Bioinformatics I
WS 15/16
Tutor: Alexander Seitz

1	2	3	$\sum$

Jonas Ditz
& Benjamin Schroeder

Assignment 4

(Abgabe am 9. November 2015)

Theoretical Assignment - *Optimal multiple alignment*

To calculate our MSA we use the recursion written down on page 50 in the script.

$$F(i_1,i_2,i_3) = max \begin{cases} F(i_1-1,i_2-1,i_3-1) + s_{SP}(a_{1i_1},a_{2i_2},a_{3i_3}) \\ \\ F(i_1-1,i_2-1,i_3) + s_{SP}(a_{1i_1},a_{2i_2},-) \\ F(i_1-1,i_2,i_3-1) + s_{SP}(a_{1i_1},-,a_{3i_3}) \\ F(i_1,i_2-1,i_3-1) + s_{SP}(-,a_{2i_2},a_{3i_3}) \\ \\ F(i_1-1,i_2,i_3) + s_{SP}(a_{1i_1},-,-) \\ F(i_1,i_2-1,i_3) + s_{SP}(-,a_{2i_2},-) \\ F(i_1,i_2,i_3-1) + s_{SP}(-,-,a_{3i_3}) \end{cases}$$

Such a recursion results in a three-dimensional matrix. Since such a matrix is really difficult to sketch, we split it into three two-dimensional matrices.

0	0	C	C	T
0	**0**	-2	-4	-6
C	-2	**-2**	-4	-10
T	-4	-8	-8	-6
T	-6	-10	-12	-10

T	0	C	C	T
0	-2	-4	-6	-8
C	-4	-4	-6	-10
T	-6	-8	**-4**	2
T	-8	-10	-8	2

C	0	C	C	T
0	-4	-6	-8	-10
C	-6	2	2	-8
T	-8	-6	-2	0
T	-10	-8	-6	**-6**

Figure 1: DP matrix of MSA, traceback is shown in red

If we fill the DP matrix using this recursion, we get several optimal alignments. One of them is the following:

$$\begin{pmatrix} C & T & T \\ - & T & C \\ C & C & T \end{pmatrix} \tag{1}$$

with score $\alpha_{SP}(A^*) = S(A,B) + S(A,C) + S(B,C) = -2 + 2 + (-6) = -6$.

Theoretical Assignment - *Progressive alignment*

Progressive alignment methodes are one way to create multiple sequence alignments (MSA), in this task we had to accomplish manualy a progressive alignment. The first step for the progressive alignment was to generate the pairwise global distance matrices. Afterwards the global optima were entered into a new table, which was used to generate the guide tree in the second step.(figure 2).

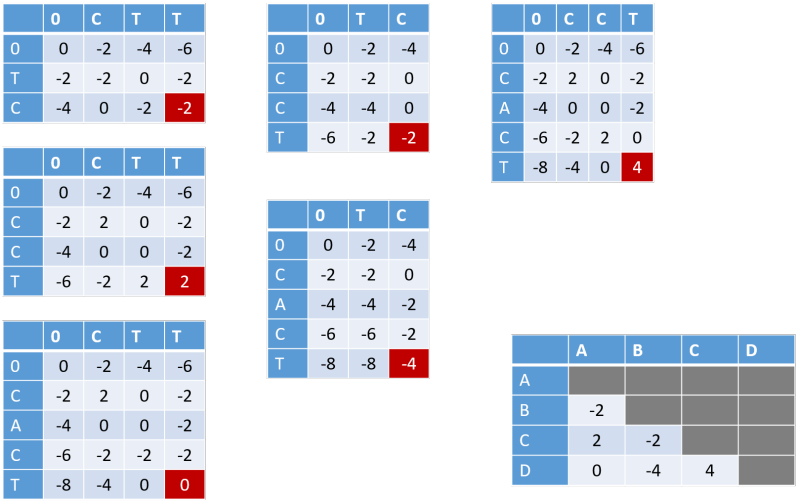

Figure 2: Global alignments lead to the basis table used to generate the guide tree

In step 2 the guide tree was created by the UPGMA methode in 3 substeps. After every step, which added a new cluster, the distance table was updated. The result from this step was the guide tree, which was used to apply the 'complete alignment' methode in the last step, the progressive alignment step.(figure 3)

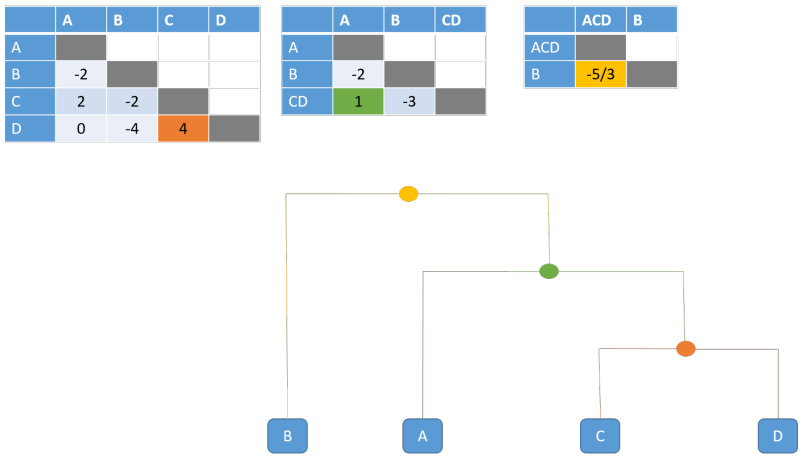

Figure 3: The guide tree is made from the final table from step 1

The last step was creating the actual MSA. The 'complete alignment' methode was used. This means after every step, the distance of the new sequences was determined to each other sequence in the cluster. The optimal resulting score was used to get the best alignment for the new Sequence.

