Ancestral Reconstruction Module-Guide

Introduction

Ancestral reconstruction allows us to trace back the evolutionary history of genes to see what the previous DNA sequence looked like and what sequence changes occurred over time to give rise to the gene we see today. In order to do this you need to compare the gene's sequence to the homologous (i.e. comparable) sequence in other species. Ancestral reconstruction can tell us interesting things such as how certain mutations persist and how sequences develop different properties and structures over time.

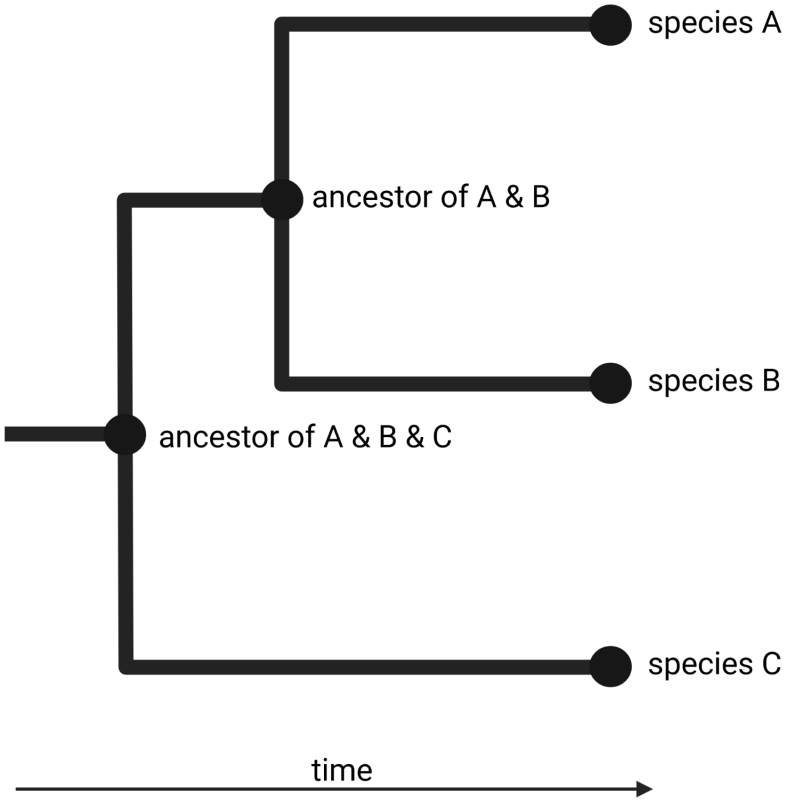


Figure 1: Phylogenetic tree for three hypothetical species A, B and C. Phylogenetic trees show the common ancestry and divergence of species over time.

To see how the sequence of our proto-gene evolved over time we need to find the homologous DNA sequence to our proto-gene in other species (i.e. the DNA sequence that eventually became the proto-gene). If we were to use only the sequence of our proto-gene and compare it across species (and blasted it across species) we probably wouldn't find anything useful, because the sequence did not necessarily look like this before. Instead of using the proto-gene sequence to identify homologous sequences, we use the neighboring genes. For example, looking at Figure 2, to find the homologous DNA to this proto-gene (yellow box), we would find the genes homologous to genes 1 and genes 2 in the other species. If we find that the orientation of gene 1 and gene 2 homologous in other species is the same, we can infer that the DNA region in between them is the homologous sequence to our proto-gene; i.e. the region we are looking for! Looking at the conservation of gene order is called **synteny**.

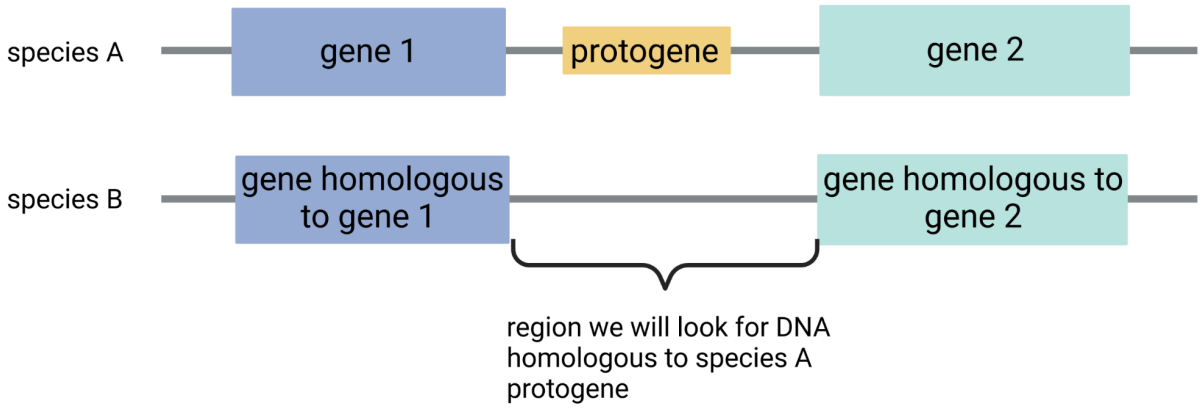


Figure 2: To find DNA sequence homologous to species A proto-gene in species B, we can use synteny. Since the position of gene 1 and gene 2 is the same in species B, we can assume the region in between the genes is the DNA homologous to the proto-gene in species A.

This region in between contains more than just the nucleotides making up the future proto-gene, and that is good because we want to be able to see how the whole region has changed over time to get a better estimate of how the sequence has changed and which sequences are more similar so having more letters gives us a better estimate. Next we will use a software called PRANK to align the sequences while keeping in mind the evolutionary relationships between the species using a guide tree (Figure 3). Guide trees are created by comparing the entire genome of species to determine how similar they are to one another. Guide trees are then used in multiple sequence alignments to help group more similar sequences based on their evolutionary relationships, making the alignment process more efficient and accurate.
We will then use the alignment from PRANK to view what the DNA looked like in each species as the proto-gene emerged.

