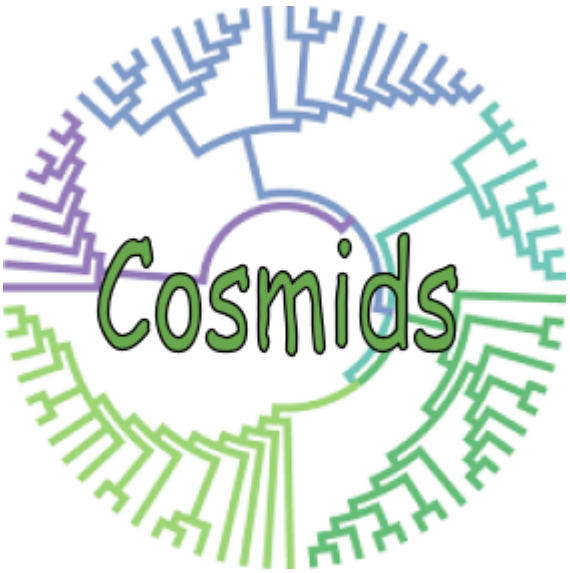
Assignment - Phylogenetic Trees Report

CS4742 - Bioinformatics



Group Members:
170406P - R. A. K. C. Nilanga
170521M - R. M. H. N. Rathnayake
170527L - S. H. R. Rukshani
170610K - M. P. J. S. Sumanapala

Step 1: Common_bacteria_set

All 21 bacteria are common for the given protein set
Common_bacteria_set = {'NZ_CP014692.1 ', 'NZ_CP023657.1', 'NZ_CP022699.1', 'NZ_CP014687.1', 'NZ_LN606600.1', 'NZ_CP011120.1', 'NZ_CP015164.1', 'NZ_CP015168.1', 'NZ_CP021524.1', 'NZ_CP022374.1', 'NZ_AP018515.1', 'NZ_CP023189.1', 'NC_017100.1', 'NC_017121.1', 'NC_017125.1', 'NC_017146.1', 'NZ_LN609302.1', 'NC_017111.1', 'NC_017150.1', 'NC_017108.1', 'NZ_AP014881.1'}

Step 3: Start and end positions of proteins

		Proteins			
Species	**Start**	**ABC transporter permease**	**LysR family transcriptional regulator**	**helix-turn-helix domain-containing protein**	**efflux transporter outer membrane subunit**
	End				
NZ_AP018515.1	38328	379517	359001	1427079
	39287	380431	359579	1428593
NC_017100.1	138760	44664	315904	155470
	139578	45029	316398	156954
NZ_AP014881.1	136243	42099	294358	151284
	137061	42464	294852	152768
NZ_CP022699.1	20827	18042	40006	117718
	21678	18938	40275	119154
NZ_CP021524.1	240268	754589	99530	294338
	241473	755500	100810	295903
NZ_CP023189.1	207292	149259	29399	141770
	208851	150152	29662	143359
NZ_CP015164.1	32236	121923	748077	98
	33840	122864	749357	1621
NZ_CP023657.1	109663	333	17020	559625
	110712	1220	17271	561190

NZ_CP014687.1	56129	14424	202443	768608
	57256	15347	202718	770161
NZ_CP014692.1	214989	63174	55106	999992
	215756	64115	57319	1001515
NC_017125.1	138760	44664	315904	155470
	139578	45029	316398	156954
NC_017108.1	138760	44664	315904	155470
	139578	45029	316398	156954
NC_017150.1	138760	44664	315910	155470
	139578	45029	316404	156954
NZ_CP022374.1	136111	42006	311284	152018
	136929	42371	311778	153502
NC_017121.1	138760	44664	315904	155470
	139578	45029	316398	156954
NZ_LN606600.1	80900	481992	1311595	136496
	82048	482918	1312500	138073
NZ_CP015168.1	22271	248637	1086506	561625
	23056	249503	1086736	563154
NZ_CP011120.1	139385	45285	320717	154424
	140203	45650	321211	155908
NZ_LN609302.1	51798	963	289340	303038
	52970	1892	289609	304606
NC_017146.1	138760	44664	315904	155470
	139578	45029	316398	156954
NC_017111.1	138760	44664	315904	155470
	139578	45029	316398	156954

