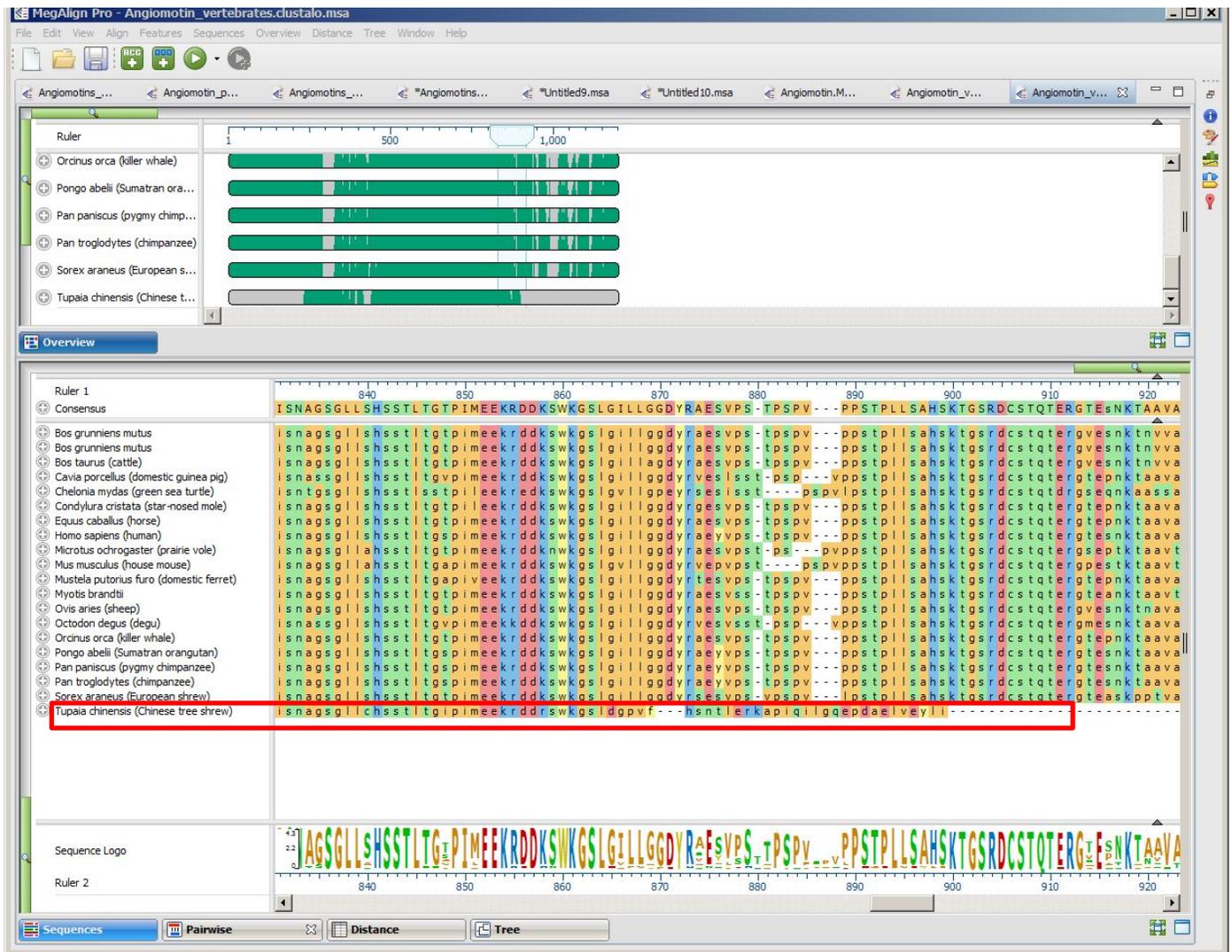


MegAlign Pro Pairwise Alignment Tutorials

All demo data for the following tutorials can be found in the *MegAlignProAlignments.zip* archive [here](#).

