KB8019 - Final Project Report Dimitri Wirjowerdojo
Group 2 Elinor Löverli

Summary

The reconstruction of phylogenetic trees between multiple organisms have been extensively used to study the evolutionary relationship between them. Usually, phylogenetic trees are constructed using orthologous sequences to create a species tree or other type of relationships to show how different genes or proteins in a large gene family are connected. The aim of this study is to analyze the genomes of five different species based on their nucleotide frequency, GC content and distance matrices to find the evolutionary relationship between them. The other goals include the identification of the possible open reading frames from each genome and the comparison of such analyses with the evolutionary relationship constructed using different methods.

We identified possible open reading frames, using different approach for eukaryotes and prokaryotes, for *B. thetaiotaomicron, C. aurantiacus, F. nucleatum, T. maritima* and *S. cerevisiae* chr XVI. The number of ORFs which we identified differ significantly to other methods due to fundamental difference in algorithm. We then calculated the genome and protein statistics, specifically the GC content, the nucleotide and dinucleotide frequencies and their cosine similarity and also the amino acid and diamino acid frequencies and their cosine similarity. Multiple distance matrices were generated based on these statistics and subsequently used for the reconstruction of dendrogram in order to be compared to the previously generated phylogenetic trees constructed using other methods. We found that the strongest agreement between our dendrogram constructed from the cosine similarity of the diamino acid frequencies and the orthologous cluster metagene-approach tree.

Finding ORF

The identification of the taxonomic classifications of our five species was done using NCBI's blastn program on default settings: Seq03 - *Bacteroides thetaiotaomicron*, strain 7330; Seq06 - *Chloroflexus aurantiacus* J-10-fl; Seq10 - *Fusobacterium nucleatum* subsp. polymorphum strain ChDC F319; Seq20 - *Thermotoga maritima*, strain Tma100; and Seq36 - *Saccharomyces cerevisiae* S288C chr XVI. As there are both prokaryotic and eukaryotic organisms in our samples, we constructed different ORF finder algorithm for each taxonomic domain.

In short, we first identified the all the non-overlapping start and stop codons in all six reading frames. For our prokaryotic samples, additional start codons (GTG and TTG) were also integrated as prokaryotes have been reported to also use them (Sacerdot et al., 1982) and so does one of the established program for such use, Glimmer (Delcher, et al., 2007). Due to the complexity of coding for overlapping ORFs and technical limitations, the polycystronicity of prokaryotes was not considered. From here, different algorithms to filter down the possible open reading frames were used. For the yeast, the filtering step was limited to checking the presence of adenine at -3 upstream of ATG (Hamilton, Watanabe and de Boer, 1987; Kozak, 1987). However, for the prokaryotes, additional filtering was done by checking the presence of the translation consensus/Shine-Dalgarno sequence at around -10 upstream of the start codon (Vimberg et al., 2007) or the transcription consensus (-35 box and -10/Pribnow box) (Harley and Reynolds, 1987) upstream of the Shine-Dalgarno sequence or the combination of the two. The -10 box is also known as the TATA box in eukaryotic organisms, though it is only present in about 16-25% of yeast genome (Erb and van Nimwegen, 2011). Additionally, each ORFs was checked for a minimum base pair length of 90 (Hyatt et al., 2010), even for the eukarya though it is known that the size of genes is in the range of thousands of base pairs. The results of these filtering algorithms are presented in Table 1 below, alongside comparison with the results from the number of ORFs as reported in other databases.

As can be seen in Table 1, there are discrepancies between the number of genes as reported by NCBI Taxonomy browser, from using the ORFfinder program from NCBI (Sayers et al., 2009; Benson et al., 2008), Glimmer v3.0 and Genscan, and also from our algorithms. Glimmer, which was designed with stronger emphasis for use in bacteria, archaea and viruses, works by using interpolated Markov model, resulting in higher than 99% specificity (Delcher et al., 2007). Conversely, Genscan was designed for eukaryotes, with a similar architecture to a generalized hidden Markov model. The model comes with additional filtering parameters such as the Maximal Dependence Decomposition which was designed to take into account signaling positions by modeling functional signals in DNA or protein sequences. On their test set, Genscan was shown to be able to exactly identify 75-80% of the exons (Burge and Karlin, 1997).

