Final Project:

Comparative Genomic Analyses on Five Different Genomes

Group 2
Dimitri Wirjowerdojo & Elinor Löverli

Overview

- ➔ Results from practical (lab 1)
- ➔ ORF Finder
- ➔ Genome Statistics
- ➔ Tree Comparison

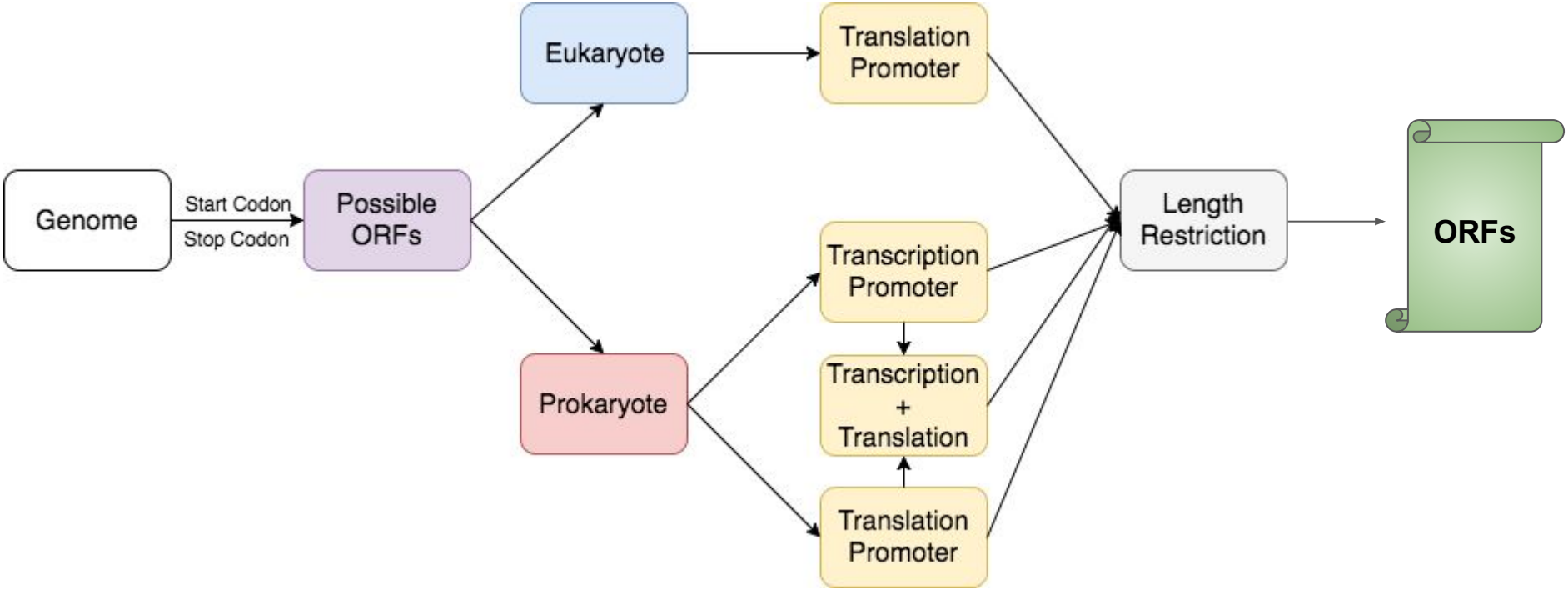
Practical results

Species	Kingdom	Number of genes	Size (Mb)	Fun fact
03. *Bacteroides thetaiotaomicron*, strain 7330	Bacteria	chromosomal genes 4 864	6.26	Circular genome, possesses plasmid with 38 genes
06. *Chloroflexus aurantiacus*, J-10-fl	Bacteria	8 184	5.2	Circular genome
10. *Fusobacterium nucleatum* subsp. polymorphum, strain ChDC F319	Bacteria	2 062	~2.3	Circular genome

Practical results

Species	Kingdom	Number of genes	Size (Mb)	Fun fact
20. *Thermotoga maritima*, strain Tma100	Bacteria	1 928	1.86	Circular genome
36. *Saccharomyces cerevisiae*, S288C chromosome XVI	Fungi	497	~0.96	Only chromosome 16 (linear)