Step 4: Phylogenetic trees

I) The extracted homologous gene sequences of each protein were separately written into 4 fasta files. Say the protein is p, the file that stored gene sequences that contains protein p was named as p.fasta. The content of the files were key value pairs; accession id as the key and sequence as the value. Then each fast file contains 21 key-value pairs. An example is shown below.

```
>NC_017108.1
TATTTGGGGTTGCTCATTGTTAGCATAAGCTCAAGAATTTTTCGGCGCATTTGCGTATCTGGAAT
ACTGTAATATGCACGCACCAGTTCCAGTGTTTCCTGCGTTTTGAATACGCTTTTTGTATCTGTTG
TAGGCGCTGCGGGCGTATTTTTGTCTGATGTTGTGCCAAAGCTTTTACGTGTTTCCGAAAACTGC
GCATGTTTGTTGGAAACAGCCCCGACGGGGTGCATATCATCAAAAAAGAAGCTGATCGGTACATC
CAGCACACAGGCAATATGATATAATCGCGAAGCACCTACCCGGTTTGCTCCGCGTTCATATTTTT
GTATTTGCTGGAAGGTAATTCCCAAAGCTTCTCCCAGCTTTTCCTGAGAAAGTCCCAGCAATGTG
CGGCGTAGTCTTATGCGTTTGCCAACATGGGCATCAATAGCACTGGCAGCAGCATGAGGCTGGAT
CATGGTGGTGTCGATCGGTTCCTGCTCCGTATGTTTCAC

>NC_017150.1
TATTTGGGGTTGCTCATTGTTAGCATAAGCTCAAGAATTTTTCGGCGCATTTGCGTATCTGGAAT
ACTGTAATATGCACGCACCAGTTCCAGTGTTTCCTGCGTTTTGAATACGCTTTTTGTATCTGTTG
TAGGCGCTGCGGGCGTATTTTTGTCTGATGTTGTGCCAAAGCTTTTACGTGTTTCCGAAAACTGC
GCATGTTTGTTGGAAACAGCCCCGACGGGGTGCATATCATCAAAAAAGAAGCTGATCGGTACATC
CAGCACACAGGCAATATGATATAATCGCGAAGCACCTACCCGGTTTGCTCCGCGTTCATATTTTT
GTATTTGCTGGAAGGTAATTCCCAAAGCTTCTCCCAGCTTTTCCTGAGAAAGTCCCAGCAATGTG
CGGCGTAGTCTTATGCGTTTGCCAACATGGGCATCAATAGCACTGGCAGCAGCATGAGGCTGGAT
CATGGTGGTGTCGATCGGTTCCTGCTCCGTATGTTTCAC

>NZ_CP022374.1
TATCTGGATACGCTCATTGTTTGCATAAGCTCAAGAATTTTTCGGCGCATTTGCGCGTCTGGAAT
GCTGTAATACACGCGTACCAGCTCCAGTGTTTCCTGCTTTTTGAATATGTTTTTTGTATCCGATG
TAGGTGCTACGCTAACATTCTTGTCCGATGGTGCACCAAAGCTTTTGCGTGTTTCCGAAAAATGT
GCCTGCTTATTGGAAACAACTTCCGCAGAATGCATATCATCAAAAAAGAAGCTGATCGGCACATC
CAACGCACAGGCAATATGATACAATCGCGAAGCGCCTACCCGGTTTGCTCCACGTTCATATTTCT
GAATCTGCTGGAAGGTAATTCCCAAAGCTTCTCCCAGTTTTTCCTGAGAGAGTCCCAACAATGTG
CGGCGTAGTCTTATGCGTTTGCCAACATGGGCATCAATGGCACTGGCAGCAGCATGAGGCTGGAT
CATGGTGGTGTCGATCGGTTCCTGCTCCGTATGTTTCAC

>NZ_LN609302.1
CAAAGGGTCGCCTTTTCGGCCGGAGTCAGCGGGCGGGCGGGGAAAACACGGCAGGGTGTCTCGTC
GGTCGTGGACCTGCGGACGCCGTTCTCGGTCCAGCCCTGAAGGTCGTAAAGGGTATAGACGATGC
GGCCACCAATCTTGCGATAGGTCGGGCCGGTGCCATAGGTGCGGTGTTTTTCCAGCGTGCGGATG
GAAATACCGAGGAAGCGCGCCGCGTCGGGCGTGCGCAGGAAACGGGGCGGCAACCCCGTCTTGGG
ATCGAGCAT
```

