CS4742 - Bioinformatics

Assignment - Phylogenetic Trees

February 18, 2021

Phylogenetic trees may be constructed using data from a signal gene across different species to make inference about the evolution of that gene. However in genomics, data from many genes are used to infer the evolutionary history of the species (by constructing a species tree). Unfortunately, trees constructed for different genes may conflict with each other. This is referred to as Between-genes phylogenetic incongruence and causes challenges in constructing a species tree.

1 Procedure to follow

In this assignment, you are given several bacterial proteins from a family of bacteria named **Acetobacteraceae**. There are 21 different bacterial species that belong to this family as listed below.

Accession ID	**Name of the species**
NZ_CP014692.1	Acetobacter aceti
NZ_CP023657.1	Acetobacter pomorum
NZ_CP022699.1	Acetobacter tropicalis
NZ_CP014687.1	Acetobacter persici
NZ_LN606600.1	Acetobacter senegalensis
NZ_CP011120.1	Acetobacter oryzifermentans
NZ_CP015164.1	Acetobacter ascendens
NZ_CP015168.1	Acetobacter ascendens
NZ_CP021524.1	Acetobacter ascendens
NZ_CP022374.1	Acetobacter oryzifermentans
NZ_P018515.1	Acetobacter orientalis
NZ_CP023189.1	Acetobacter pomorum
NC_017100.1	Acetobacter pasteurianus
NC_017121.1	Acetobacter pasteurianus
NC_017125.1	Acetobacter pasteurianus
NC_017146.1	Acetobacter pasteurianus
NZ_LN609302.1	Acetobacter ghanensis
NC_017111.1	Acetobacter pasteurianus
NC_017150.1	Acetobacter ghanensis
NC_017108.1	Acetobacter pasteurianus
NZ_AP014881.1	Acetobacter pasteurianus

Let's call the set of proteins given to you as the $protein\_set$. The $protein\_set$ assigned to your group can be found in the file "*Supporting_Documents\protein_sets_for_groups.pdf*"

For the proteins in $protein\_set$, you are required to carryout the following steps in order to build phylogenetic trees to infer their evolution among the species of the Acetobacteraceae family.

1. Out of the above 21 bacterial species, find which ones have all the proteins in $protein\_set$ present in their genomes. Let's call this set of bacteria the $common\_bacteria\_set$. You can use the Excel file called the "*Supporting_Documents\protein_tables.xlsx*", which is given to you, to find this $common\_bacteria\_set$. Each sheet in this excel file represents one of the bacterial species belonging to the Acetobacteraceae family mentioned above and lists the proteins present in them. The sheet name is the accession ID of the species. The protein names are given in the column "*Protein name*".

   E.g., if $protein\_set = \{FUSCfamilyprotein, MFStransporter, YraNfamilyprotein\}$ then,

   $common\_bacteria\_set$= {All 21 Acetobacteraceae bacterial species above}– $\{NZ\_CP015168.1\}$

   because "*FUSC family protein*" and "*MFS transporter*" are present in all of the above bacterial species and "*YraN family protein*" is present in all except NZ_CP015168.1 (Acetobacter ascendens).

   - **Note:** If a protein name in the protein table of a species contains the entire name of the protein you are looking for, then that protein is considered present in the genome of that species.
   - **Note:** The proteins given to you are likely to be present on all 21 Acetobacteraceae bacterial species above.

2. Retrieve the genomes of all the bacterial species in $common\_bacteria\_set$. I.e., for each species $s \in common\_bacteria\_set$, extract the genome sequence of s. Please refer to "*Supporting_Documents\sequence_download_procedure.pdf*" for guidelines on how to download these genome sequences.

3. For each protein $p \in protein\_set$, extract its corresponding gene sequence from the genome of each species $s \in common\_bacteria\_set$ (you retrieved these genomes in step 2 above). We will refer to the set of gene sequences corresponding to protein p as $homologous\_gene\_sequences(p)$. To do this extraction of gene sequences, you will need to find out the start and end position of p in the genome of each species $s \in common\_bacteria\_set$ by referring to the sheet named by the accession ID of s in the Excel file "*Supporting_Documents\protein_tables.xlsx*" and locating the protein p in it. The columns named "*start*" and "*stop*", indicate the start and end position of genes respectively. Then, you can extract the sequence segments from the start to stop positions of the respective genomes (you may have to write a small script to do this extraction).