ORF Finder Workflow



ORF Finder Result

Seq	Number of Genes (as reported by NCBI Taxonomy database)	Number of Open Reading Frames				
		NCBI ORFfinder	Glimmer/ GenScan¤	Transcription Only	Translation Only	Both
03	4903	56561	7035	12968	14140	3056
06	3991	63116	7041	6725	20282	2194
10	2130	11569	2755	4764	3110	1139
20	5805	17704	3137	3107	6165	1077
36	497	8029	338¤	-	1353	-

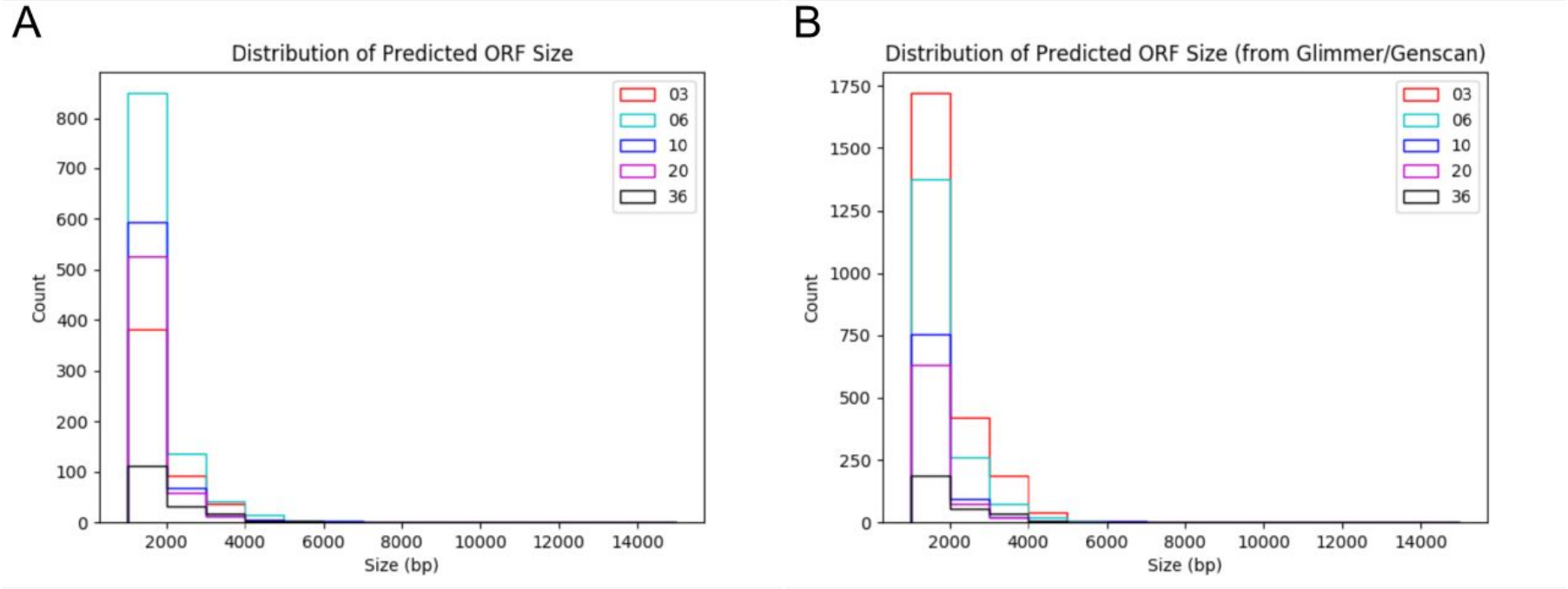
03 *B. thetaiotaomicron* | 06 *C. aurantiacus* | 10 *F. nucleatum* | 20 *T. maritima* | 36 *S. cerevisiae* chr. XVI

ORF Distribution

Seq	ORF Size (bp)	
	Max / Min	Median / Mean
03	4398 / 90	144 / 237
06	16641 / 90	165 / 290
10	13200 / 90	447 / 636
20	5073 / 90	180 / 376
36	7470 / 90	144 / 410

03 *B. thetaiotaomicron* | 06 *C. aurantiacus* | 10 *F. nucleatum* | 20 *T. maritima* | 36 *S. cerevisiae* chr. XVI

ORF Distribution



03 *B. thetaiotaomicron* | 06 *C. aurantiacus* | 10 *F. nucleatum* | 20 *T. maritima* | 36 *S. cerevisiae* chr. XVI