Tutorial 1: Multiple versus pairwise alignments

1. Extract the contents of the *MegAlignProAlignments.zip* folder and double-click on *Angiotenin_vertibrates.clustalo.msa* to launch it in MegAlign Pro. This project contains a collection of Angiotenin proteins from a diverse set of vertebrate species, already multiply-aligned with Clustal Omega. The figure below shows the alignment with the Sequences and Overview views scrolled to show the two tree shrew sequences.



2. Notice that the *Tupaia* sequence (box) seems to end abruptly with an unaligned run of amino acids. If you can't see the abrupt end, move the Sequence view's green "horizontal zoom slider" to the left until the Sequences view appears as in the image above.
3. To investigate this using pairwise alignment, click on the name of the *Tupaia* sequence to select it, and then Ctrl-click (Win) or Cmd-click (Mac) on the name of the *Sorex* sequence to add it to the selection. You can select these sequences in either the Overview or the Sequences view.
4. Next right-click on either name and select "Align Pairwise ..."
5. In the Align Pairwise dialog box, select "Global: Needleman-Wunsch" from the "Using:" drop-down menu. Since we are specifically interested in the C-terminal ends of the sequences, the "Global" option is ideal.
6. Keep the default settings for all other dialog options and click OK. The alignment will run and a Pairwise View will appear.
7. Scroll down the Pairwise View to the end of the alignment and notice that it is now clear that the *Tupaia* sequence contains a large deletion and that the C-terminal ends of the two shrew sequences are nearly identical.

Pairwise alignments can also be used to help interpret multiple alignments with gaps that might obscure the relationships between sequences. While the Pairwise view is still open, try some Global alignments between different species pairs. This is easily done by changing the sequence(s) in the drop-down menus at the top of the Pairwise view.

8. Change the sequence in the right drop-down menu to to *Homo sapiens* (human).
9. Scroll through the updated alignment and observe that the two sequences match pretty well, despite some mismatches and gaps. The header shows a similarity calculation (%Similar) of 91.3%.

Next, change the sequence shown in the left drop-down to *Pan troglodytes* (chimpanzee). The human and chimp sequences are 98.4% similar, with only an 11-residue gap differentiating them. This similarity can be seen with a much greater degree of clarity than could be seen looking at the Overview of the original multiple alignment



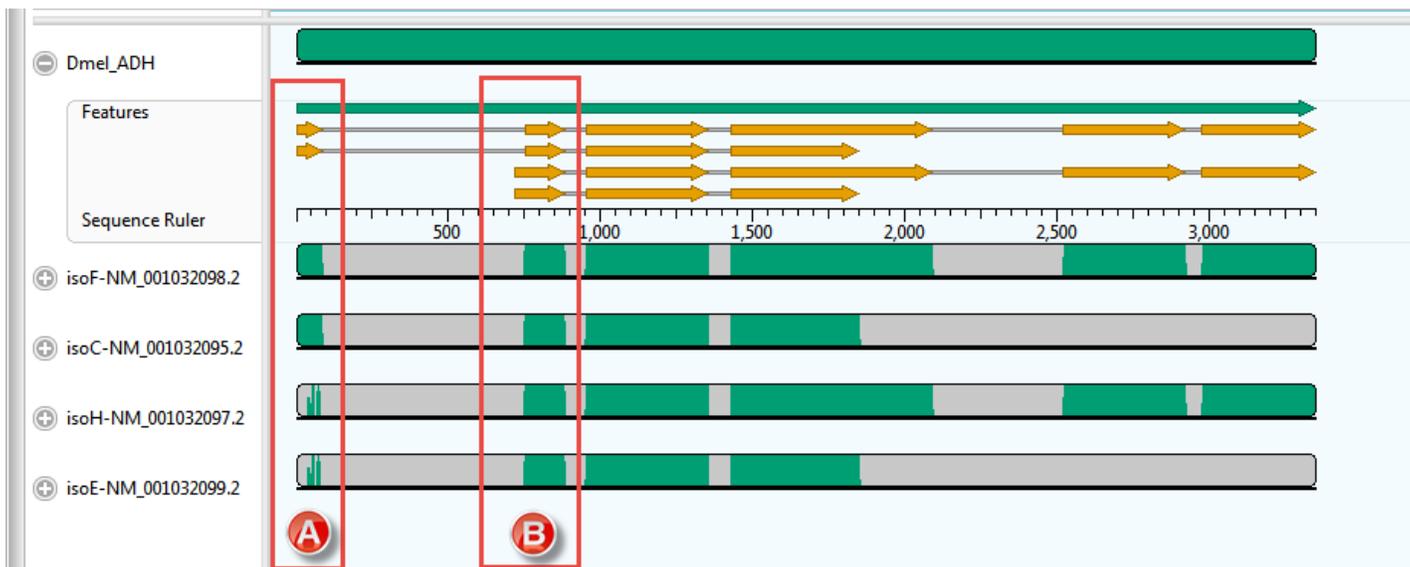
Tutorial 2: Aligning transcripts to genes