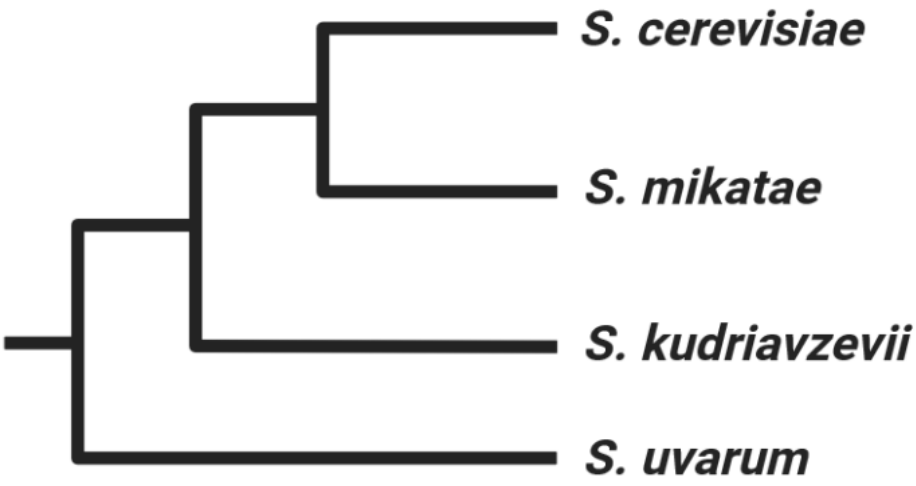


Figure 3: Guide tree for Saccharomyces species. This tree represents how the different *Saccharomyces* species are related to each other based on the grouping of the branches. Where, for example, *S. cerevisiae* is more closely related to *S. mikatae* than it is to *S. kudriavzevii*.

Goals: In this module, participants will learn

1) To use phylogeny to reconstruct the ancestry of a non-annotated ORF
2) Determine the evolutionary relationships between species for a specific chromosomal region.

Objectives: After completing this module, participants will be able to:

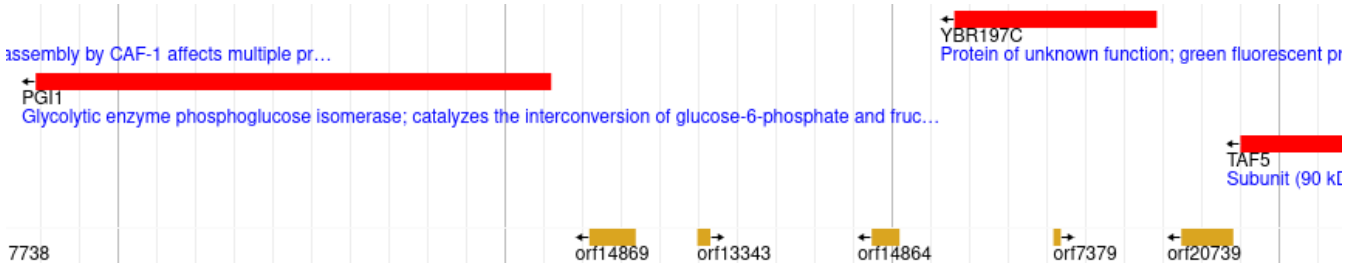
1) Generate a phylogenetic tree for an ORF sequence to other sequences in the *Saccharomyces* clade.
2) Align nucleotide sequences between species for a specific genomic region
3) Determine the percent nucleotide and amino acid sequence similarity of an *S. cerevisiae* proto-gene to homologous sequences from other closely related species.

This guide will walk through investigating an example proto-gene using a genome browser. Follow along and answer the questions in the accompanying worksheet for your proto-gene. During the module if you run into errors check out the ‘Troubleshooting’ section at the end.

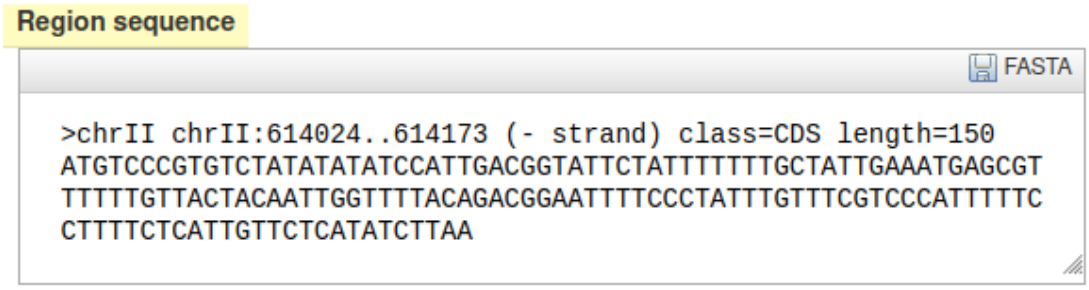
Activity

1. Find your proto-gene on Jbrowse. Go to the ORF information app: https://carvunislab.csb.pitt.edu/shiny/coexpression/
   Type the ORF ID for your proto-gene in the box, select ‘Sequence’ in the ‘Results type’ drop down menu and click on the ‘Genome Browser’ link on the results page.

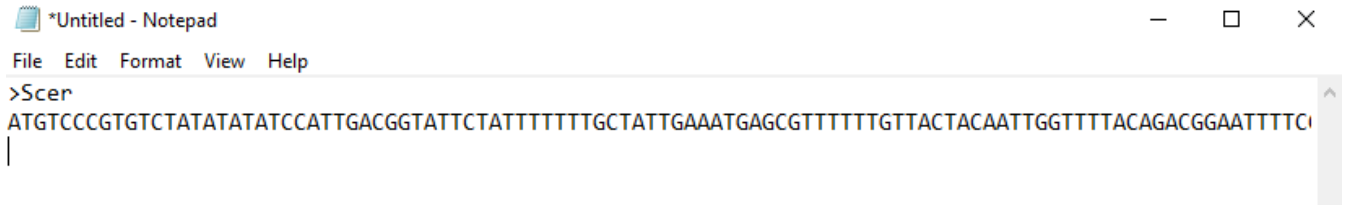
   In this example the proto-gene is orf14869.



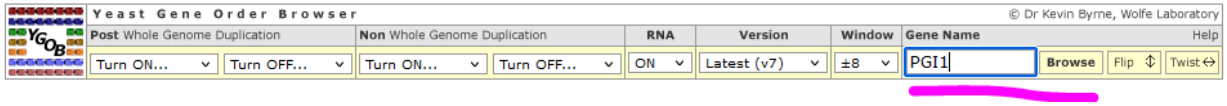
2. Identify adjacent annotated conserved genes. In this example, the genes PGI1, YBR197C and TAF5 are all adjacent conserved genes to the proto-gene orf14869. (YBR197C is a gene of unknown function, but is a verified ORF, and thus can be used as a conserved gene).
3. Click on your proto-gene in the Genome Browser and copy its DNA sequence. (or use the ORF informaton app https://carvunislab.csb.pitt.edu/shiny/coexpression/)