II) Then the content of each above file was given as inputs to the clustal omega web interface to generate Multiple Sequence Alignment. The generated alignments were obtained as CLUSTAL files.

```
CLUSTAL O(1.2.4) multiple sequence alignment


NZ_LN609302.1      ------------------------------------------------------------
NZ_LN606600.1      ----------TGCCCC-CAAAACCCGAGATTCCGAGCTTTCTGACATATGGCATCAACA
NC_017108.1        ------------------------------------------------------------
NC_017150.1        ------------------------------------------------------------
NC_017121.1        ------------------------------------------------------------
NC_017125.1        ------------------------------------------------------------
NC_017146.1        ------------------------------------------------------------
NC_017100.1        ------------------------------------------------------------
NC_017111.1        ------------------------------------------------------------
NZ_AP014881.1      ------------------------------------------------------------
NZ_CP022374.1      ------------------------------------------------------------
NZ_CP011120.1      ------------------------------------------------------------
NZ_CP014687.1      ------------------------------------------------------------
NZ_CP023657.1      ------------------------------------------------------------
NZ_CP023189.1      ------------------------------------------------------------
NZ_AP018515.1      ------------------------------------------------------------
NZ_CP021524.1      -----TAG-------------------------------------------------------
NZ_CP022699.1      ------------------------------------------------------------
NZ_CP015164.1      ------------------------------------------------------------
NZ_CP014692.1      TGGGTCTGTTCGGCACCATTAAACCCGGGCTGAGGAAATCTCAG-------GCGGCTATT
NZ_CP015168.1      ------TGATC-----------------------------------------------------


NZ_LN609302.1      ------------------------------------------------------------
NZ_LN606600.1      CAAAAACTGAT---------GACATAGGCTTT----GTGCATGTCGAAA--------ACG
NC_017108.1        ------------------------------------------------------------
NC_017150.1        ------------------------------------------------------------
NC_017121.1        ------------------------------------------------------------
NC_017125.1        ------------------------------------------------------------
NC_017146.1        ------------------------------------------------------------
NC_017100.1        ------------------------------------------------------------
NC_017111.1        ------------------------------------------------------------
NZ_AP014881.1      ------------------------------------------------------------
NZ_CP022374.1      ------------------------------------------------------------
NZ_CP011120.1      ------------------------------------------------------------
NZ_CP014687.1      ------------------------------------------------------------
NZ_CP023657.1      ------------------------------------------------------------
NZ_CP023189.1      ------------------------------------------------------------
```

III) The alignments generated in III) were given as the input to generate phylogenetic trees and tree data using the clustal phylogeny web application. The generated trees are shown in the following diagrams.

