Results for HW1: Finding Long ORFs

Summary: All but one of our prokaryotes are bacteria, and only about 24% are pathogenic. The genome size and number of protein-coding genes range from 1.30 Mbp and 1192 genes (Candidatus Liberibacter solanacearum) to 7.60 Mbp and 7272 genes (Mesorhizobium loti, our community prokaryote). The GC content ranges from 34% (Nitrosopumilus maritimus) to 64% (both Thermanaerovibrio acidaminovorans and Thermomicrobium roseum). The average protein length is about 311 amino acids. The genome with greatest percent devoted to protein-coding genes, at 94%, is Sulfurimonas autotrophica, while the genome with the smallest coding percentage, 75%, is Haloquadratum walsbyi, our only Archaeon. Sn and sSn have good averages of 0.59 and 0.81, respectively, with relatively narrow ranges around these means. PPV and sPPV are much more variable: PPV ranges from 0.22 to 0.79 with an average of 0.59, and sPPV ranges from 0.42 to 0.98 with an average of 0.81. The gene prediction worked best -- considering both sensitivity and positive predictive value -- on the two Legionella bacteria, which both had sSn of 0.86 and sPPV of 0.96. The worst performance was on Mesorhizobium loti, also our largest prokaryote, with sSn of 0.83 and sPPV of 0.42.

Cells with this shading may not be correct, based on incorrect results turned in for the community prokaryote. These values have not been included in the summary statistics near the bottom of the spreadsheet.

Species	Arch / Bact	Pathogenic?	Habitat	Genome (in Mbp)	GC%	Proteins	Avg Protein Length	Coding Percent	Sn	sSn	FOR	PPV	sPPV	FDR
Aliivibrio salmonicida LFI1238	B	Pathogen causing Hitra disease in Atlantic salmon and rainbow trout	Aquatic	4.65	39	3915	308	78	0.6845	0.8208	0.1792	0.7362	0.8828	0.1172
Alteromonas macleodii str. 'Deep ecotype' chromosome	B	N\A	Marine, sea-water	4.45	45	4084	314	86	0.6486	0.8053	0.1947	0.7564	0.9392	0.0608
Brucella Ovis	B	Inflamation of the epididymis and placenta in sheep	HostAssociated Habitat (infects Sheep tissue)	3.26	57	2890	298	79	0.3571	0.6626	0.3374	0.2462	0.4569	0.5431
Candidatus Liberibacter solanacearum	B	Causes Zebra Chip disease of potato	Host Associated	1.30	35	1192	282	80	0.5206	0.7322	0.2678	0.6866	0.9657	0.0343
Clostridium phytofermentans ISDg uid58519	B	Not pathogenic	Forest soil (originally in Massachusetts)	4.85	35	3902	337	82	0.2596	0.6284	0.3716	0.0243	0.0587	0.9413
Cyanothece ATCC 51142	B	Non-Pathenogenic Nitrogen-Fixing	Aquatic	5.46	38	5304	297	87	0.5549	0.7294	0.2706	0.7454	0.9800	0.0200
Cyanothece PCC 7425	B	Not pathogenic to humans	Aquatic	5.79	51	5327	307	85	0.6270	0.7753	0.2247	0.6982	0.8633	0.1367
Escherichia coli BL21 Gold DE3 pLysS AG	B	N/A	multiple habitats	4.60	51	4228	310	86	0.6388	1.4664	0.1729	0.5688	1.3055	0.2632
Hahella chejuensis KCTC 2396	B	Not a known pathogen	Marine sediment	7.22	54	6773	313	88	0.6170	0.7914	0.2086	0.5162	0.6621	0.3379
Haloquadratum walsbyi DSM 16790	A	No	Aquatic	3.18	48	2647	298	75	0.1561	0.4977	0.5023	0.0157	0.0502	0.9498
Legionella pneumophila str. Lens	B	Pathogenic, Legionnaire's disease.	HostAssociated	3.40	38	2934	336	87	0.6970	0.8596	0.1404	0.7785	0.9600	0.0400
Legionella pneumophila str. Paris	B	Pathogen, Legionnaire's disease in humans (resulting pneumonia-like disease)	Multiple habitats, but usualy found growing inside other organisms such as protozoans in aquatic environments.	3.63	38	3166	333	87	0.6892	0.8585	0.1415	0.7743	0.9645	0.0355
Mesorhizobium loti MAFF303099	B	No, symbiotic nitrogen-fiuxing bacteria	Soil	7.60	63	7272	300	86	0.5256	0.8298	0.1702	0.2672	0.4218	0.5782
Neisseria lactamica 020 06	B	Commensal species	Host nasopharynx	2.22	52	1972	309	83	0.5984	0.8119	0.1881	0.3932	0.5335	0.4665
Nitrosopumilus maritimus SCM1	B	No	Aquatic	1.65	34	1796	274	90	0.5718	0.7539	0.2461	0.7253	0.9562	0.0438
Paenibacillus polymyxa E681	B	not pathogenic	Terrestrial	5.40	46	4805	320	86	0.5457	0.8248	0.1752	0.5201	0.7862	0.2138
Shewanella pealeana ATCC 700345	B	no	isolated from the accessory nidamental gland of the squid	5.17	45	4241	338	83	0.5223	0.8156	0.1844	0.6187	0.9662	0.0338
Sulfurimonas autotrophica DSM	B	No	Marine, Sediment	2.20	35	2158	312	94	0.6742	0.8434	0.1566	0.7543	0.9435	0.0565
Thermanaerovibrio acidaminovorans	B	non-pathogenic	multiple habitats	1.84	64	1738	326	92	0.4217	0.8653	0.1346	0.2560	0.5253	0.4746
Thermomicrobium roseum DSM 5159	B	no	Specialized	2.92	64	2859	303	89	0.4216	0.5998	0.4002	0.2249	0.3199	0.6801
Yersinia pseudotuberculosis YPIII	B	NA	NA	4.70	48	4192	312	84	0.6324	0.8056	0.1935	0.6953	0.8857	0.1143
Minimum				1.30	34	1192	274	75	0.4216	0.7294	0.1346	0.2249	0.4218	0.0200
Maximum				7.60	64	7272	338	94	0.6970	0.8653	0.2706	0.7785	0.9800	0.5782
Mean				4.07	47	3685	311	85	0.5898	0.8073	0.1914	0.5876	0.8131	0.1919