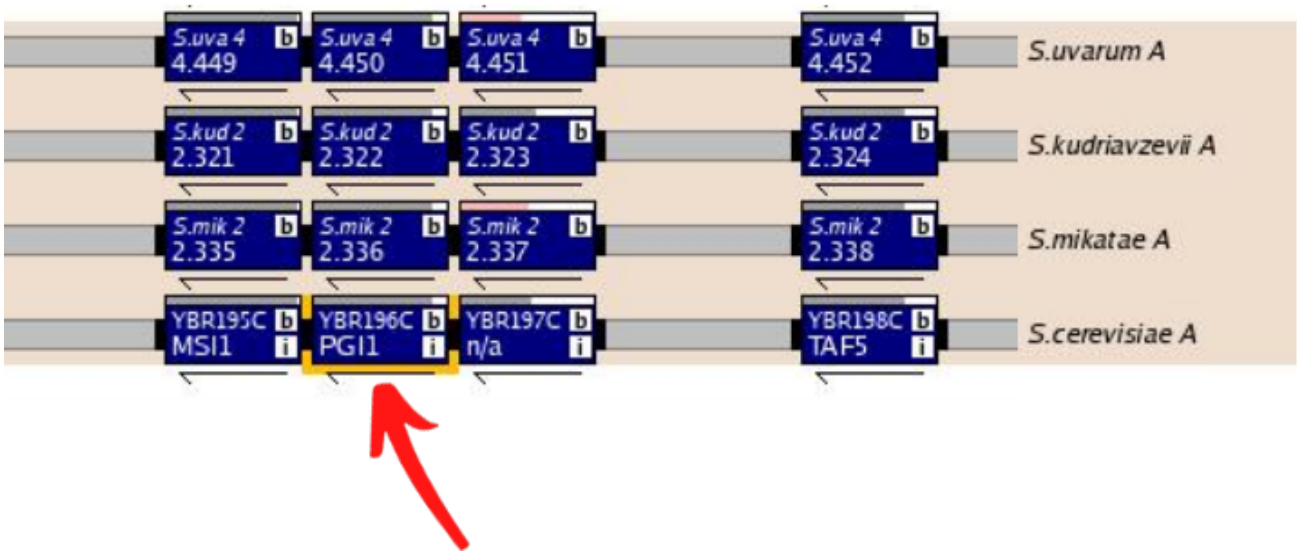
4. Next we need to make a file that contains the DNA sequence of our proto-gene and the homologous regions in other species. Open any text editor (check the troubleshooting section for more details). In the first line type '>Scer' then hit enter and paste the DNA sequence of your proto-gene. (this is FASTA nomenclature, a file format used for sequence data in bioinformatics)



5. Now we need to get the DNA sequences for the other species. Go to the Yeast Gene Order Browser http://ygob.ucd.ie/. At the bottom of the screen there is a toolbar. In the toolbar select "TURN ALL ON" from the "Turn ON…" drop down menu. Next type in the name of one of the adjacent conserved genes in the "Gene Name" box at the bottom of the page and click "Browse".

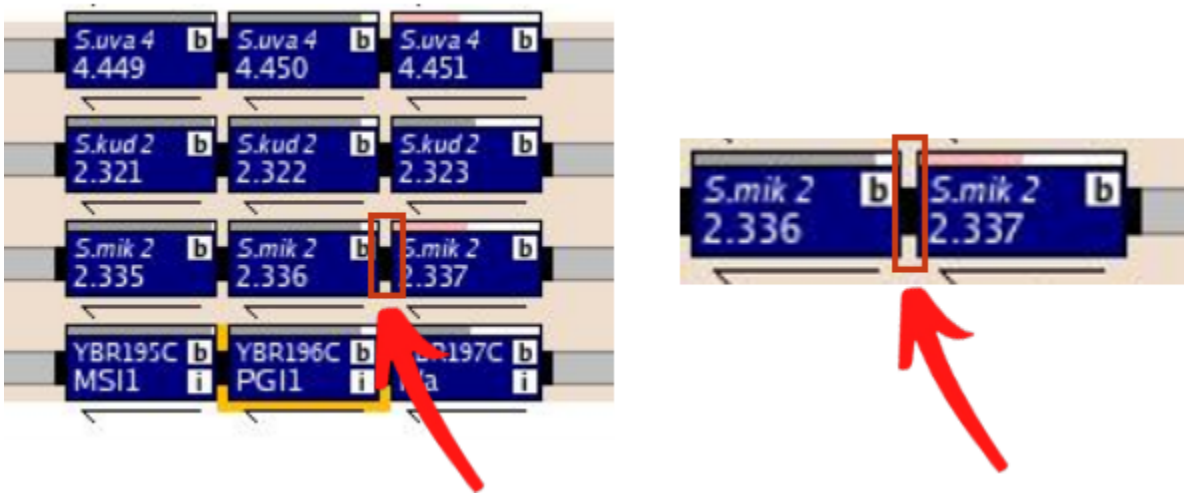


6. The resulting page shows the homologous genes that have conserved position and order across all the different yeast species. Each row represents a species and each box represents a gene, underneath each gene is an arrow showing which strand the gene is located on. Look for the row that says *S. cerevisiae*. The gene you searched for should be highlighted in yellow. Below is a zoomed in example locus, PGI1.



Now we want to see if the genes surrounding the region where our proto-gene is located are the same in the three other species (same meaning that the order and direction of the genes are the same). In this example we see that the neighboring genes PGI1 and YBR197C have homologous genes in the same order and direction in the other species. If synteny at the locus of interest is only present in some species refer to the Troubleshooting section of this document for help.

7. Now click on the line in between the two upstream and downstream neighboring genes to your proto-gene to get the *S. mikatae* DNA sequence.

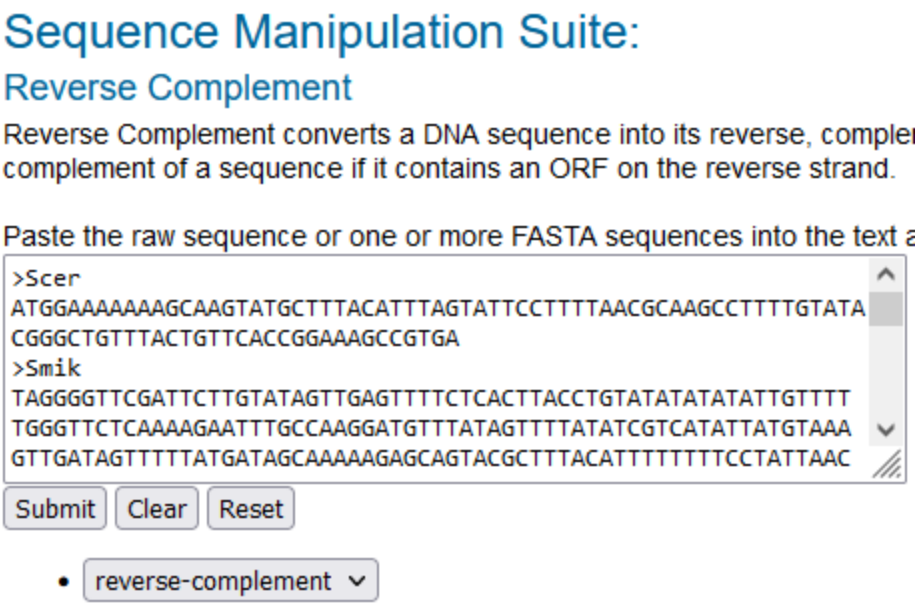


8. This will bring up a new window with a DNA sequence. Copy the DNA sequence only (i.e. not the header) and go back to your text editor and add a new line with '>Smik' hit enter and then paste the DNA sequence. It is important for later steps that each sequence is named using the first letter of the genus and then species abbreviation (ie: Scer, Smik, Skud, Suva).