Table 1. Statistics regarding the number of open reading frames for each species and the identified open reading frames.

Seq	Number of Genes (as reported by NCBI Taxonomy database)	Number of Open Reading Frames					ORF Size (bp)	
		NCBI ORFfinder	Glimmer/ GenScan¤	Transcription Only	Translation Only	Both	Max/ Min	Median/ Mean
03	4903	56561	7035	12968	14140	3056	4398 / 90	144 / 237
06	3991	63116	7041	6725	20282	2194	16641 / 90	165 / 290
10	2130	11569	2755	4764	3110	1139	13200 / 90	447 / 636
20	5805	17704	3137	3107	6165	1077	5073 / 90	180 / 376
36	497	8029	338¤	-	1353	-	7470 / 90	144 / 410

* *Number of genes from plasmid.*

The approaches of these established programs are significantly different to our algorithms that did not use any machine-learning method, which may definitely have been the cause for the major difference in the number of predicted ORFs. Ultimately, for subsequent use in this study, we decided only to consider the ORFs predicted with the translational promoter filtering. Our reasoning includes the fact that this approach was applicable to all of our genomes in order to keep the consistency and that in terms of number they are not irrationally different.

From the translational promoter-filtering predicted ORFs, we generated the proteome for each species. Subsequently, we obtained the reference proteome for the species from the UniProt proteome database (The UniProt Consortium, 2017) and ran blastp for each species against their corresponding reference proteome. For yeast, we ran it only against the proteome for chromosome XVI. The settings for blastp was left on default settings, except for the number of hits (flag `-max_hsps`) which was limited to 1 to expedite the process. The results are divided into those with expectation values equal or less than 0.001 or larger than 0.001 as shown in Table 2 below.

Table 2. Number of hits from blastp against the reference proteome for each species, based on the expectation value.

Seq. No.		03	06	10	20	36
Number of hits based on E-value	≤ 0.001	1632	2874	545	1172	146
	> 0.001	12315	17221	2512	4901	1176
No hits		193	187	53	92	31

The expectation value signifies the number of match one should expect to find when running a search against a certain database, or in other words, it acts as random noise in the searching process (Sayers et al., 2009; Benson et al., 2008). The use of 0.001 as the e-value threshold is justified by a study showing that 99% of hits below this threshold were considered as homologs according to PFAM clans (Boekhorst and Snel, 2007). As the blast search was run against the corresponding reference species' proteome, we made the assumption that the 'homology' could be interpreted as a very likely protein-coding gene for each hit under the threshold.

Subsequently, we constructed a histogram to see the distribution of the size of the ORFs (Fig. 1A). Compared to the distribution plot obtained from Glimmer/Genscan (Fig. 1B), there is a strong agreement between them, with the size of most ORFs being around 1kbp. However, the numbers of ORFs in the smallest bin size (1000) appear to be underreported by our algorithms. Additionally, consistent with the number of predicted ORFs, species 06 also have the highest number of ORFs in the smallest bin size and species 36 with the lowest.

