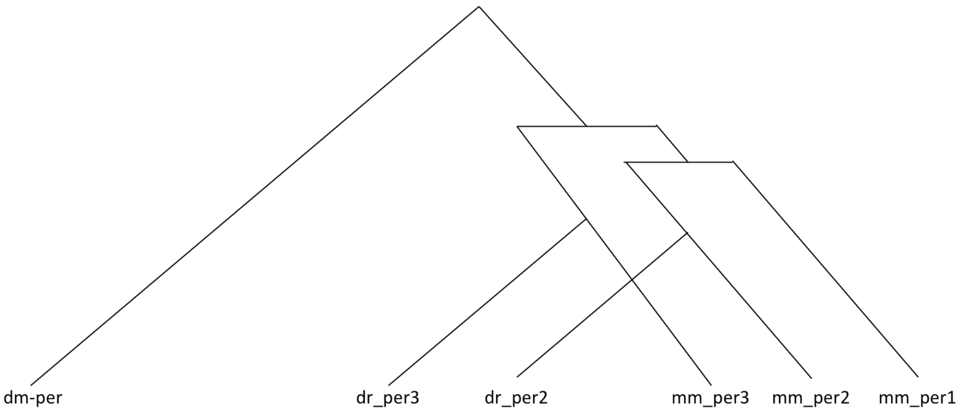
Exercises - Week 6

Genomics and bioinformatics

1 Phylogenetic trees

1. Start with the highest BLAST score (984) and link the two corresponding proteins (mm_per2 and dr_per2). Then continue with the next highest scores until the tree is completed. Because of the query and reference lengths, you noticed that BLAST scores are not symmetric, but following this rule there is a unique solution tree.



The gene tree

On this graph duplications events are represented with horizontal lines. One can differentiate speciation and duplication events by the fact that we see the same species (but another version of the gene) on two branches separated by a duplication (as for **mm_per1** and **mm_per2**). When there is a speciation event, species are all different on both parts after the event (as for **dm_per** and **dr_per3**).

2. - Orthologous pairs correspond to reciprocal best hits *between two different species.* Here is the way to find them:
   Choose one line of the table, say **dm_per**. Compare with the fish: on our line, the best score between **dr_per2** and **dr_per3** is 125 for **dr_per3**. Then take the **dm_per** column, and verify that the same gene has the best score among dr genes. In this case, **dr_per3** is best with 123, compared to 114. So **dm_per** and **dr_per3** are reciprocal best hits, thus they are orthologs by definition.
   Here is the list of orthologs:
   dm_per and **dr_per3**
   dm_per and **mm_per1**
   dr_per2 and **mm_per2**
   dr_per3 and **mm_per3**

Note: in particular with this definition there are genes issued from speciation events that are not formally orthologs. Closest pairs (such as **dr_per2** and **mm_per2**) are always orthologs.

- Non-reciprocal best hits are considered as paralogous pairs. In the same species: **mm_per1** and **mm_per2**.
- A paralogous pair in different species: **mm_per1** and **dr_per2**.

2 Another HMM

2.1 The Viterbi algorithm

Characters T_1, T_2 and T_3 are respectively A, T and C. Choose a cell in the table: its position defines n and s. Then replace emission and transition probabilities in the Viterbi formula.

Be careful here that if there is 0.8 probability of emitting one of A-T, there is 0.4 probability of emitting A and 0.4 for T; similarly for G-C.

	-	A	T	C
S	1	0	0	0
N	0	↖ 0.2	← 0.079	← 0.0078
I	0	0.125	0.028	0.0063

The most probable hidden sequence associated to ATC is NNN.