Part A: Comparing results from three multiple alignment methods

1. Extract the contents of the *MegAlignProAlignments.zip* folder and double-click on *DmADH.muscle.msa* to launch it in MegAlign Pro.

This document contains a default MUSCLE alignment of the *Drosophila melanogaster* ADH gene and the mRNA transcripts for isoforms F,C,E and H. For this example, some of the annotations of the original GenBank entry, NT_033779.5, were removed for the sake of clarity.

The Overview for this alignment is shown below.

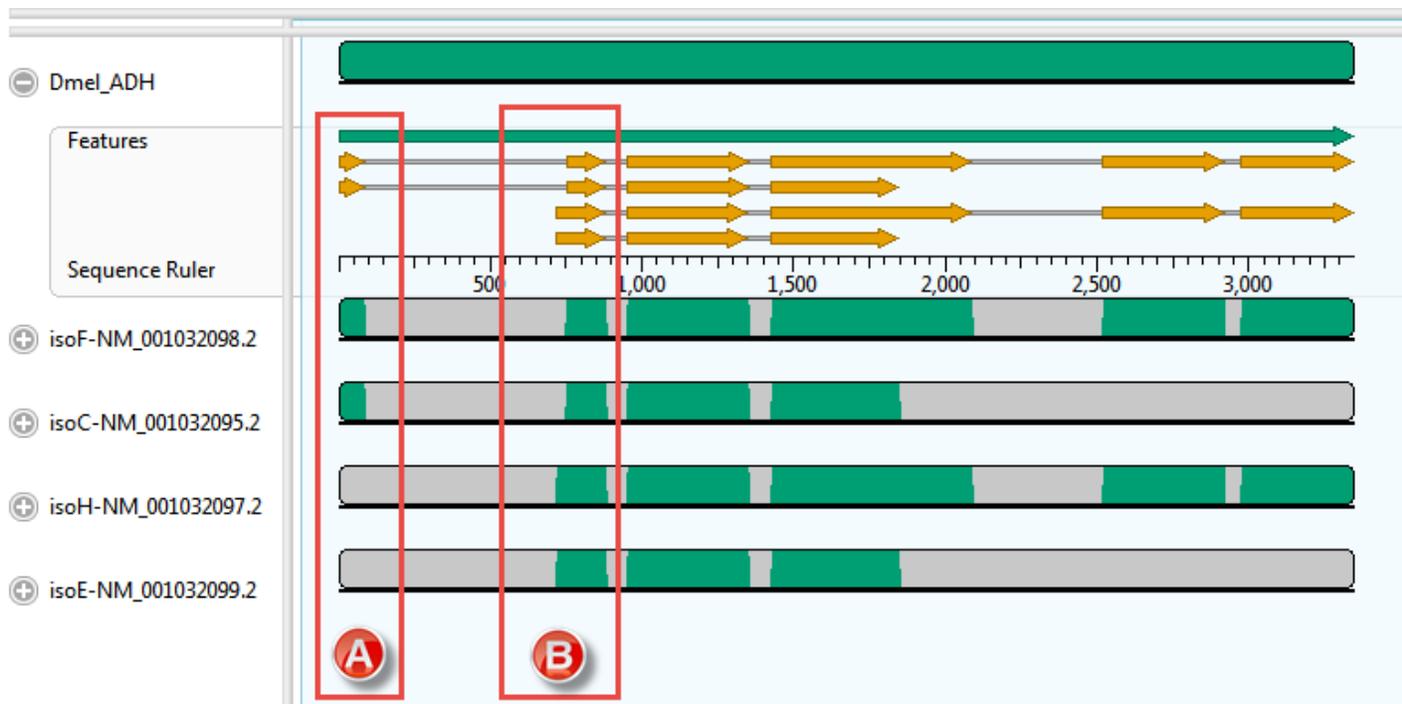


For this example, the transcripts, represented by the four lower sequences, have already been arranged in the same vertical order as the mRNA-features for the ADH gene, represented by orange arrows in the upper sequence. Hover your mouse over some of the orange mRNA features and verify that they are in the same order as the transcripts represented by the lower four sequences.

There is, as expected, a good correspondence of the aligned regions of the transcripts to the exons of the gene. But a closer examination reveals a few problems, one of which is indicated by the red boxes in the figure above. Notice that that isoF and isoC are annotated to code for Exon I (box A), which matches the multiple alignment result. However the 5' end of isoH and isoE are also aligned, somewhat poorly, to this exon. How can we tell? Note