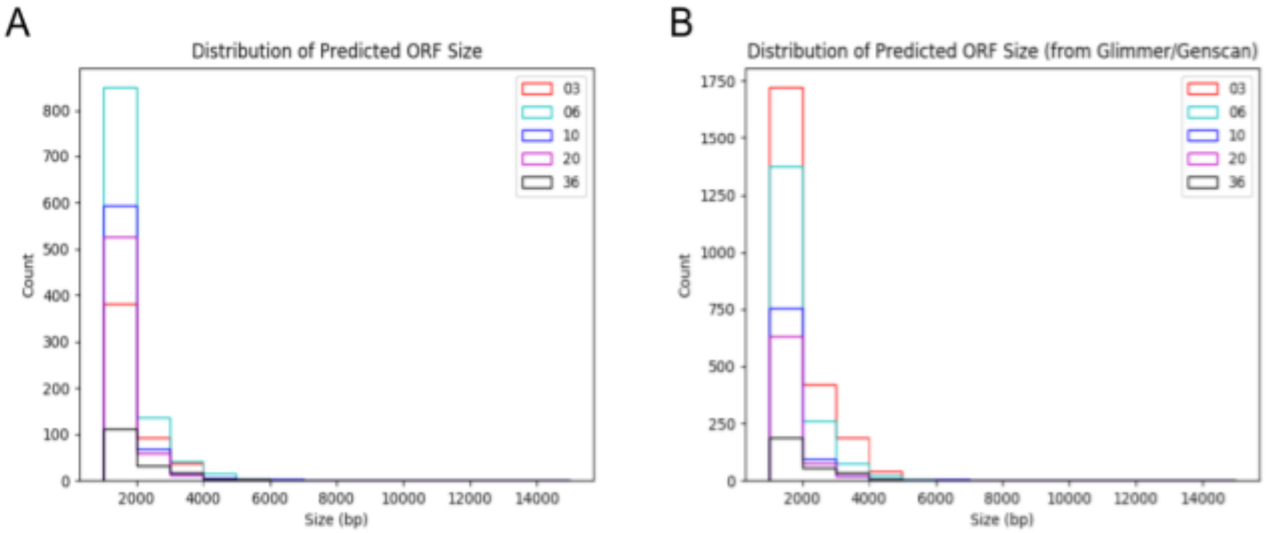


Fig. 1. The distribution of predicted open reading frame size constructed using ORFs predicted (A) from translation-only filtration (B) by Glimmer/Genscan.

Genome & Proteome Statistics

The statistics that were determined from the genome of our sample species include the GC content and the nucleotide and dinucleotide frequencies and their cosine similarity. The GC content for each species are shown in Table 3 below.

Table 3. GC content from the genome of five species.

Seq.	03	06	10	20	36
GC Content	43%	57%	27%	46%	38%

Additionally, for each species, the nucleotide, dinucleotide, amino acid and diamino acid frequencies are calculated. Except for the diamino acid frequency, they are illustrated in Figure 2, 3 and 4, respectively.

The GC content in a genome is associated with genes in prokaryotes and introns in eukaryotes. GC rich regions are low in translational stop codons (TAA, TAG and TGA) and their reverse complements (TTA, CTA and TCA) and therefore the probabilities for ORFs are higher in these regions (Oliver and Marín, 1996). Additionally, in prokaryotic genomes the GC content is highly variable, ranging from 17% to 75% (Brocchieri, 2014). Our results seem to be in agreement: as can be seen in Table 3, the GC content for the prokaryotic genomes varied between 27% and 57%. Conversely, for the eukaryotic sequence *S. cerevisiae* chromosome XVI (Seq36) the GC content was 38%.

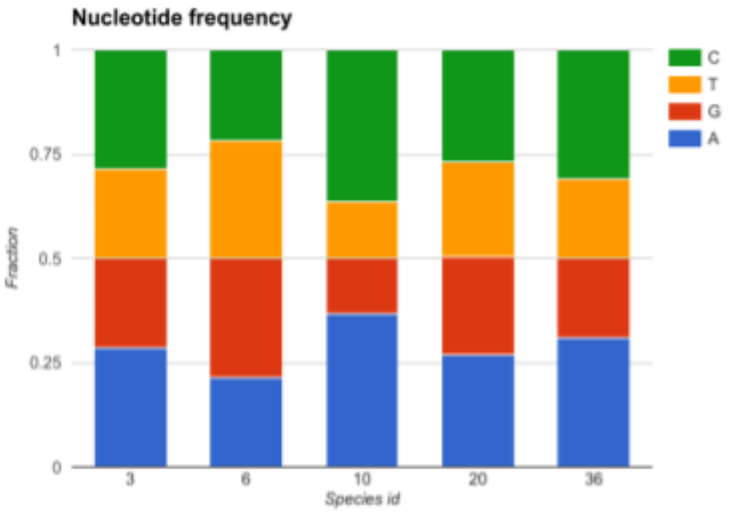


Figure 2. Nucleotide frequency for all five genomes. Each staple represent a genome with the nucleotide frequency shown on the y-axis.

