny

Learning objective: *The goal is to learn how to generate multiple sequence alignments, how to interpret them e.g. regarding sequence conservation and their usefulness for different types of questions. Additionally, you are going to apply the Sankoff algorithm and learn how to work with phylogenetic trees.*

Exercise 3.1: Homologous sequences, conserved domains and phylogenetic trees

Tools for generating multiple alignments: http://www.ebi.ac.uk/Tools/msa

a) Save the sequence of protein **Q38856** together with 9 homologous sequences in multi–fasta–format.

b) Find highly conserved parts of the sequences with a tool of your choice.

c) Do all amino acids have to be highly conserved in order to conclude that the proteins are homologous?

d) Let's assume that you want to locate the active centre of a protein but only have the protein sequence without the corresponding structure. How can a multiple sequence alignment help you to solve this problem?

e) Generate a multiple sequence alignt of 50 homologous sequences with the same tool.

f) What differences do you observe between the two alignments?

g) Look at the phylogenetic tree of the sequences in 3.1.e) and find three biological groups (plants, fungi and animals).

Exercise 3.2: Comparison of various tools

The following multiple sequence alignments were generated with different tools:

Tool	Protein	Alignment
ClustalW	FOS_Rat	MMFSGFNADYEASSSRCSSASPAGDSLSYYHSPADSFSSMGSPVNTQDFC
	FOS_MOU	MMFSGFNADYEASSSRCSSASPAGDSLSYYHSPADSFSSMGSPVNTQDFC
	FOS_CHIC	MMYQGFAGEYEAPSSRCSSASPAGDSLTYYPSPADSFSSMGSPVNSQDFC
	FOSB_MOU	-MFQAFPGDYDS-GSRCSS-SPSAESQ--YLSSVDSFGSPPTAAASQE-C
	FOSB_HU	-MFQAFPGDYDS-GSRCSS-SPSAESQ--YLSSVDSFGSPPTAAASQE-C
		:.. .:*:: .***** **:.:* * *..***.* :.. :*: *
MAFFT	FOS_Rat	MMFSGFNADYEASSSRCSSASPAGDSLSYYHSPADSFSSMGSPVNTQDFC
	FOS_MOU	MMFSGFNADYEASSSRCSSASPAGDSLSYYHSPADSFSSMGSPVNTQDFC
	FOS_CHIC	MMYQGFAGEYEAPSSRCSSASPAGDSLTYYPSPADSFSSMGSPVNSQDFC
	FOSB_MOU	-MFQAFPGDYD-SGSRCSS-SPSAES--QYLSSVDSFGSPPTAAASQE-C
	FOSB_HU	-MFQAFPGDYD-SGSRCSS-SPSAES--QYLSSVDSFGSPPTAAASQE-C
		:.. .:*: ..***** **:.:* * *..***.* :.. :*: *
MUSCLE	FOS_Rat	MMFSGFNADYEASSSRCSSASPAGDSLSYYHSPADSFSSMGSPVNTQDFC
	FOS_MOU	MMFSGFNADYEASSSRCSSASPAGDSLSYYHSPADSFSSMGSPVNTQDFC
	FOS_CHIC	MMYQGFAGEYEAPSSRCSSASPAGDSLTYYPSPADSFSSMGSPVNSQDFC
	FOSB_MOU	-MFQAFPGDYD-SGSRCSS-SPSAESQ--YLSSVDSFGSPPTAAASQE-C
	FOSB_HU	-MFQAFPGDYD-SGSRCSS-SPSAESQ--YLSSVDSFGSPPTAAASQE-C
		:.. .:*: ..***** **:.:* * *..***.* :.. :*: *
Clustal Omega	FOS_Rat	MMFSGFNADYEASSSRCSSASPAGDSLSYYHSPADSFSSMGSPVNTQDFC
	FOS_MOU	MMFSGFNADYEASSSRCSSASPAGDSLSYYHSPADSFSSMGSPVNTQDFC
	FOS_CHIC	MMYQGFAGEYEAPSSRCSSASPAGDSLTYYPSPADSFSSMGSPVNSQDFC
	FOSB_MOU	-MFQAFPGDYDSGS-RCSSSPSA---ESQYLSSVDSFGSPPTA-AASQEC
	FOSB_HU	-MFQAFPGDYDSGS-RCSSSPSA---ESQYLSSVDSFGSPPTA-AASQEC
		:.. .:*:: * ****: * : * * .***.* : ::: *

Compare the MSAs.

a) Are there differences regarding the gap arrangement?

b) Does this change the degree of conservation in the coloured columns?