that there are only two yellow arrows in box A, but there are green bars in all four transcript rows for box A, the bottom two of which are much sparser than the others. This part of the lower two transcripts—which have no corresponding yellow arrows—should really have aligned with the start of Exon II (box B).

As a reminder, the example above used the MUSCLE alignment method. Before we proceed to see how pairwise alignment can resolve this problem, let's spend a few moments exploring how some of the other three multiple alignment engines handle these particular sequences.

2. Multiply realign the sequences using Clustal Omega (**Align > Realign Using Clustal Omega**)
Examine the alignments of transcript sequences to the first two exons of the gene's sequence. In this case, there are no green bars at all in box A.
3. Realign again, this time using MAFFT (**Align > Realign Using MAFFT**). For this example data set only MAFFT gives a result that is consistent with the annotations.



Bear in mind, that you may not always have annotations to rely on for validating multiple alignments, so it's often a good idea to use pairwise alignment in this situation.

Part B: Using a local pairwise alignment method

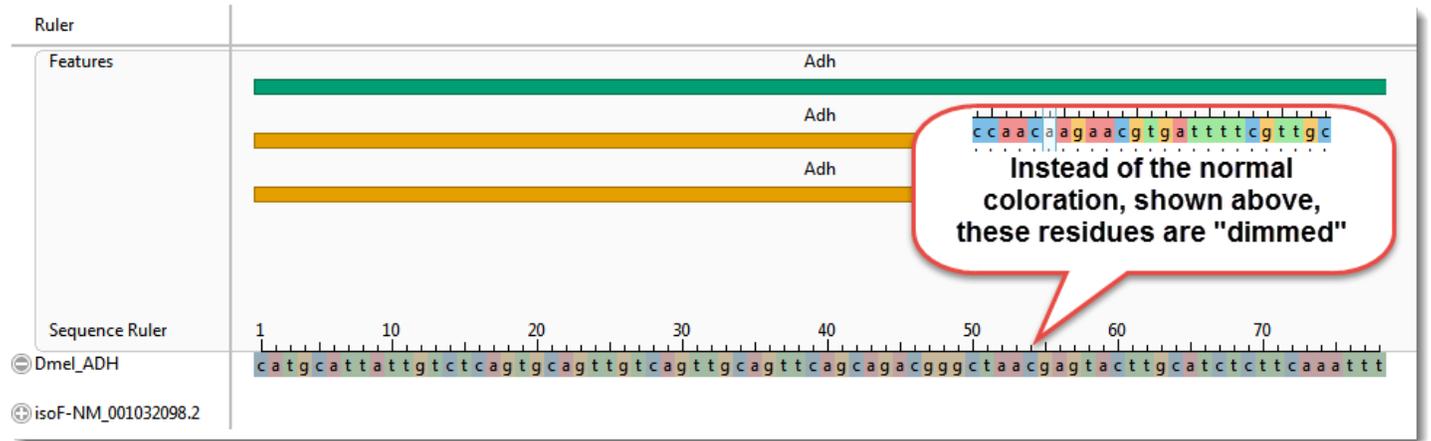
In this part, we will use a local pairwise alignment to resolve the correct mapping of the first intron in isoF and isoC.