As can be seen in Table 1, the reported genes by NCBI Taxonomy database somewhat agrees with the GC content in Table 3. Genome 10 with the lowest GC content also had the lowest amount of genes among the prokaryotic genomes. However, in the genomes 03, 06 and 20 the association is not as clear, but the GC content is roughly similar (43%-57%) and so are the NCBI gene count (3991-5805). Genome 36 had the lowest amount of genes, only 497, and also a low GC content (38%).

The nucleotide content is different for each genome as seen on Figure 2. The distribution of the nucleotides are approximately equal in both species 3 and 20 whereas species 10 has the greatest nucleotide distribution. It was also observed that there is a 1:1 distribution

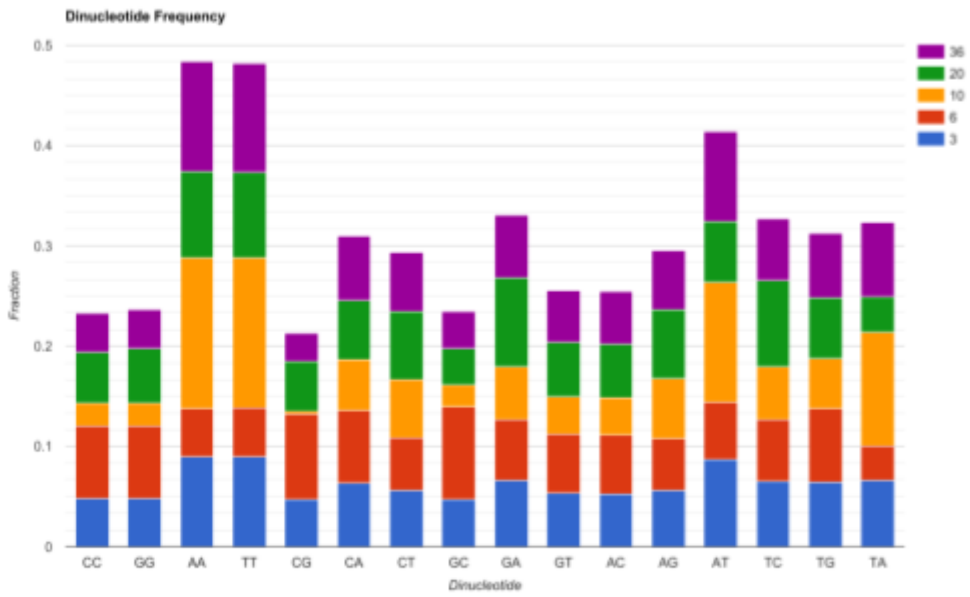


Figure 3. Dinucleotide frequency for all five genomes. Each staple represent a dinucleotide pair with the fraction for each genome is represented on the y-axis.

between purines and pyrimidines in all species.

It is harder to observe the trends in the dinucleotide frequency. However, except for Seq06, all genomes seem to be in rich AA and TT dinucleotides. In addition, despite being of different taxonomic domain, there seems to be an agreement in the distribution between Seq10 and Seq36 except for the dinucleotide CG, which is significantly underrepresented.

The challenge in trend observation is also found with the amino acid fraction, as shown in Figure 4. There are four amino acids that have higher distribution than the others: leucine, serine, isoleucine and arginine. Conversely, on the lower scale, there are methionine, tryptophan and cysteine.

Phylogenetic Tree/Dendrogram Reconstruction

The reconstruction of dendrograms were based only on the prokaryotic organisms in order to maintain consistency with the results of the lab exercises. In addition, attempts at incorporating the genome of yeast chr. XVI only resulted in trees that were skewed. Multiple distance matrices were constructed from several different metrics, as mentioned in the previous section. The formula to calculate the distance matrices can be found in Appendix A.