Blastp Result

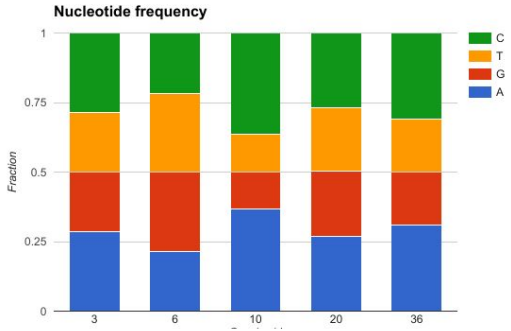
Default settings

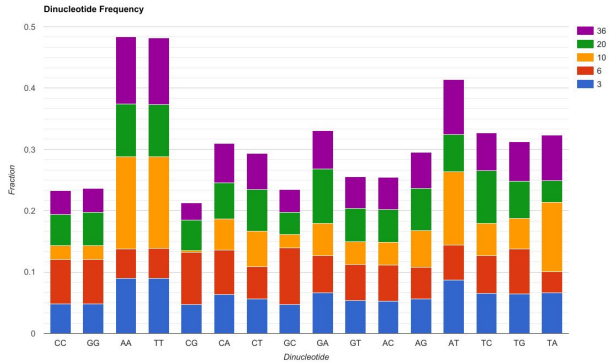
Against itself (reference proteome from UniProt)

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

03 *B. thetaiotaomicron* | 06 *C. aurantiacus* | 10 *F. nucleatum* | 20 *T. maritima* | 36 *S. cerevisiae* chr. XVI

Nucleotide and dinucleotide frequency



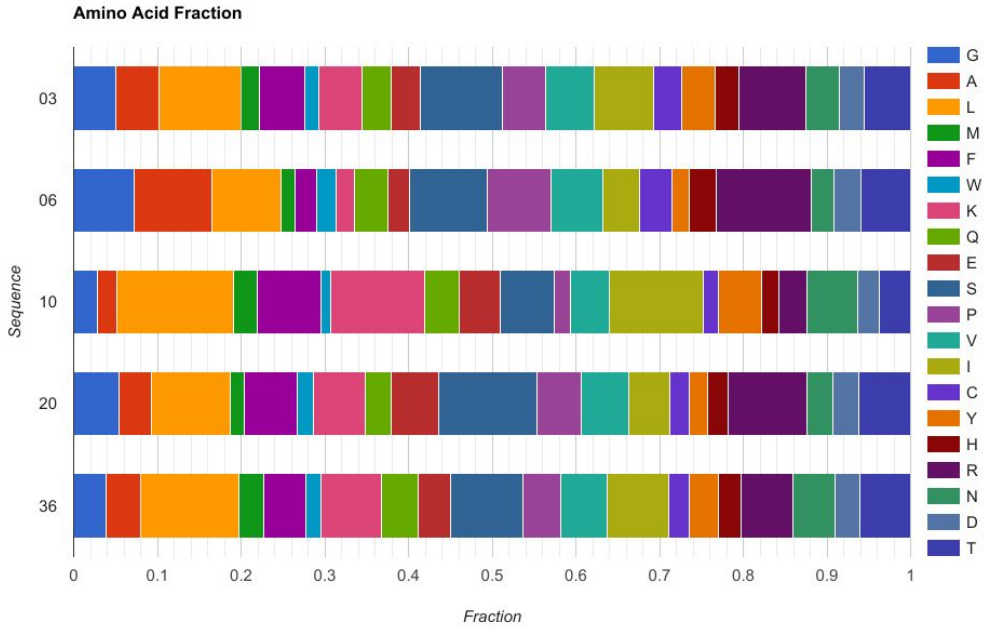


Seq.	03	06	10	20	36
GC Content	43%	57%	27%	46%	38%

- Purine and pyrimidine symmetric distribution
- AT combinations overrepresented
- GC combinations underrepresented
- Exception: Sequence 6

03 *B. thetaiotaomicron* | 06 *C. aurantiacus* | 10 *F. nucleatum* | 20 *T. maritima* | 36 *S. cerevisiae* chr. XVI

Amino acid frequency



High-frequency amino acids in all 5 genomes:

- Leucine
- Serine
- Isoleucine
- Arginine

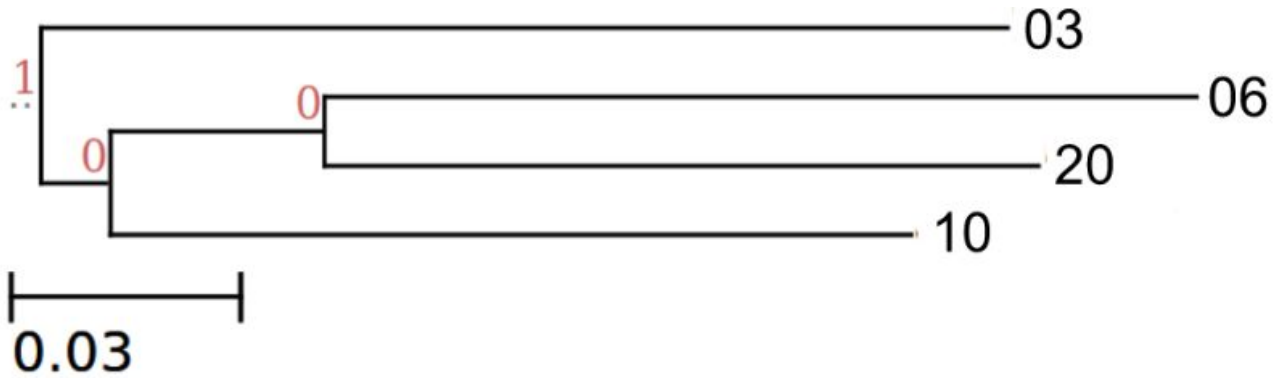
Low-frequency amino acids in all 5 genomes:

- Methionine
- Tryptophan
- Cysteine

03 *B. thetaiotaomicron* | 06 *C. aurantiacus* | 10 *F. nucleatum* | 20 *T. maritima* | 36 *S. cerevisiae* chr. XVI

Dendrograms vs. phylogenetic trees

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

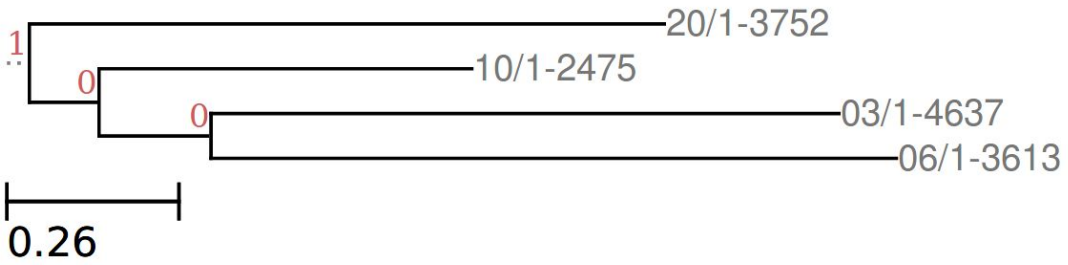


Phylogenetic tree reconstruction from 16S rRNA using neighbor-joining method.

03 *B. thetaiotaomicron* | 06 *C. aurantiacus* | 10 *F. nucleatum* | 20 *T. maritima* | 36 *S. cerevisiae* chr. XVI

Dendrograms vs. phylogenetic trees

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

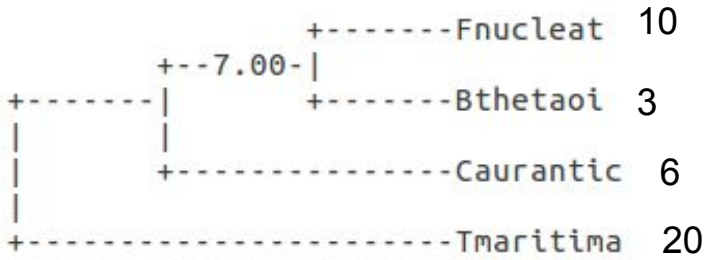


Phylogenetic tree reconstruction from orthologous cluster metagene.

03 *B. thetaiotaomicron* | 06 *C. aurantiacus* | 10 *F. nucleatum* | 20 *T. maritima* | 36 *S. cerevisiae* chr. XVI

Dendrograms vs. phylogenetic trees

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



Consensus tree from orthologous cluster genes, using phylip.

Conclusion

Identified species:

- 03. *Bacteroides thetaiotaomicron,*
- 06. *Chloroflexus aurantiacus,*
- 10. *Fusobacterium nucleatum,*
- 20. *Thermotoga maritima,*
- 36. *Saccharomyces cerevisiae*, chromosome XVI.

ORF:

- Significant difference to other methods due to fundamental difference in algorithm.
- The size of most ORFs peak at ~1kbp.
- True positive against itself through blastp ranges between 20%-72%
- True negative between 1.6%-6%

Trees:

- Tree from diamino acid frequency agrees most with metagene from orthologous cluster,
- No overall agreement found.