1. Click on Pairwise view tab. Look at the left drop-down menu at the top of the view to see that the uppermost sequence (Dmel_ADH) has automatically been selected as the *reference*.
2. In the right drop-down menu, choose "isoF-NM_001032098.2". The default *Local: Smith Waterman* method is used to create the alignment

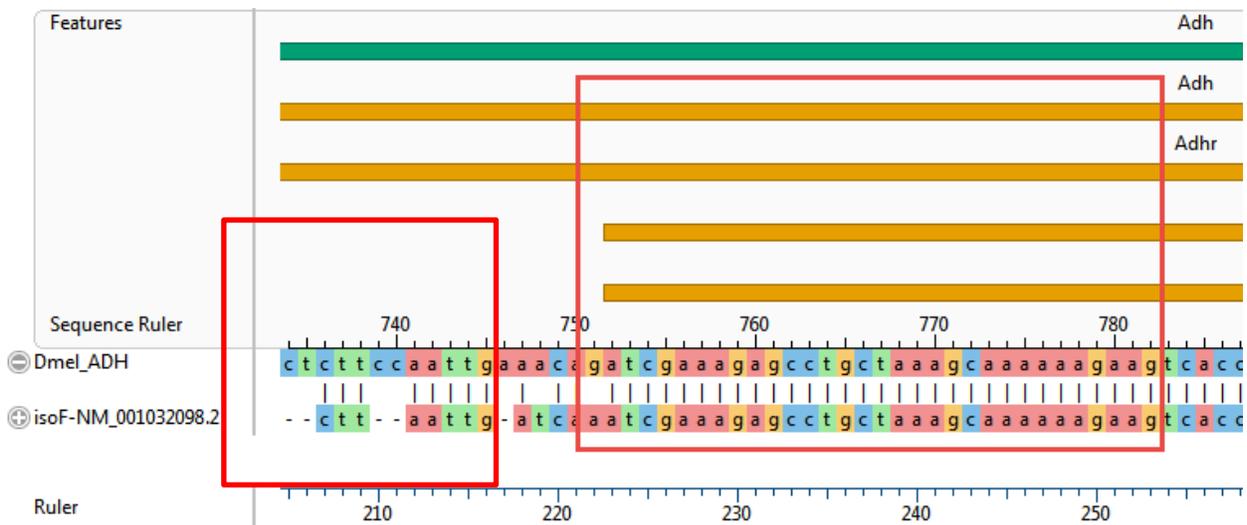
To make sure all features will be displayed, open the **Tracks** side-panel. Under "Pairwise details," click once on the word "Features." In the Options area below, change the Features "Rows" from 3 to 5.

3. In the Pairwise view, click on the plus-sign icon next to the name "Dmel_ADH" to reveal the Features and Sequence Ruler tracks.

A correct alignment would show the 5' end of the isoF sequence aligned with the two orange arrows that begin at the very start of the Dmel_ADH sequence. Instead, however, the beginnings of both the ADH reference sequence and the mRNA are shown as dimly colored *context*; in other words, as unaligned flanking sequence.