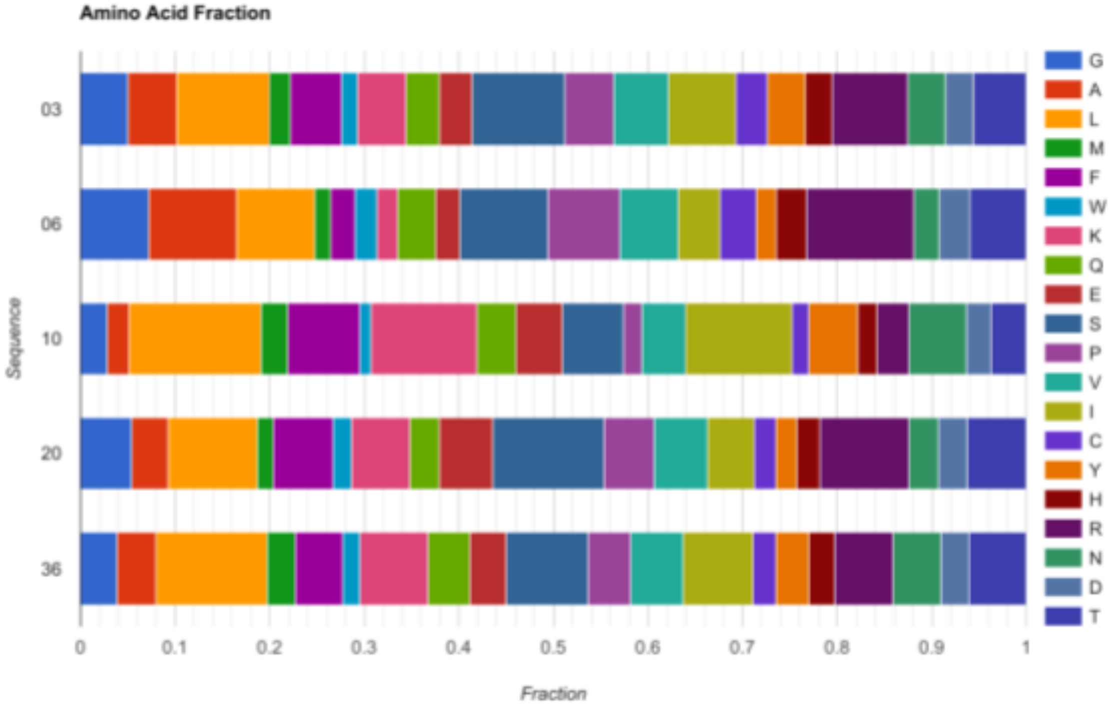


Figure 4: Amino acid frequency in sequence 03, 06, 10, 20 and 36.

The dendrograms were reconstructed using the program Belvu v2.2-1-gdd68 (Sonnhammer and Hollich, 2005) using the `-T R` flag to force reconstruction based on a distance matrix. In total, we created 9 dendrograms, each for each distance matrix. However, for the analysis we narrowed down to only comparing with the trees reconstructed based on amino acids/proteomes as this was the approach that was taken in the lab exercises (other dendrograms can be found in Appendix B). Additionally, we believe by using the diamino acid frequency, the precision is higher as it takes into account two amino acids at the same time. The dendrograms for the diamino acid and its cosine similarity are illustrated in Figure 5.

