Abstract

This study investigates the evolutionary relationships and environmental preferences of five newly sequenced bacterial genomes (GXS004, GXS005, GXS006, GXS013, and GXS015) through gene annotation, metabolic pathway analysis, and clustering techniques. Open reading frames (ORFs) were identified and translated into amino acid sequences for functional annotation using BLASTp, UniProt, and KEGG. Then, evolutionary relationships were explored through BLASTn-based taxonomic identification and clustering analysis. Results show that GXS013 and GXS015 share multiple functional pathways, particularly in sulfur metabolism, indicating potential adaptation to sulfur-rich host-associated environments. GXS006, with enzymes related to osmotic stress, may be linked to marine environments. BLASTn-based results also infer that GXS006 is closely related to a marine family within the *Shewanella* genus. GXS004 and GXS005 exhibit functional distinctions, with no clear environmental associations identified. The neighbor joining tree, Non-metric Multidimensional Scaling (NMDS), and hierarchical clustering reveal that GXS013 and GXS015 are closely related genetically. Although GXS004 and GXS005 appear genetically similar to GXS013 and GXS015—GXS004 based on clustering analysis and GXS005 based on the neighbor-joining tree—both exhibit distinct functional profiles, suggesting separate evolutionary trajectories. The phylogenetic and clustering analyses suggest that functional similarities do not always align with evolutionary distance, pointing to the complex interplay between genomic evolution and environmental adaptation. Further research is required to refine taxonomic classifications and explore environmental relationships through additional pathway and gene-specific studies to validate these findings.

Methods

To identify genes and associated metabolic pathways from the assembled genomes, open reading frames (ORFs) were first located and then translated into amino acid sequences. This was because most downstream gene annotation tools (e.g., BLASTp, UniProt) operate on protein sequences. Firstly, the ORF search and threshold estimation implementation discussed in class was used with slight modifications (Bromberg, 2025). Since the provided genomes consisted of multiple scaffolds concatenated together with palindromic markers ("NNNNNACTGNNNGTCANNNNN"), ORFs that contained the markers were excluded to avoid generating non-biological or chimeric coding regions. For each genome, a statistical threshold for ORF length was estimated by randomly shuffling the input sequence 100 times and finding the ORFs within each shuffle. From each iteration, the minimum threshold length corresponding to a significance level of $\alpha = 0.05$ was recorded. The final ORF length threshold was set as the median of these 100 minimum lengths, which balances the risk of false positives and ensures biological relevance. ORFs above this threshold were extracted in both forward and reverse strands and saved in a FASTA format.

To identify gene functions, the nucleotide ORFs were translated to amino acid sequences and queried against the UniProtKB/Swiss-Prot database using BLASTp (Camacho et al., 2023). Other parameters were default. For genomes GXS006 and GXS013, sequences were split into two files and two BLASTp runs each to accommodate upload size constraints on the BLASTp web interface. From the BLASTp outputs, only hits with e-values ≤ 1e-50 were retained to ensure high confidence. UniProt accessions were extracted (with version suffixes

removed) and submitted to the UniProt ID mapping service to retrieve gene names, EC numbers, and KEGG pathway annotations (Bateman et al., 2024). EC numbers were then mapped to metabolic pathways using the KEGG Mapper with the default settings (Minoru Kanehisa et al., 2024). Common and unique EC numbers and genes were compared between genomes. Specific pathways and genes related to an environmental preference were identified through a literature review.

To investigate evolutionary relationships, two complementary approaches were used: multiple sequence alignment (MSA) and sequence similarity-based clustering. For alignment-based inference, all concatenation markers were first removed to prevent artifacts during MSA. MSA was then performed using the MAFFT online version with default settings except for selecting “Leave gappy regions” (Rozewicki et al., 2019). A neighbor joining tree was constructed with the default settings. (Neighbor joining was performed because the genomes do not have strong similarity.) Secondly, the full genome sequences (with markers retained) were queried using BLASTn (Core Nucleotide database, default settings) to identify likely species or close relatives (Camacho et al., 2023). Hits with e-values ≤ 1e-50 were retained, and pairwise similarity was computed using the formula: similarity = (% identity / 100) × alignment length. The resulting similarity scores were normalized using z-scores to allow cross-comparison across genomes. The normalized scores were merged based on the “Scientific Name” column (species names), with unmatched values retained. For any duplicate species names, only the first occurrence was kept. To assess clustering based on species genomic similarity, both Non-Metric Multidimensional Scaling (NMDS) and clustering techniques were applied with random state = 1. Several distance metrics (Euclidean, Cosine, Pearson) and clustering algorithms (K-means, Hierarchical) were tested (scikit-learn, 2025; scikit-learn, 2019; scikit-learn, n.d.). The best-performing method was selected using silhouette scores, which quantify how well each genome fits within its assigned cluster relative to others (scikit-learn, 2019b).

Results/Discussion

The estimated ORF length thresholds are 190, 174, 258, 192, and 222 for GXS004, GXS005, GXS006, GXS013, and GXS015, respectively. From the UniProt ID Mapping results, only one EC number is returned for each of GXS004 (EC 3.6.5.3) and GXS005 (EC 2.7.3.9). According to the information on KEGG, EC 3.6.5.3 encodes a GTPase enzyme critical for protein synthesis by driving ribosomal movement during translation. EC 2.7.3.9 is driving ribosomal movement during translation. However, neither enzyme mapped to any known KEGG pathway, so no additional environmental inference could be made.