i) ABC transporter permease

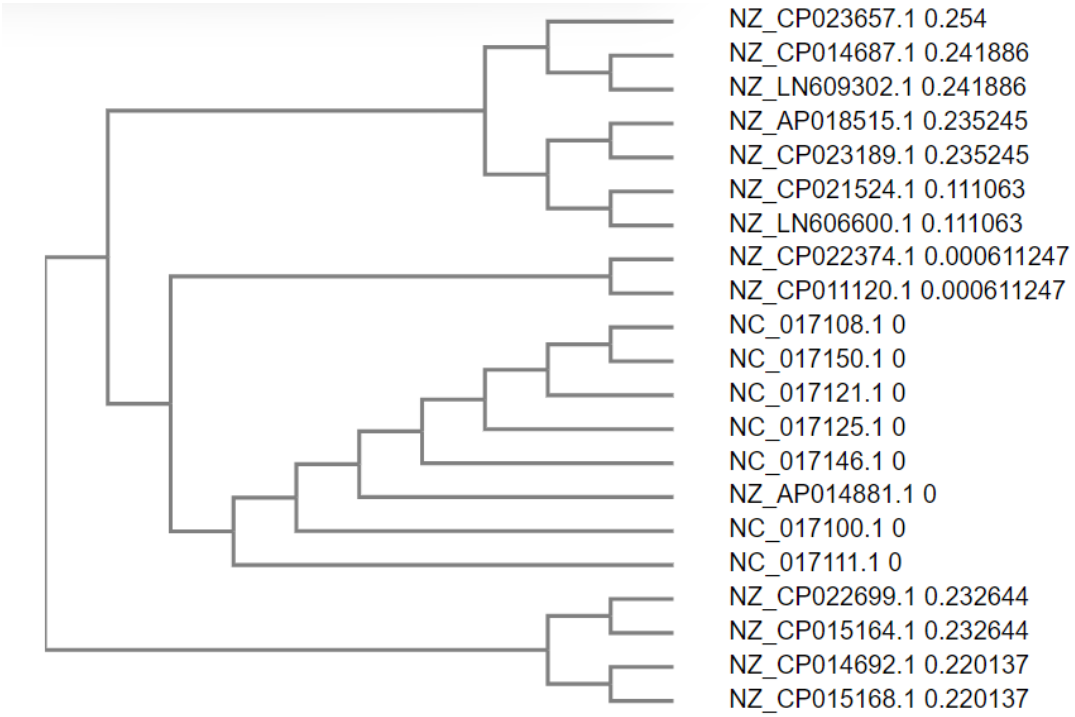


Figure 1: ABC transporter permease tree

ii) LysR family transcriptional regulator

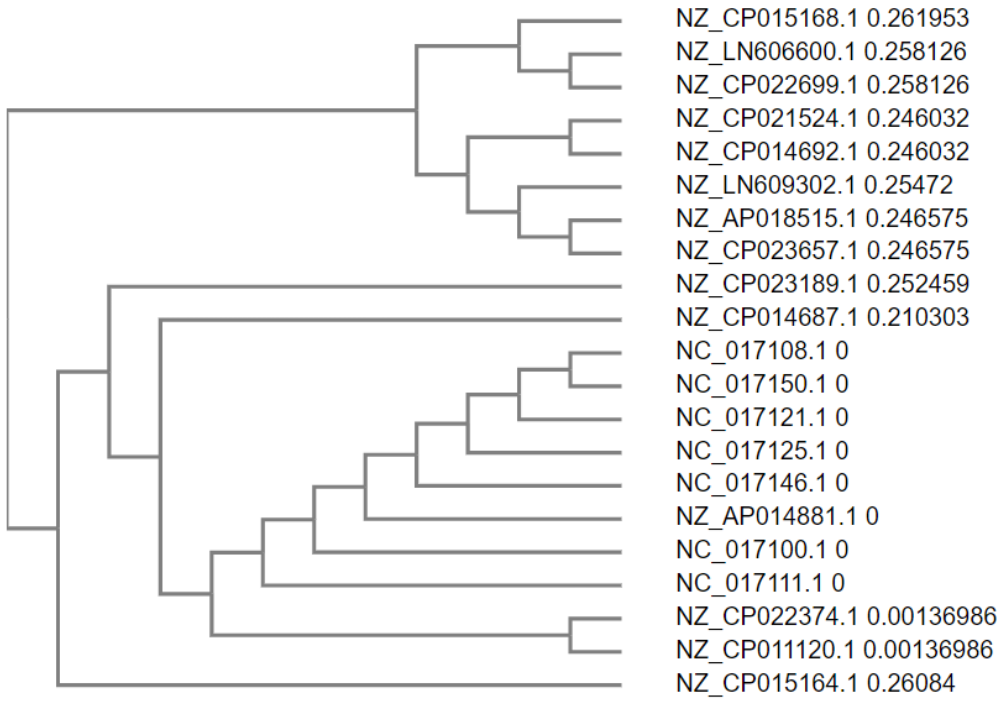


Figure 2 : LysR family transcriptional regulator tree

iii) helix-turn-helix domain-containing protein

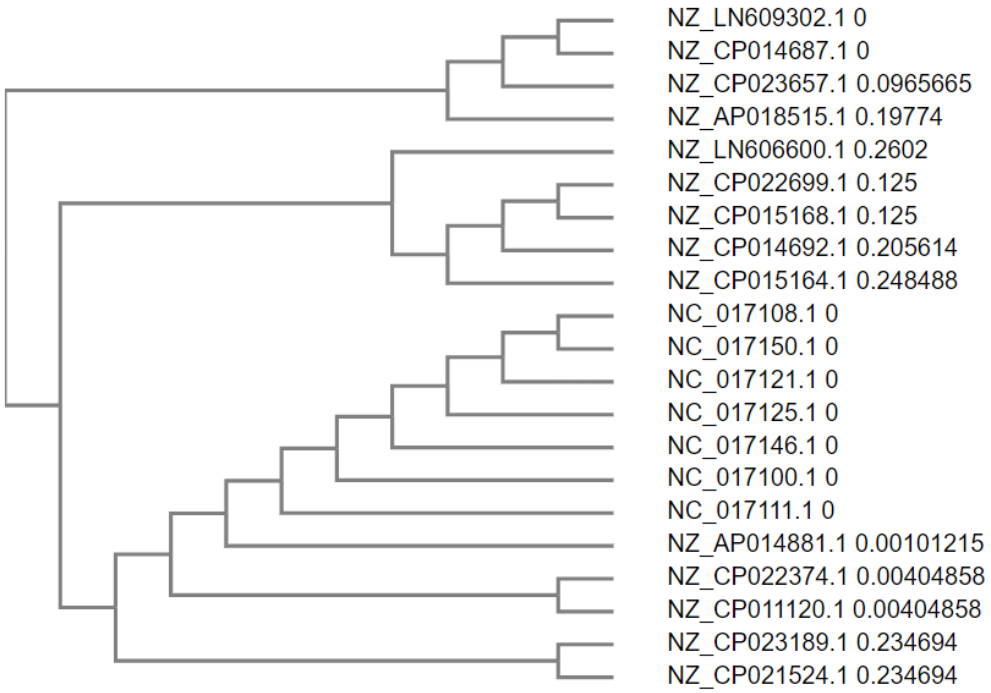


Figure 3 : helix-turn-helix domain-containing protein tree

iv) efflux transporter outer membrane subunit

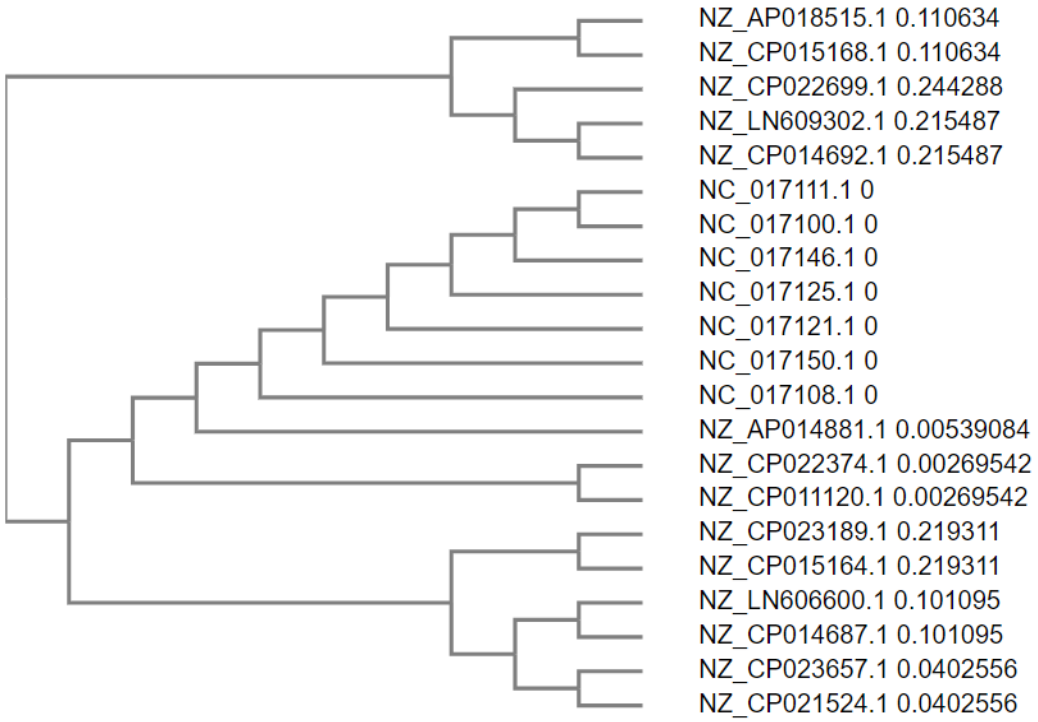


Figure 4 : efflux transporter outer membrane subunit tree

Step 5 : Computation of the Robinson-Foulds Distances

Robinson-Foulds Distance

Robinson-Foulds distance [1] between two unrooted leaf labeled trees T1 and T2 is equal to the normalized count of the bipartitions induces by one tree and not the other tree, that is,

$$\mathcal{D}_{RF}(T_1, T_2) = \frac{1}{2}((|\Gamma(T_1) - \Gamma(T_2)|) + (|\Gamma(T_2) - \Gamma(T_1)|)).$$

where,

$$\Gamma(T) = \{\pi_e \mid e \in E(T)\}$$

$\Gamma(T)$ - The set of bipartitions