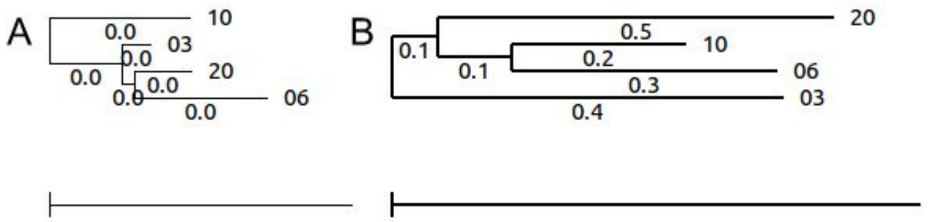
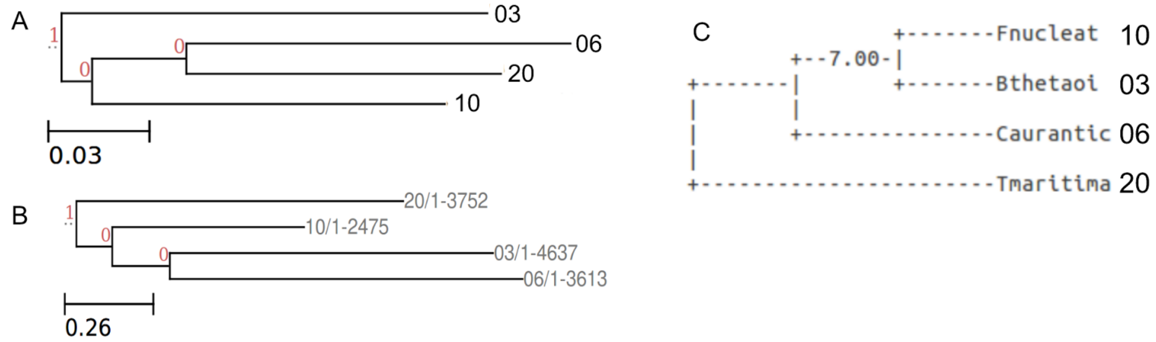


Figure 5. Dendrograms reconstructed using (A) diamino acid frequency and (B) cosine similarity of the diamino acid frequencies.



```
                    +-------Fnucleat  10
          +--7.00-|
+-------|           +-------Bthetaoi  03
|         |
|         +---------------Caurantic 06
|
+----------------------Tmaritima 20
```

Figure 6. Phylogenetic tree reconstruction from (A) 16s rRna using neighbor-joining method and (B) orthologous cluster metagene. (C) Consensus tree created from orthologous cluster genes using phylip.

As can be seen in Figure 5 and 6, there are similarities and differences between the dendrograms and the trees obtained previously in the lab. The closest similarity can be found between the cosine similarity of the diamino acid frequencies (Figure 5B) and the phylogenetic tree reconstructed from orthologous cluster metagene (Figure 6B, lab 4). The outgroup of these two trees is also the outgroup in the consensus tree that was created from orthologous cluster genes using the phylip consense program (Figure 6C, lab 4). On the other hand, the consensus tree (Figure 6C) seems to be in exact disagreement with the dendrogram reconstructed using diamino acid frequency (Figure 5A). However, these differences are expected as essentially the method of reconstruction is different. Furthermore, Zvelebil and Baum (2007) argued that the construction of phylogenetic trees is based under the assumption that the sequences derived from a single ancestral sequence and hence reconstruction of species tree should only use orthologous sequences.

Scripts

The scripts were developed on python v3.6 and can be found at https://github.com/dmtr13/KB8019. Backwards compatibility not tested.

References

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Sayers, E.W. 2009. GenBank. *Nucleic Acids Research*. Jan;37(Database issue):D26-31. Epub 2008 Oct 21.

Boekhorst, J. and Snel, B. 2007. Identification of homologs in insignificant blast hits by exploiting extrinsic gene properties. *BMC Bioinformatics*. **8**(356).

Brocchieri L. 2014. *The GC Content of Bacterial Genomes.* Journal of Phylogenetics & Evolutionary Biology. **2(**e108)

Burge, C. and Karlin, S. 1997. Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology*. **268**(1).

Delcher, A. L., Bratke, K. A., Powers, E. C. and Salzberg, S. L. 2007. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics*. **23**(6), PP.673-679.

Erb, I. and van Nimwegen, E. 2011. Transcription Factor Binding Site Positioning in Yeast: Proximal Promoter Motifs Characterize TATA-Less Promoters. *PLoS ONE*. **6**(9): e24279.

Hamilton, R., Watanabe, C. K. and de Boer, H. A. 1987. Compilation and comparison of the sequence context around the AUG startcodons in *Saccharomyces cerevisiae* mRNAs. *Nucleic Acids Research*. **15**(8), pp.3581-3593.

