BIF701 Lab 5 Evolutionary Processes: Placing the Archaea in the Tree of Life

By Christopher Eeles

Introduction and Purpose

In this lab we will utilizing web-based bioinformatics tools to develop phylogenetic trees exploring the relationships between two species from each of the Eukaryotic, Bacterial, and Archaeal domains of the tree of life.[1] Comparisons will be made using sequence alignment of the heat-shock protein Hsp70 gene between each species and across their respective domains.[1] Hsp70 is highly conserved throughout the tree of life and thus can be used as a molecular clock—a metric for estimating the evolutionary distance between two or more species.[1, 2] The resulting alignment data will be used to derive distance matrices and their associated phylogenetic trees for several settings from each online tool; these settings represent the different statistical methods (*i.e.* UPGMA, NJ) used for clustering species into groups based on evolutionary relatedness.[2] Each clustering method comes with its own set of assumptions, strengths, and weaknesses and therefore selection of the best settings for a task will depend on the hypothesis in question.[2] Through this lab we will become familiar with the output from online phylogenetic analysis tools and use this information to derive conclusions about the validity of molecular clocks as an approximation of evolutionary relation.

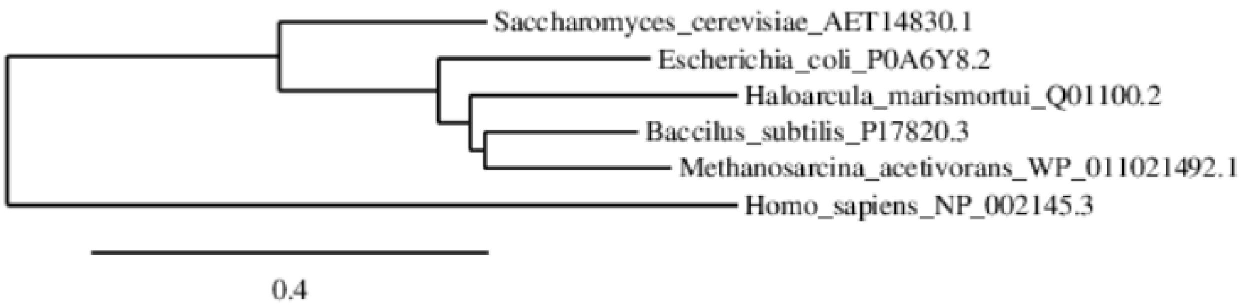
Method and Results

The first step in our analysis requires identifying the species associated with the provided gene accession numbers from the NCBI protein database. The results are as follows:

Domain	Accession number	Protein Source
Eukaryotic:	NP_002145	*Homo sapiens*
	AET14830	*Saccharomyces cerevisiae*
Bacterial:	P0A6Y8	*Escherichia coli*
	P17820	*Bacillus subtilis*
Archaeal:	WP_011021492	*Methanosarcina acetivorans*
	Q01100	*Haloarcula marismortui*

The FASTA file from each protein was placed into a text file which was input into the "*a la carte*" phylogeny.fr analysis tool. This system chains together programs for the identification of homologous sequences, their multiple alignment, phylogenetic reconstruction and graphic representation of the inferred tree; it is designed to be accessible to biologists with no experience in phylogeny, but powerful enough for specialized use.[5] The tree created using the **(1)** BioNJ setting by phylogeny.fr is:

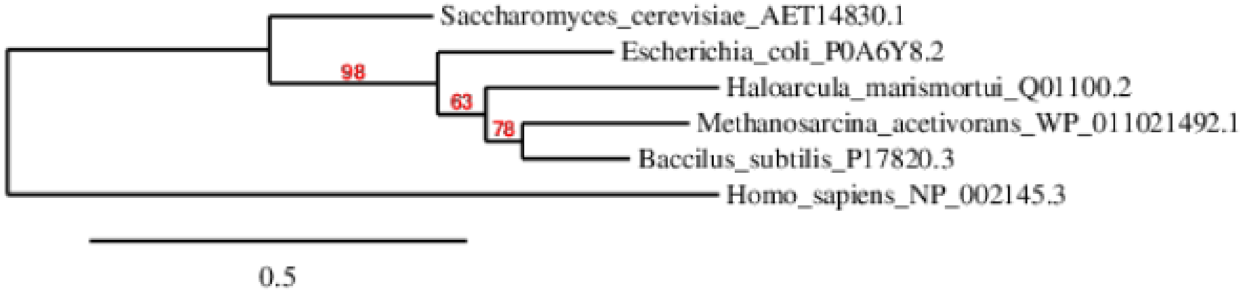
Figure 1


Saccharomyces_cerevisiae_AET14830.1
Escherichia_coli_P0A6Y8.2
Haloarcula_marismortui_Q01100.2
Baccilus_subtilis_P17820.3
Methanosarcina_acetivorans_WP_011021492.1
Homo_sapiens_NP_002145.3
0.4

BIF701 Lab 5 Evolutionary Processes: Placing the Archaea in the Tree of Life
By Christopher Eeles

The tree created using the **(2)** PhyML setting by phylogeny.fr is:

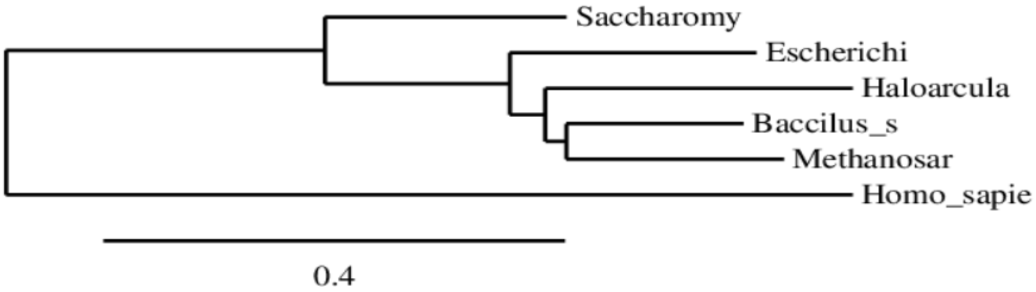
Figure 2


Saccharomyces_cerevisiae_AET14830.1
Escherichia_coli_P0A6Y8.2
Haloarcula_marismortui_Q01100.2
Methanosarcina_acetivorans_WP_011021492.1
Baccilus_subtilis_P17820.3
Homo_sapiens_NP_002145.3
98
63
78
0.5

The .dist file in Phylips format from **(1)** was copied and used as input to the EMBOSS fneighbor tool available from the National Health Research Institute (NHRI). The European Molecular Biology Open Software Suit (EMBOSS) is a collection of freeware that automatically processes a wide range of data types for myriad analyses.[3] Fneighbor is a program from EMBOSS specialized for phylogenetic analysis by modified NJ or UPGMA methods.[4] The NJ method produces an unrooted tree without the assumption of a clock, while the UPGMA assumes a clock and therefore could generate a rooted tree given the proper context.[4] For this analysis we have input the distance matrixes from

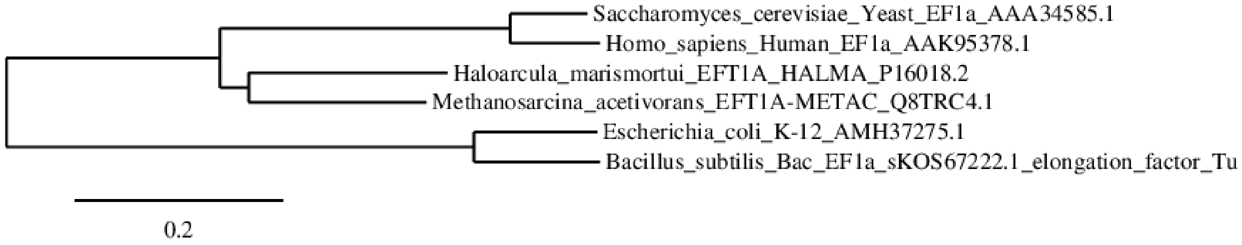
The fneighbor program only takes distance matrices as its input and therefore only the results of the BioNJ phylogeny.fr analysis are included. Taking the distance matrix from **(1)** as an input into fneighbor and using the BioNJ setting resulted in:

Figure 3



For comparison, the final tree from the Lab 5 example with elongation factor as the molecular clock is:

Figure 4


Saccharomyces_cerevisiae_Yeast_EF1a_AAA34585.1
Homo_sapiens_Human_EF1a_AAK95378.1
Haloarcula_marismortui_EFT1A_HALMA_P16018.2
Methanosarcina_acetivorans_EFT1A-METAC_Q8TRC4.1
Escherichia_coli_K-12_AMH37275.1
Bacillus_subtilis_Bac_EF1a_sKOS67222.1_elongation_factor_Tu
0.2

BIF701 Lab 5 Evolutionary Processes: Placing the Archaea in the Tree of Life
By Christopher Eeles

Discussion and Conclusions

Analysis of the heat shock protein Hsp70 data using both the BioNJ and PhyML resulted in trees with slight variations in the scale of relationships, but ultimately with the same phylogenetic conclusions. While these results are unremarkable at first glance, they do provide additional confidence in the structure of the generated phylogenetic trees. The BioNJ distance matrix was also taken to EMBOSS fneighbor for refinement; this yielded no apparent changes in either scale or structure of relations but was necessary to ensure comparability with the elongation factor analysis conducted during class. Given the lack of a common ancestor these trees are unrooted. This conclusion is supported by the tool research discuss in the Methods and Results section: use of the NJ methodology for phylogenetic analysis produces unrooted trees. For comparison, it would have been interesting to generate a UPGMA distance matrix for analysis in fneighbor, which supports generation of trees using this data, but phylogeny.fr did not provide an option for UPGMA based analysis.

Comparing the results to the elongation factor analysis, it seems that our secondary analysis with Hsp70 brings the initial results, and the hypothesis domain hypothesis suggested by Carl Woese, into question[1]. While **Fig. 4** shows the expected divergence of the Bacterial proteins into first the Archaeal and then Eukaryotic ones, **Fig. 1**, **2** and **3** all show the Bacterial and Archaeal Hsp70 proteins to be too closely related to clearly resolve the split between these domains. This weakens the argument for the validity molecular clock hypothesis but does not undermine it completely. It is possible, for example, that variations in mutation rate at this locus between species resulted in changes to the Hsp70 gene which distort the evolutionary time passage. If this is the case, it weakens the argument for the use of Hsp70 as a molecular clock in this analysis. This makes sense, as random chance plays a large role in gene mutation. It would therefore be necessary to compare a large number of proposed molecular clock genes in order to come to relational conclusions with a high degree of certainty. Given the superficiality of this analysis, and our lack a research background in evolutionary biology, it is likely that this is already the case if one were to examine the literature.

References

1. School of Biological Sciences and Applied Chemistry. (2018). *BIF701 Lab 5 Evolutionary Processes: Placing the Archaea in the Tree of Life*. Toronto, ON: Seneca College.

2. School of Biological Sciences and Applied Chemistry. (2018). *BIF701 Evolutionary Processes.* Toronto, ON: Seneca College.

3. Rice , P., Longden , I. and Bleasby, A. (2000). *EMBOSS: The European Molecular Biology Open Software Suite*. Trends in Genetics 16(6) (pp. 276-277). Retrieved from http://emboss.sourceforge.net/what/.

4. Felsenstein, J. (2004). *EMBOSS: fneighbor Manual.* Retrieved from http://bioinfo.nhri.org.tw/cgi-bin/emboss/help/fneighbor.

5. Dereeper A. *et al.* (2008). Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res*. 1(36) (pp. 465-469). Retrieved from https://academic.oup.com/nar/article/36/suppl_2/W465/2505761.