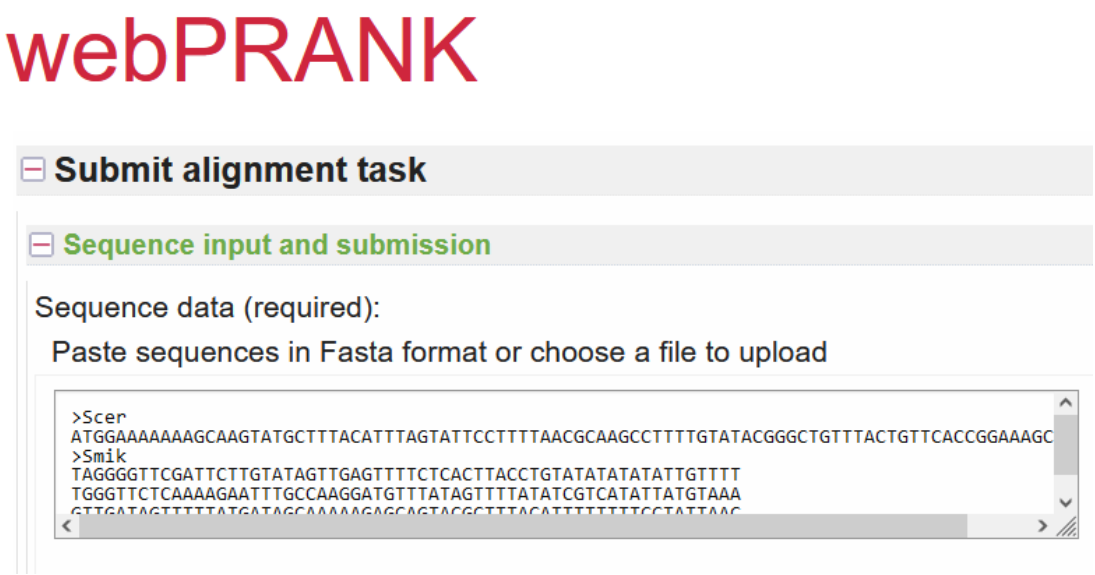
9. Follow steps 7 and 8 for the other two species *S. kud* and *S. uva*.

10. If your proto-gene is located on the negative strand (-), you will need to reverse complement the DNA sequences of the other species (ie Suva, Smik, Skud). You do not need to reverse complement the DNA sequence of your proto-gene in Scer. If your proto-gene is on the forward strand (+) you can skip this step entirely.

    Go to the following website: https://www.bioinformatics.org/sms2/rev_comp.html
    Paste the DNA sequences for Suva, Smik and Skud into the text box, make sure 'reverse-complement' is selected and click submit. Replace the sequences in your text file with these new reverse-complemented sequences. If you copy the entire output, be sure to update the headers for each sequence so that they say '>Smik' only and not '>Smik reverse complement'.

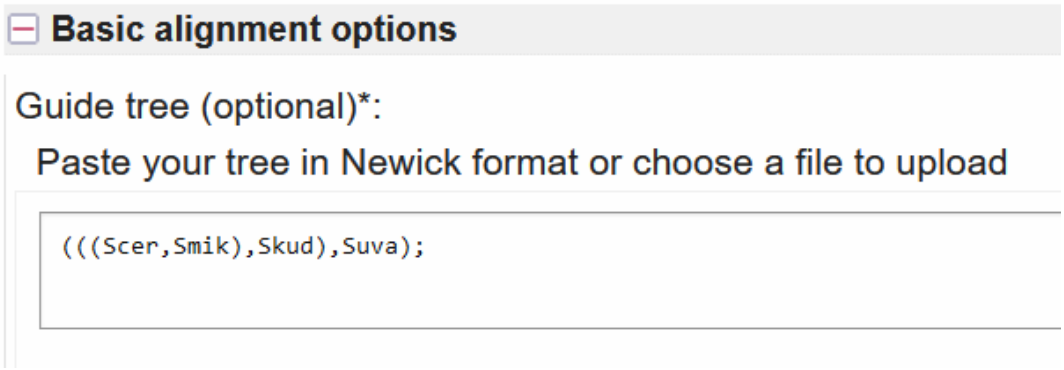


11. We are now ready to do ancestral reconstruction. Go to the https://www.ebi.ac.uk/goldman-srv/webprank/ website. Either upload your sequence file or paste it into the text box.



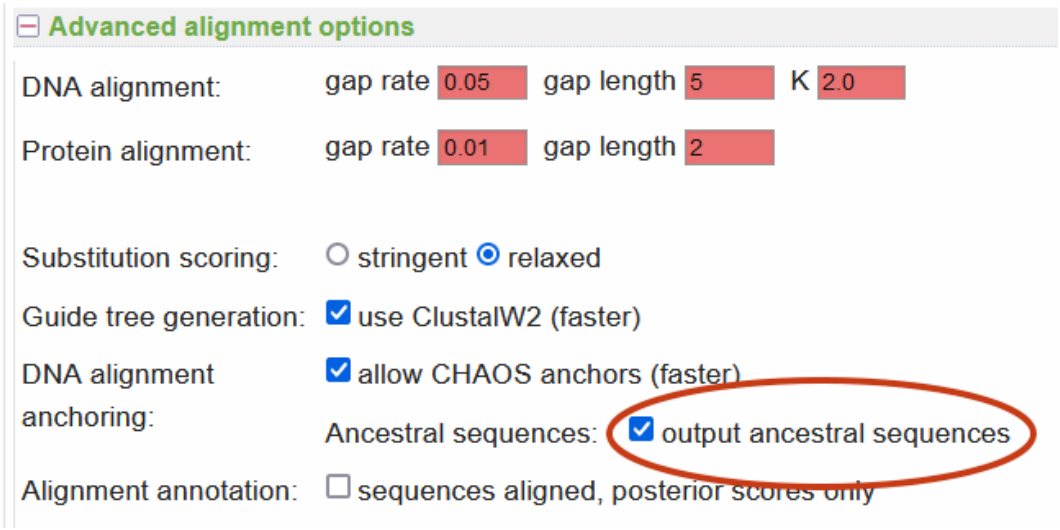
12. Scroll down and expand the ‘Basic alignment options’ menu. In the Guide tree section paste the following Saccharomyces species tree into the box.