Harley, C. B. and Reynolds, R. P. 1987. Analysis of E.coli Promoter sequences. *Nucleic Acids Research*. **15**(5), pp.2343-2361.

Hyatt, D., Chen, G., LoCascio, P. F., Land, M. L., Larimer, F. W. and Hauser, L. J. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*. **11**(1).

Kozak, M. 1987. An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Research*. **15**(20), pp.8125-8148.

Oliver, J. L. and Marín, A. 1996. A relationship between GC content and coding-sequence length. *Journal of Molecular Evolution*. **43**(3), pp.216-223.

Sayers, E.W., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., DiCuccio, M., Edgar, R., Federhen, S., Feolo, M., Geer, L.Y., Helmberg, W., Kapustin, Y., Landsman, D., Lipman, D.J., Madden, T.L., Maglott, D.R., Miller, V., Mizrachi, I., Ostell, J., Pruitt, K.D., Schuler, G.D., Sequeira, E., Sherry, S.T., Shumway, M., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusova, T.A., Wagner, L., Yaschenko, E., and Ye, J. 2009. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*. Jan;37(Database issue):D5-15. Epub 2008 Oct 21.

Sacerdot, C., Fayat, G., Dessen, P., Springer, M., Plumbridge, J. A., Grunberg-Manago, M. and Blanquet, S. 1982. Sequence of a 1.26-kb DNA fragment containing the structural gene for *E. coli* initiation factor IF3: presence of an AUU initiator codon. *The EMBO Journal*. **1**(3), pp.311-315.

Sonnhammer, E.L.L. and Hollich, V. 2005. *Scoredist*: A simple and robust protein sequence distance estimator. *BMC Bioinformatics*. **6**(108).

The UniProt Consortium. 2017. UniProt: the universal protein knowledgebase. *Nucleic Acids Research*. **45**(D1), pp.D158-D169.

Vimberg, V., Tats, A., Remm, M. and Tenson, T. 2007. Translation initiation region sequence preferences in *Escherichia coli*. *BMC Molecular Biology*. **8**(100).

Zvelebil, M. and Baum, J. O. 2007. Understanding Bioinformatics. 1st ed. New York: Garland Science.

Appendix A
Formulae

Distance based on GC-content:

$$D_{GC} = \sqrt{(GC_i - GC_j)^2}$$

where, GC is the GC content for one genome, and *i* and *j* are the two genomes being compared.

Euclidean distance (applicable for nucleotide, dinucleotide, amino acid and diamino acid relative frequencies):

$$D = \sqrt{\sum_{k=1}^{l} (z_k^i - z_k^j)^2}$$

where, z is each element in the vector of the relative frequencies, *i* and *j* are the two genomes/proteomes being compared, and *l* the length of the vector (4 for nucleotide, 16 for dinucleotide, 20 for amino acid and 400 for diamino acid).

Distance based on cosine similarity (normalized dot product):

$$D(i,j) = \frac{z^i z^{j^{\top}}}{||z^i|| \, ||z^j||}$$

where, *z* is a vector of the relative frequencies and *i* and *j* are the two genomes/proteomes being compared.

Appendix B
Other Dendrograms Reconstructed using the Generated Distance Matrices

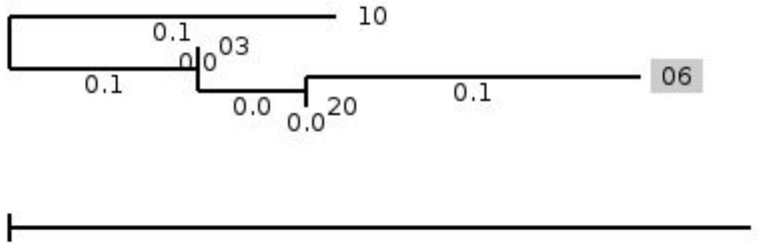


Figure B1: Whole genome dendrogram based on GC content.