after 3 steps the MSA was completed. (figure 4)

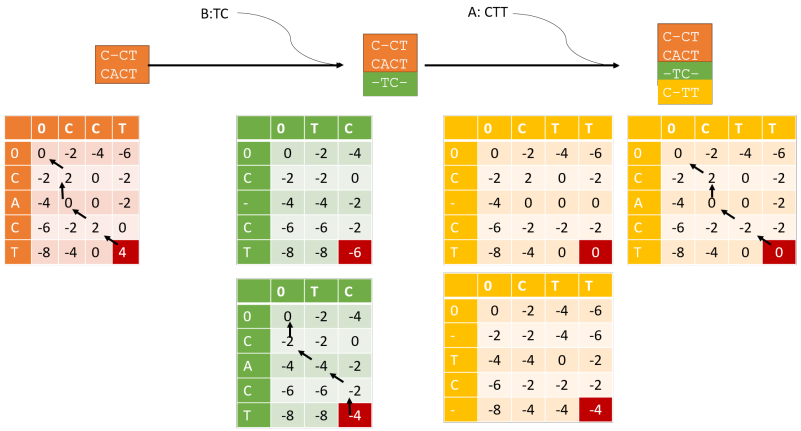

Figure 4: Complete alignment methode is used to get an MSA from the guidetree from step 2

Practical Assignment - *Comparing multiple alignment*

Multiple sequence alignments (MSA) are very important to detect similarities between several sequences and e.g. determine whether a sequence belongs to a certain family or not. Hence, many different programs, such as ClustalW, were developed to calculate a MSA. We calculated a multiple sequence alignment for three files (BB11007, BB20002 and Prolyl-tRNA), each file contained several sequences, with different programs to compare the results. The used programs are MAFFT [1], MUSCLE [2, 3] and ClustalΩ [4]. The resulting MSAs were compared with a gold standard provided by BaliBase [5]. This database provide a large number of reference MSAs as well as a program "BaliScore", which can be used to calculate different scores to evaluate your MSA. For this assignment we are interested in the total column score, which is the percentage of accurately reconstructed columns of the reference MSA, and the sum-of-pair score, which is for BaliScore the percentage of pairs of aligned residues that are similar in the reference and the reconstructed MSA. Furthermore, a important information is the runtime of different programs. We used the bash command *time* to collect this information.

The resulting total column scores can be found in figure 5. All calculated sum- of-pairs scores are written down in figure 6 and the runtime of all programs can be found in figure 7.

total column score	**ClustalΩ**	**MUSCLE**	**MAFFT**
BB11007	$c_{1,1}$	$c_{2,1}$	$c_{3,1}$
BB20002	$c_{1,2}$	$c_{2,2}$	$c_{3,2}$
Prolyl-tRNA	$c_{1,3}$	$c_{2,3}$	$c_{3,3}$

Figure 5: total column score of all used programs for example files

SP-score	**ClustalΩ**	**MUSCLE**	**MAFFT**
BB11007	$s_{1,1}$	$s_{2,1}$	$s_{3,1}$
BB20002	$s_{1,2}$	$s_{2,2}$	$s_{3,2}$
Prolyl-tRNA	$s_{1,3}$	$s_{2,3}$	$s_{3,3}$

Figure 6: sum-of-pairs score of all used programs for example files

runtime	**ClustalΩ**	**MUSCLE**	**MAFFT**
BB11007	$t_{1,1}$	$t_{2,1}$	$t_{3,1}$
BB20002	$t_{1,2}$	$t_{2,2}$	$t_{3,2}$
Prolyl-tRNA	$t_{1,3}$	$t_{2,3}$	$t_{3,3}$

Figure 7: runtime of all used programs for example files

References

[1] K. Katoh, K. Kuma, H. Toh, and T. Miyata, "Mafft version 5: improvement in accuracy of multiple sequence alignment. nucleic acids research," Nucleic Acids Research, vol. 33, no. 2, 2005.

[2] R. C. Edgar, "Muscle: multiple sequence alignment with high accuracy and high throughput," Nucleic Acids Research, vol. 32, no. 5, 2004.

[3] R. C. Edgar, "Muscle: a multiple sequence alignment method with reduced time and space complexity," BMC Bioinformatics, vol. 5, no. 113, 2004.

[4] F. Sievers, A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, J. D. Thompson, and D. G. Higgins, "Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega," Molecular Systems Biology, vol. 7, no. 1, 2011.

[5] A. Bahr, J. Thompson, J.-C. Thierry, and O. Poch, "Balibase (benchmark alignment database): enhancements for repeats, transmembrane sequences and circular permutations," Nucleic Acids Research, vol. 29, no. 1, 2001.