    (((Scer,Smik),Skud),Suva);



This tree, represented in Figure 3, shows the relationship between species based on their entire genome sequences. The tree helps the PRANK algorithm to determine how the nucleotide sequences you are inputting are related to each other to make the alignment more accurate and efficient. The above text is a way to represent this tree.

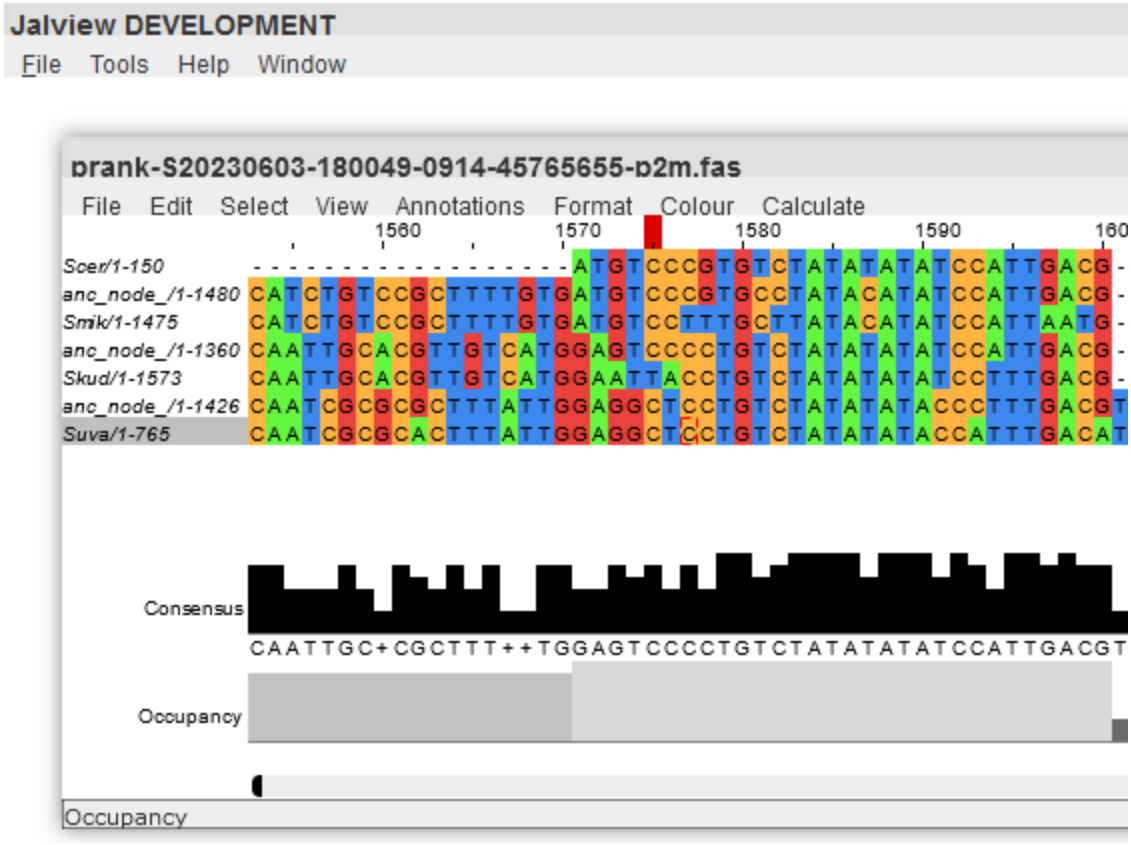
13. Scroll down to the ‘Advanced alignment options’ and check the ‘output ancestral sequences’ box



14. Now scroll back up and click ‘Start alignment’. If you get an error message try refreshing the page. When the reconstruction is complete you will see a window like this. Select ‘FASTA ancestral’ from the drop down menu and click the ‘download’ button’. This will save a .fas file onto your computer.

15. Go to https://www.jalview.org/jalview-js/JalviewJS.shtml/
    Click on File > Input Alignment > From File. Upload (click on load) your .fas alignment file from webPrank. (Fyi this website can be a little slow)

16. A new window will open with the alignments. If you would like to change the color of the letters click ‘Colour’ then select ‘Nucleotide’.



Each row corresponds to a species or ancestor in the same order as the species tree you provided with the ancestors in between each extant species, as shown in figure 4.

Figure 4: *left*: Representation of guide tree used in the sequence alignment process. *right*: Each row in the alignment corresponds to either an extant species or the ancestral species. The rows in the alignment are ordered in the same way as the guide tree with the ancestral sequences in between each extant sequence. Looking at this alignment, specifically at the start codon (ATG) we see that ancestor 1 and *S. mikatae* have a start codon in the same place as the proto-gene in *S. cerevisiae.* In ancestor 2 and *S. kudriavzevii* there is a start codon present but it is shifted over 3 positions to the left (and therefore is in the same frame as the ORF in S.cer). In ancestor 3 and

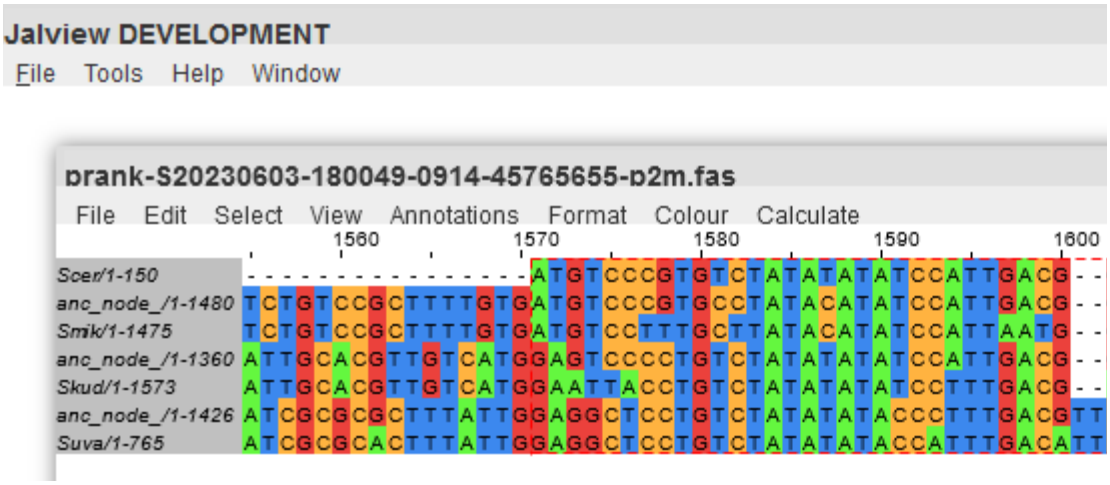
S. uvarum there is no start codon. Therefore by looking at the start codon, we would say the proto-gene first arose in ancestor 2.