$E(T)$ - The set of internal edges

Due to a tree with n leaves having at most n-3 nontrivial bipartitions, n-3 is the largest possible Robinson-Foulds distance between two trees.

The following code snippet is used to calculate the Robinson-Foulds distance between each pair of phylogenetic trees with the help of 'ETEtoolkit' python library [2].

```
from ete3 import Tree

tree_data = ['/content/drive/Shareddrives/Cosmids/Data/Tree Data UPGMA/ABC transporter permease.txt',
             '/content/drive/Shareddrives/Cosmids/Data/Tree Data UPGMA/LysR family transcriptional regulator.txt',
             '/content/drive/Shareddrives/Cosmids/Data/Tree Data UPGMA/helix-turn-helix domain-containing protein.txt',
             '/content/drive/Shareddrives/Cosmids/Data/Tree Data UPGMA/efflux transporter outer membrane subunit.txt']

for idx, a in enumerate(tree_data):
  for b in tree_data[idx + 1:]:
    t1 = Tree(a)
    t2 = Tree(b)

    rf, max_rf, common_leaves, parts_t1, parts_t2, x4, x5 = t1.robinson_foulds(t2, unrooted_trees=True)

    print(a, b)
    print("RF distance is %s over a total of %s" %(rf, max_rf))
```

The following table contains the Robinson-Foulds distance between each pair of phylogenetic trees.

	ABC transporter permease	LysR family transcriptional regulator	helix-turn-helix domain-containing protein	efflux transporter outer membrane subunit
ABC transporter permease		18	16	30

LysR family transcriptional regulator	18		22	30
helix-turn-helix domain-containing protein	16	22		28
efflux transporter outer membrane subunit	30	30	28	

Max Robinson-Foulds distance = 36

Step 6 : Explanation

Possible reasons for phylogenetic incongruence between trees constructed for different genes among the same set of species

1. Failure to recover the correct gene trees or errors occur in tree construction.
    a. Stochastic errors - Due to insufficient sequence length.
    b. Systematic errors - Due to deviation from the model assumptions. [3]
2. Biological Reasons

