AlignmentViewer

Working together Jan and Lola used the OMG and part of the ATP gigantic pipelines at the Salk to identified 250 genes to use for building a phylogenetic tree of the 133 species. Overviews of each pipeline is on gigantic.fish.

For OMG: Each gene was identified in each genome. However, in some cases a given gene was absent in a given genome. This absence might be due to loss of the gene in evolution of the species - or the gene might be present in the species but absent in the current version of the genome. A total of nearly 1000 genes were identified as useful and protein sequences for each gene in each species were handed off to Lola. Identification of these genes was done by Jan.

For ATP: To start, Lola reduced the number of sequences to 250 so that things would not take more than a few hours for alignment and for tree building. Next, for the 250 sequences: Protein sequences representing each gene in a given species were joined together into a single super-long sequence per species - around 12,000 amino acids in length per species. These superlong sequences are what you downloaded from gigantic.fish and used to build the alignment in MAAFT. Reducing the number of genes to 250 per specis and building the superlong sequences was done by Lola.

The alignment you uploaded to AlignmentViewer is a random section out of the Species-133 MAFFT alignment made on CIPRES by Lola.

On AlignmentViewer:

1. In msa view: Set Sequence order to **identity 1**. Scroll to the down to see the different species. For the last few species, scroll along the length of the alignment. Why is it some species are missing sequence in sections of the alignment. There are several possible explanations.

There are 3 kinds of changes that commonly occur in the evolution of gene/protein sequences over time. For a protein that is 100 amino acids in length, we can imagine: 1) Transformation at a given position in the protein going from one amino acid to different amino acid (in this case, the sequence has changed but length is still 100), 2) loss of 1 (or more) amino acid at that position (in this case, amino acids are the same except one is missing and length is now 99), or 3) insertion of 1 (or more) amino acids at that position (in this case, amino acids are the same plus there is new sequence and length is 101).

2. Looking at the alignment vs gap and conservation bar char vs sequence logo - can you identify changes in the sequence, deletions in the sequences, and insertions in the sequences?

3. You may have noticed, its really hard / impossible to distinguish insertions from deletions. To do this accurately, we really need to use a phylogenetic tree. Why is this? These question may be challenging but hopefully by the end of the lab in 3 weeks the

answers will be more apparent.

It's worth noting that because it's not possible to separate insertions from deletion in an alignment, they are referred to as **indels**.