Scroll horizontally until you see the sequence of your proto-gene in the 'Scer' row. (i.e. until you see the letters 'ATG' in the top row instead of '- - -').

Answer question 1 in the worksheet: Describe what you see in the DNA sequence at the start of your proto-gene in the other species/ancestors. Do you see the start of an ORF in the other species (i.e an ATG codon)? If there is one present, is it the same or different frame as the ORF in *S. cerevisiae*?

Answer question 2 in the worksheet: In what species or ancestor did the ORF first appear? (i.e. what species or ancestor furthest from *S. cerevisiae* can you find the first ATG in the same frame as the proto-gene in *S. cerevisiae*). Use the tree in Figure 4 for reference.

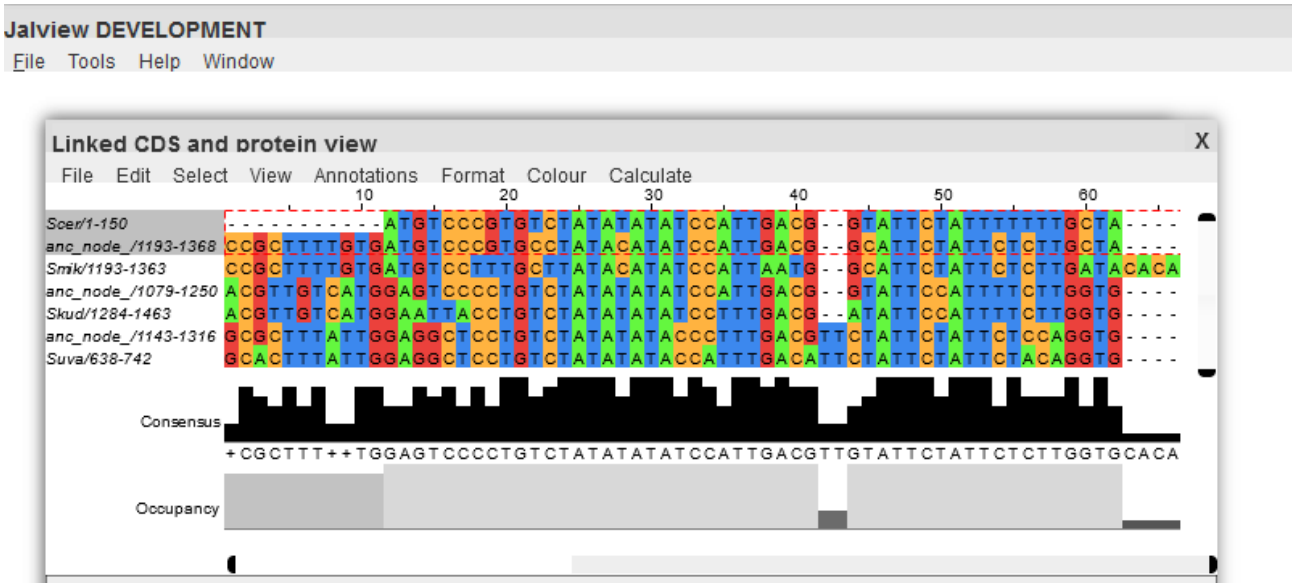
17. Next we want to compare how similar the DNA and amino acid sequences are in each species and ancestor to your proto-gene in *S cerevisiae*. To begin we will subset the alignment to contain the sequences corresponding to your proto-gene. Using your cursor, drag and select the sequence of your proto-gene and the homologous region in the other species and ancestors. I.e. begin selecting when you no longer see only dashes '-' in the top row and see the nucleotides ATG, and continue until you see a stop codon (TAA, TAG, or TGA). Then drag vertically to select all rows.



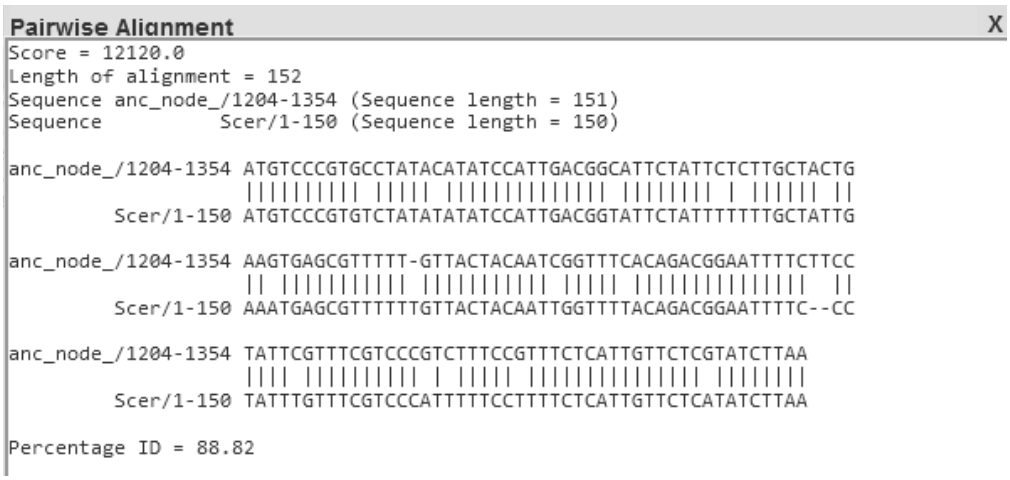
For example, in the above alignment, you would start selecting the sequences at column 1571 where the proto-gene begins and drag horizontally till the end of the proto-gene sequence. There may be a lot of gaps, so be sure to scroll to the end of the Scer sequence (i.e. confirm there is a stop codon TAA, TAG or TGA).

18. Then click 'Calculate' > 'Translate as cDNA' > '1. Standard'. A new window will pop up called 'Linked CDS and protein view'. To see how similar the ancestral nucleotide sequence for *S cerevisiae* and S. *mikatae* is to your proto-gene, Click on the 'Scer' label on the left of the sequences and then while holding down the 'ctrl' key on your keyboard click the ancestral label for the ancestral row*. The two sequences you want to compare should now be highlighted.

*If you are using a mac you will not be able to select the species. You can skip selecting the rows and go to step 19.