   - **Note:** If a given protein is present in more than one location on the same genome (multiple entries in the protein table of a given species), take the first occurrence (earliest row).

E.g., let $protein\_set = \{FUSCfamilyprotein, MFStransporter, YraNfamilyprotein\}$. Then, when you need to extract the gene corresponding to the first protein "*FUSC family protein*" in the genome of NZ_CP014692.1 (Acetobacter aceti) you will have to look in sheet NZ_CP014692.1 of the file "*protein_tables.xlsx*" , locate the protein name "*FUSC family protein*" and read the *start* and *stop* positions. In this case you would get 177331 and 179517. Then, you will have to extract the segment of the Acetobacter aceti genome (you would have retrieved this in step 2) from position 177331 to 179517. You will repeat this process for all proteins in $protein\_set$ and all species in $common\_bacteria\_set$. In this example, for each protein $p \in protein\_set$ you would have extracted 20 homologous gene sequences, which would form the set $homologous\_gene\_sequences(p)$.

4. For each protein $p \in protein\_set$, use the sequences in $homologous\_gene\_sequences(p)$ as input data to build a phylogenetic tree to infer the evolution of the genes corresponding to protein p in the Acetobacteraceae family. Use UPGMA as the tree construction algorithm. Feel free to use any software or programming language for the implementations. You can use ClustalX software or Rstudio.

5. After completing steps 1-4, you would have built several phylogenetic trees, one for each protein $p \in protein\_set$. Now you will need to check whether these trees agree with each other or not. For this, you can compute the distance (difference) between each pair of trees you built. There are several distance measures you can use. For this assignment, we will use one of the most popular ones called the Robinson-Foulds Distance. Compute the Robinson-Foulds Distance between each pair of phylogenetic trees you built in step 4 above. Based on the distance calculations, comment on how the phylogenetic trees you built agree or disagree with each other.

   You can learn about the Robinson-Foulds Distance by referring to section 2.4 of the research paper provided to you titled "*A Metric for Phylogenetic Trees Based on Matching*", which can be found in the "*Supporting_Documents*" folder. If you wish, you can refer to the original paper that proposed the Robinson-Foulds Distance too, which is D.R. Robinson and L.R. Foulds, “Comparison of PhylogeneticTrees,”Math. Biosciences,vol. 53, pp. 131-147, 1981.

6. Disagreements between trees constructed for different genes is referred to as phylogenetic incongruence and causes challeneges if you were to infer a species tree from several gene trees. E.g., if you were to infer the phylogenetic for the Acetobacteraceae family using the genes tree you built (of course for this, you would need a lot more than just 3 or 4 genes trees for this). Explain the possible reasons for phylogenetic incongruence between trees constructed for different genes among the same set of species. You may refer to the following papers.

   - Dealing with incongruence in phylogenomic analyses.
   - Inferring Horizontal Gene Transfer.

2 Deliverables

D1 - Report: Write a report that includes the following with regard to the steps in the procedure above:

- $common\_bacteria\_set$ you would find in step 2 (listing the species IDs is sufficient)
- Phylogenetic trees you would build in step 4, the steps you followed to build them, and references to any software tools you used.
- Computation of the Robinson-Foulds Distances in step 5.
- The explanation requested in step 6.

D2 - Data and code: Provide the following as files saved in a folder named "data_and_code_<your group name>"

- Homologous gene sequences you would have extracted in step 3 in FASTA format. For each protein $p \in protein\_set$, save its homologous gene sequences in a separate file.
- Any code/scripts you wrote.
- A readme file explaining the steps to follow if someone wants to run your implementation on your data.

Note: Submit the Report in PDF format. Zip the "data_and_code_<your group name>" folder and upload it as a separate file to the same assignment link.