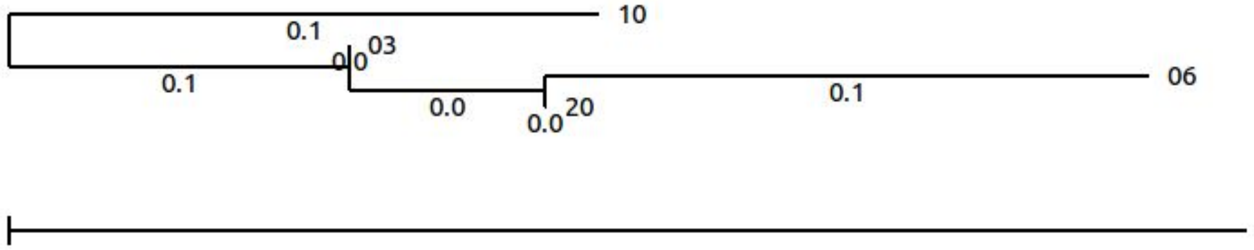


Figure B2: Whole genome dendrogram based on the nucleotide frequency.

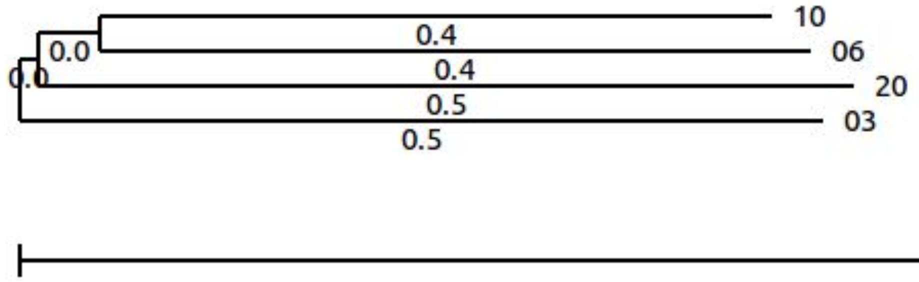


Figure B3: Whole genome dendrogram based on the cosine similarity of the nucleotide frequency.

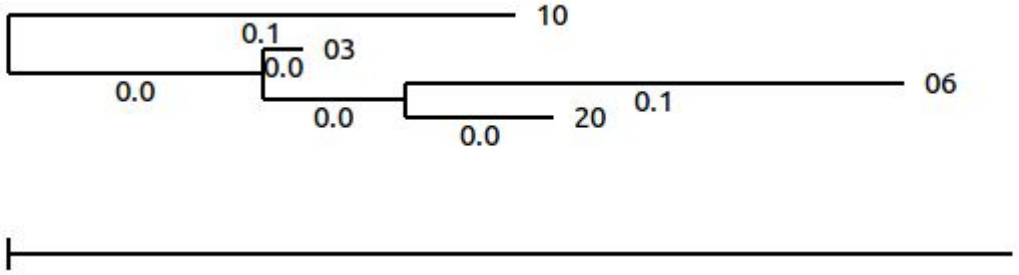


Figure B4: Whole genome dendrogram based on the dinucleotide frequency.

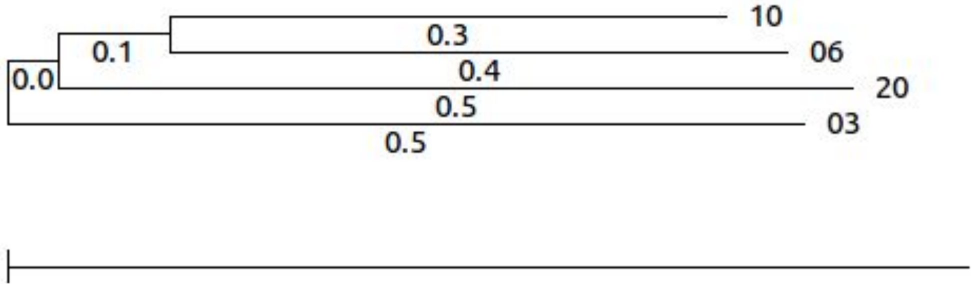


Figure B5: Whole genome dendrogram based on the cosine similarity of the dinucleotide frequency.