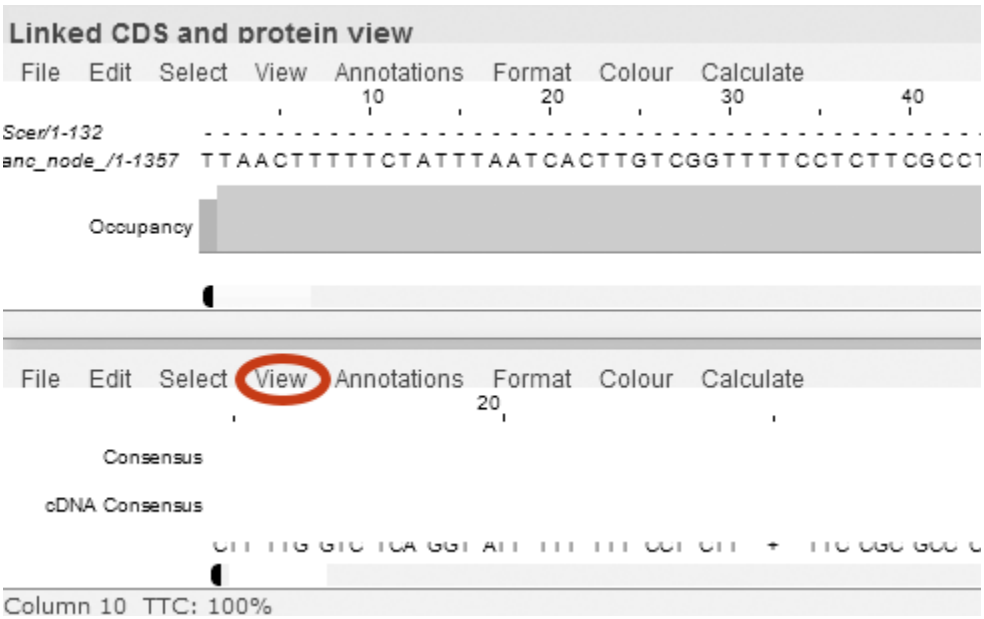
19. To calculate how similar these two nucleotide sequences are click ‘Calculate’ > ‘Pairwise Alignment’. A new window will appear showing the pairwise alignment between these two sequences as well as a ‘percentage ID’, which is how similar the two sequences are in terms of how many nucleotides align (indicated by a ‘|’).



Answer question 3 on the worksheet: Using the region homologous to your proto-gene, how similar are the DNA sequences between your proto-gene sequence in *S. cerevisiae* compared to the ancestors and other species? Fill in the percentage ID in the table.

Answer question 4 on the worksheet: Does the DNA sequence similarity between *S. cerevisiae* and the other species/ancestors decrease with evolutionary distance (ie does DNA sequence similarity decrease as you get further from *S. cerevisiae)*?

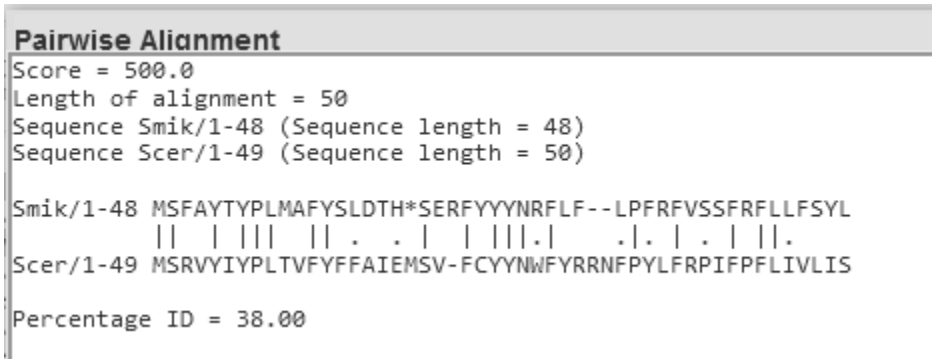
20. To view the protein sequence of your proto-gene, go back to the 'Linked CDS and protein view' window. In the bottom toolbar, click on the 'View' button (circled in red in the diagram below) and then unselect 'Nucleotide'.



You can play around with the colors for the alignment using the 'Colour' menu.
Scroll through the protein sequences. Do you see any species or ancestors that contain an ORF? ORFs are sequences that begins with a start codon (denoted as M) and ends with a stop codon (denoted as *).

If there are lots of gaps in your sequence, you remove the gaps by going to ‘Edit’ > ‘Remove all gaps’.

21. Next calculate how similar each protein sequence is to your proto-gene. For example, to see how similar the protein sequence for *S cerevisiae* and *S. Mikatae*, click on the Scer row and while holding down the ‘ctrl’ key on your keyboard, click on the Smik row. Then click ‘Calculate’ > ‘Pairwise Alignment’.



Answer question 5 on the worksheet: Using the region homologous to your proto-gene, how similar are the amino acid sequences between your proto-gene sequence in *S. cerevisiae* compared to the ancestors and other species? Fill in the percentage ID in the table.

Answer question 6 on the worksheet: Does the amino acid sequence similarity between *S. cerevisiae* and the other species/ancestors decrease with evolutionary distance? (i.e. does amino acid sequence similarity decrease as you get further from *S. cerevisiae*)

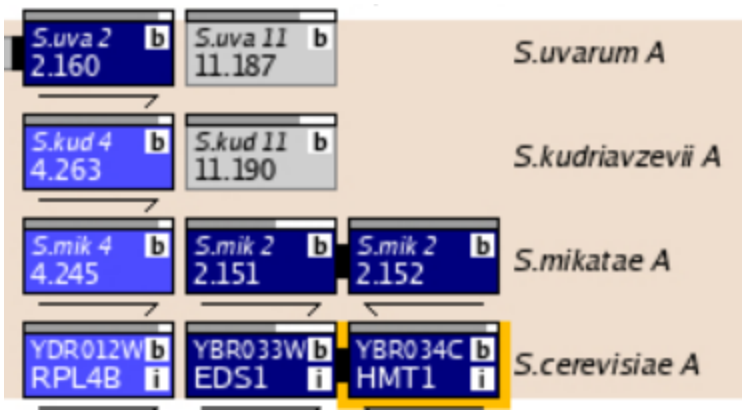
Troubleshooting

- **What is a plain text editor?**
  A plain text editor is a software that allows you to type without any auto-correcting or auto-formatting, this is useful for creating FASTA files (or other bioinformatic files) that require specific formatting. Examples of plain text editors that are already included on computers include Notepad on windows and TextEdit on mac. For TextEdit, turn plain text mode on by going to the 'Format' menu and click 'Make Plain Text'.

- **The Yeast Gene Order Browser says the gene I searched for is 'turned off'**
  The Yeast Gene Order Browser only includes certain annotated genes. If one of your adjacent genes is not found, try searching for the neighboring gene on the other side, or try one a little further away. You can also try searching by the gene's systematic name and common name (i.e. for example PGI1 is the common name and YBR196C is the systematic name).