    Phylogenetic incongruence can occur when gene trees are truly different from each other. The three major evolutionary mechanisms resulting in true phylogenetic discordance between genes are incomplete lineage sorting, hidden paralogy and horizontal gene transfer (HGT). [3]

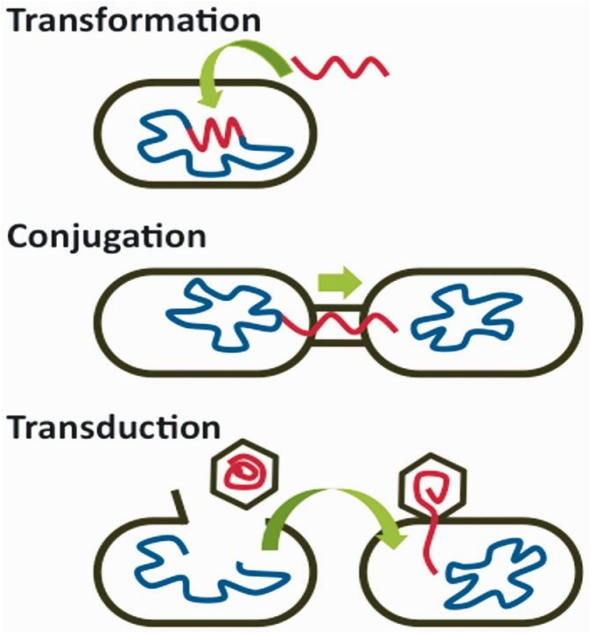
    a. Incomplete lineage sorting - Occurs when an ancestral species experiences many speciation events (lineage segregation, all species die, natural calamities, slaughter) in a short period of time. Because of incomplete lineage sorting, there is no assurance that alleles will sort in a lineage to match the overall species pattern. Speciation events assort gene patterns and utinamatly patterns do not match to the species pattern. These issues then lead to gene trees that are "discordant" with the species tree. As a result, the extracted trees may differ from one another.

    b. Hidden paralogy - Some copies of paralogous genes can be retained in the genome after their duplication, but some copies can also be lost. These paralogous copies and loss patterns can produce complex patterns that make interpreting phylogenetic trees challenging, often leading to incorrect judgments about species relationships. As a result, the true phylogeny will reflect the duplication history of the gene, which is independent of the species' divergence history. [3]

c. Horizontal gene transfer (HGT) - If there are genetic exchanges (not a sexual manner) among species, the phylogeny of individual genes will be altered by the frequency and character of transfers.

There are three major genetic mechanisms which cause horizontal gene transfer.

- Transformation: Bacteria take up DNA from their environment
- Conjugation: Bacteria directly transfer genes to another cell
- Transduction: Bacteriophages (bacterial viruses) move genes from one cell to another [4][5]



Using the above three genetic mechanisms, bacteria can transfer and receive genetic material from other species. Which causes gene pattern mismatch and ultimately causes gene tree mismatches.

References

[1] Lin, Yu, Vaibhav Rajan, and Bernard ME Moret. "A metric for phylogenetic trees based on matching." IEEE/ACM Transactions on Computational Biology and Bioinformatics 9.4 (2011): 1014-1022.

[2] “ETE Toolkit - Analysis and Visualization of (phylogenetic) trees”. [online] Available at: http://etetoolkit.org/ [Accessed 14 February 2022].

[3] N. Galtier en V. Daubin, “Dealing with incongruence in phylogenomic analyses”, Philosophical Transactions of the Royal Society B: Biological Sciences, vol 363, no 1512, bll 4023–4029, 2008.

[4] M. Ravenhall, N. Škunca, F. Lassalle, en C. Dessimoz, “Inferring horizontal gene transfer”, PLoS computational biology, vol 11, no 5, bl e1004095, 2015.

[5] A. R. Burmeister, “Horizontal Gene Transfer”, Evolution, medicine, and public health, vol 2015, no 1, bll 193–194, Jul 2015.