Scrolling down through the alignment observe that the 5' end of the transcript sequence has aligned, albeit poorly, with a region beginning within the first intron. Meanwhile, a solid ungapped alignment doesn't start until 5' end of the second exon annotated for either isoF or isoC (Fig 2B, indicated with box).



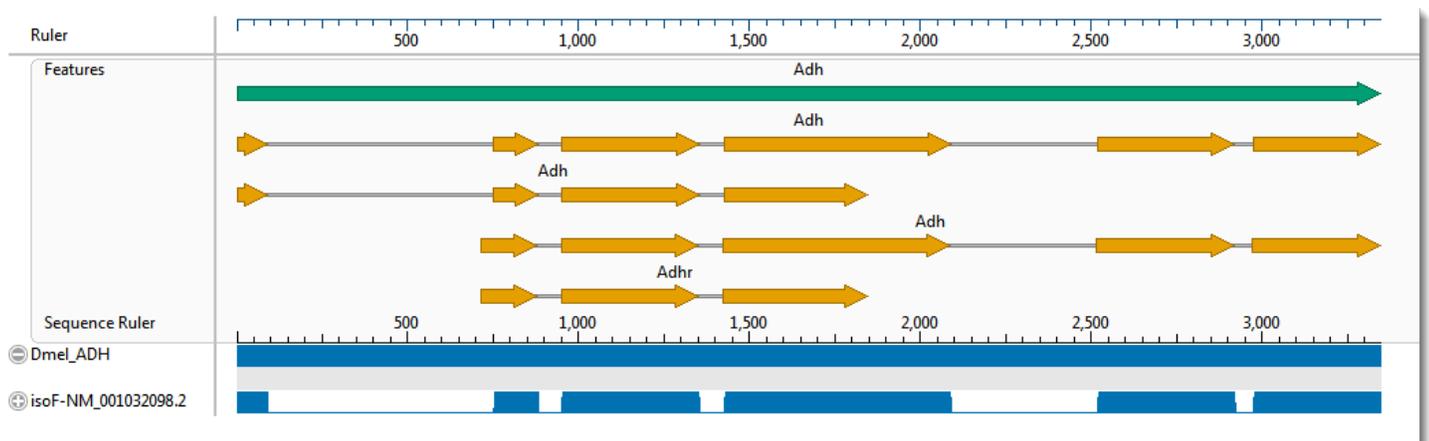
Clearly, a local pairwise alignment was not an improvement over the first two multiple alignments tried in Part A.

Most likely, introducing a sufficient number of gaps to span the intron would reduce the score so much that the aligner can find a higher scoring segment that contains numerous short gapped regions.

Part C: Using a global pairwise alignment method

Let's try again, this time using global alignment. As you may recall, this type of alignment forces the ends of both sequences to be aligned.

1. Click the gear icon on the Pairwise view toolbar and select **Global: Needleman-Wunsch** from the "Using" drop-down menu. Then press **Align**.
2. After aligning, the display options that you set above remain in effect, but the sequences' detail tracks will collapse. Expand the Dmel_ADH track again by clicking on its plus sign.
3. With long alignments like this one, it's easier to see the "big picture" if you zoom all the way out using the green "zoom slider" above the view. Do that now. The pairwise alignment now appears similar to the Overview. You can see that the global alignment using isoF matches Dmel_ADH's annotations.



MegAlign Pro allows you to open additional pairwise views so that you can compare alignments with different settings, or even different pairs of sequences. We can use this to compare any of the other three mRNA sequences to Dmel_ADH, while leaving the previous alignment untouched. An easy way to do this is to use the "Clone this view" tool, which is situated at the far right side of the toolbar. The two views will start out the same, but they are completely independent.

4. Clone the current Pairwise view.
5. Now select isoE-NM001032099.2 as the query sequence, keeping the reference sequence set to Dmel_ADH. Does this new Global pairwise alignment give the expected results? (Yes).