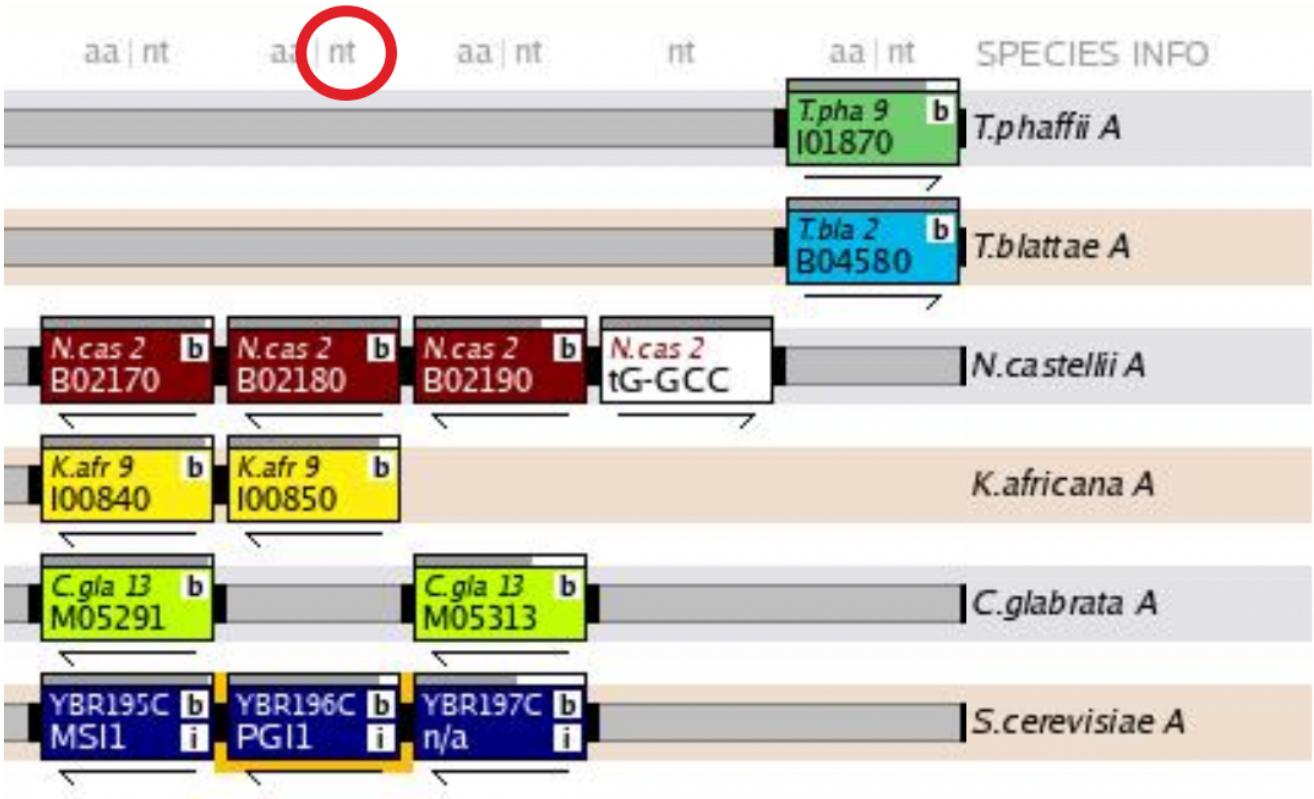
- **Having trouble with clicking to get intergenic sequence**
  If you are having trouble clicking on the intergenic sequence between the two genes adjacent to your proto-gene because the region is too small, try zooming in on your web browser.

- **What if synteny is not preserved in all species?**
  If the adjacent genes to your proto-gene are not found in the same position across all the species, you can instead use only the species where they are. In this case you will need to update the guide tree used for PRANK to include only the species you are using, or just not use a guide tree at all.



Above is an example where synteny is not preserved across all the four species: HMT1 and EDS1 locus in *S. cerevisiae* is syntenic with the locus in *S. mikatae* but not with *S. kudriavzevii* and *S. uvarum*. If your proto-gene was located between EDS1 and HMT1, you would only use the *S. mikatae* sequence and not use a guide tree when doing ancestral reconstruction in PRANK. If your locus in *S. cerevisiae* is only syntenic with *S. mikatae* and *S. kudriavzevii* but not *S. uvarum*, you can use the sequences for *S. mik* and *S. kud* and update the guide tree to be: ((Scer,Smik),Skud);

- **What if my proto-gene is not within the intervening sequences but instead is overlapping a gene?**
  If your proto-gene is fully overlapping a gene on the opposite strand, click on the ‘nt’ button at the top of the Yeast Gene Order Browser screen above the gene box. This will open a new window that has the DNA sequences for that gene in all the species. You can then go through and copy the sequences you are interested in (ie S. uva, S. kud, and S. mik for this tutorial). You will need to account for what strand the gene is located on, such that if the gene is on the positive strand and your proto-gene is on the minus strand you will need to reverse complement the gene sequences, and if the gene is on the negative strand and your proto-gene is on the positive strand you will also need to reverse complement the gene sequences.



If your proto-gene is partially overlapping a gene, you will need to get the intergenic sequence and the gene sequence that covers your proto-gene and concatenate them. Again you will need to take into account the direction of the sequences, where the intergenic sequences are always in the forward direction, however the sequences of the genes can be in the forward or reverse direction depending which strand the gene is on.

- **Error: Names in sequence data and guide tree don't match.**
  Check to make sure the species labels in your sequence file match the species label in your guide tree string, and avoid using spaces in the species labels.

  For example if your sequence file looks like:
  >Scer
  ATAT…ATTAG
  >Skud
  ATGG…ATTAA
  ….
  >Suva
  ATGA…..TTAATA

  Then your guide tree labels should be (((Scer,Smik),Skud),Suva);

- **I only see the sequences for the extant species and none for the ancestors in the alignment.**

On the PRANK result output page, make sure you select the ‘FASTA ancestral’ option in the ‘Download the results’ dropdown menu.

- **Jalview is not loading or keeps crashing**
  Jalview is a widely used alignment tool with many great functionalities. Recently, they developed a web based version called JalviewJS (the one used in this tutorial) however it is still under development/testing and there may be temporary outages. In this case, you can install the desktop version to use. https://www.jalview.org/download/

Sources

- Byrne KP and Wolfe KH. (2005). The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species, *Genome Research*
- Löytynoja A and Goldman N. (2010). webPRANK: a phylogeny-aware multiple sequence aligner with interactive alignment browser, *BMC Bioinformatics*
- Diagrams made using biorender.com