Exercise 3.3: Conserved motifs

Use Clustal Omega to generate a multiple sequence alignment of the sequences provided on the lecture website (sequences1.fasta). Locate uninterrupted, highly conserved areas of at least length 10 and save them as potential motifs for exercise sheet 4 (based on FOSB_MOUSE).

Exercise 3.4: Outgroup

a) Generate an MSA of the sequences provided on the lecture website (sequences2.fasta).

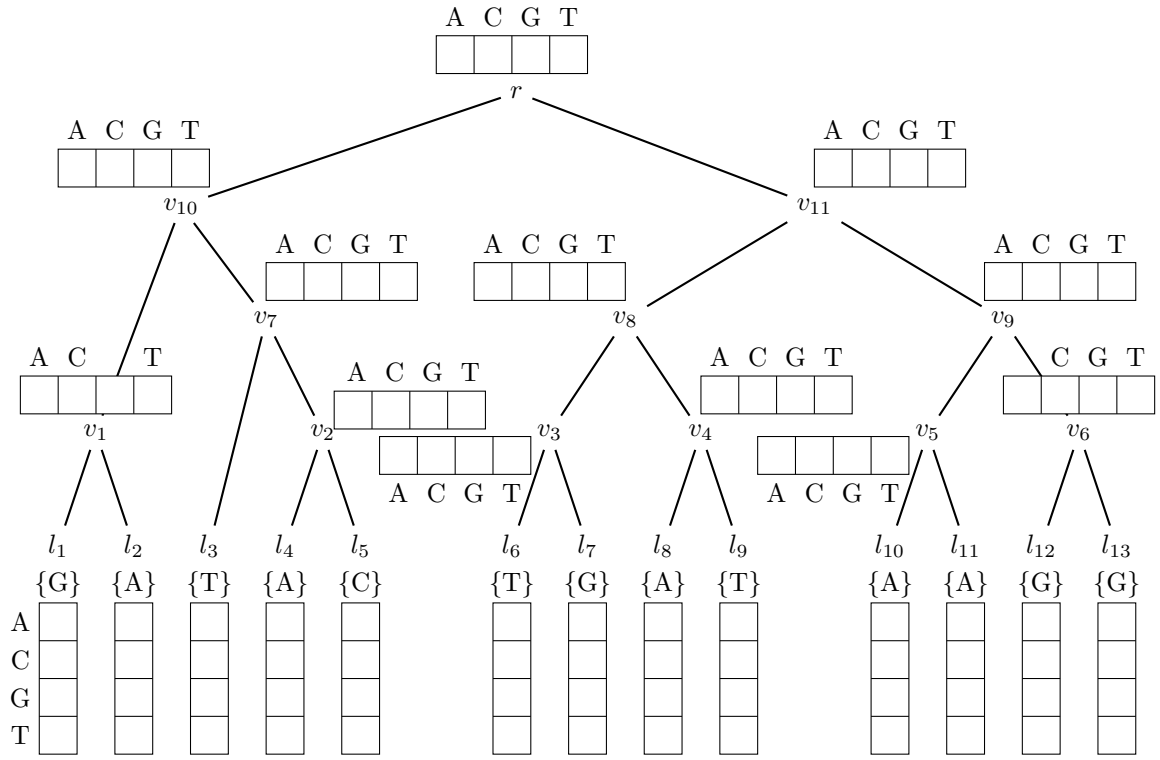
b) Is everything conserved?

c) Which species differs from the rest?

d) Construct a phylogenetic tree.

Exercise 3.5: Sankoff algorithm

Which base was likely in the ancestor sequence at the given position of the alignment? Use the Sankoff algorithm and the given cost function.



$\longrightarrow$ Base in the ancestor sequence:

Cost function:

	A	C	G	T
A	0	2	1	2
C	2	0	2	1
G	1	2	0	2
T	2	1	2	0

Have fun!