In contrast, GXS006, GXS013, and GXS015 yield substantial numbers of EC numbers and could be mapped to multiple metabolic pathways. Shared EC numbers provide some insights into relationships: GXS006 shared three EC numbers with both GXS013 and GXS015, while GXS013 and GXS015 shared ten EC numbers. Notably, GXS013 included all pathways represented in GXS015 — cysteine and methionine metabolism, selenocompound metabolism, sulfur metabolism, and secondary metabolite biosynthesis — suggesting both functional overlap and potential evolutionary relatedness. However, GXS013 appears functionally broader, involved in a wider range of core and auxiliary metabolic processes, particularly carbohydrate and sulfur-associated pathways. GXS015 was more narrowly focused, with a stronger emphasis

on sulfur-containing amino acid biosynthesis. The shared pathways suggest that both bacteria are adapted to sulfur-rich environments and may rely on sulfur-containing compounds such as methionine, cysteine, and possibly selenoproteins. The presence of secondary metabolite biosynthesis pathways further suggests that both organisms might inhabit variable or stressful environments requiring secondary metabolites (Bin, Huang and Zhou, 2017). To further explore the relatedness of these genomes, gene-level comparisons were performed. Only GXS013 and GXS015 shared a substantial number of genes (n = 23), reinforcing the hypothesis of evolutionary relatedness.

For GXS006, while none of the pathways are directly exclusive to marine microbes, the presence of choline oxidase (EC 1.1.3.17) and betaine-aldehyde dehydrogenase (EC 1.2.1.8) is functionally suggestive. These enzymes are key in the conversion of choline to glycine betaine, a well-known osmoprotectant. Accumulation of glycine betaine is a hallmark adaptation of marine and halotolerant bacteria to osmotic stress (Kiene, 1998; Boch et al., 1997). However, further direct evidence from genome taxonomy is needed to confirm this.

To further explore evolutionary relationships among the genomes, taxonomic identification was pursued using BLASTn. The top three BLASTn hits for each genome (all with e-values < 1e-50) were compiled into a comparative table (see Table 1). From this, GXS006 appears most closely related to the Shewanella genus, a member of the marine-associated Shewanellaceae family. This supports previous functional evidence for osmotic stress adaptations (Janda, 2014). For the other genomes, potential affiliations include Candidatus taxa (GXS004), Mycoplasma (GXS005), and Butyrivibrio (GXS013 and GXS015), the latter of which are commonly found in gastrointestinal microbiomes (Rodríguez Hernáez et al., 2018). However, the low query coverages (generally <50%) for these matches reduce confidence in their taxonomic assignments.

Genome	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. ident	Acc. Len	Accession
GXS004	MAG: Candidatus Tyloplasma litorale Fukuoka2020 DNA, complete genome	Candidatus Tyloplasma litorale	1050	1810	42%	0	81.39	615622	AP027078.1
	MAG: Candidatus Hepatoplasma vulgare Av-JP DNA, complete genome	Candidatus Hepatoplasma vulgare	854	1332	41%	0	79.82	662108	AP027131.1
	MAG: Candidatus Stammera capleta isolate 1 chromosome, complete genome	Candidatus Stammera capleta	505	505	18%	4.00E-137	75.4	281294	CP043983.1
GXS005	Metamycoplasma sualvi strain Mayfield chromosome, complete genome	Metamycoplasma sualvi	518	518	11%	1.00E-140	74.69	842965	CP174167.1
	Candidatus Mycoplasma mahonii isolate OP267 chromosome, complete genome	Candidatus Mycoplasma mahonii	505	740	24%	8.00E-137	72.01	796768	CP114583.1
	Mesomycoplasma lagogenitalium strain 12MS chromosome, complete genome	Mesomycoplasma lagogenitalium	462	462	7%	5.00E-124	77.82	973223	CP122979.1
GXS006	Shewanella psychropiezotolerans strain YLB-06 chromosome, complete genome	Shewanella psychropiezotolerans	17084	68411	85%	0	90.21	6449204	CP041614.1
	Shewanella psychrophila strain WP2 chromosome, complete genome	Shewanella psychrophila	13404	49584	76%	0	87.29	6353406	CP014782.1
	Shewanella violacea DSS12 DNA, complete genome	Shewanella violacea DSS12	12152	57210	81%	0	88.87	4962103	AP011177.1
GXS013	Butyrivibrio proteoclasticus B316 chromosome 1, complete sequence	Butyrivibrio proteoclasticus B316	1319	3253	6%	0	78	3554804	CP001810.1
	Butyrivibrio fibrisolvens strain D1 chromosome, complete genome	Butyrivibrio fibrisolvens	1303	1303	5%	0	77.69	4671138	CP146963.1
	Butyrivibrio fibrisolvens strain ASCUSDY19 chromosome 1, complete sequence	Butyrivibrio fibrisolvens	1293	1505	5%	0	77.62	4116214	CP065800.1
GXS015	Butyrivibrio sp. JL13D10 chromosome 1, complete sequence	Butyrivibrio sp. JL13D10	1271	4688	10%	0	83.13	3687363	CP172573.1
	Butyrivibrio hungatei strain MB2003 chromosome I, complete sequence	Butyrivibrio hungatei	941	1143	6%	0	80.73	3143784	CP017831.1
	Treponema rectale strain CHPA chromosome, complete genome	Treponema rectale	684	684	4%	0	76.74	2864011	CP031517.1

Table 1. Top BLASTn Hits and Taxonomic Assignments for Each Genome.

The neighbor-joining tree shown in Figure 1 indicates that GXS005, GXS013, and GXS015 cluster together. The grouping of GXS013 and GXS015 is consistent with previous functional and genetic analyses. However, the close association of GXS005 with GXS013 is unexpected, and its underlying basis remains unclear. In addition, clustering analysis was conducted on the normalized similarity scores derived from BLASTn alignments. From the several different combinations of distance metrics and clustering methods, the highest silhouette score (0.1184) was obtained using hierarchical clustering with Euclidean distance and three clusters. Visualization of this result is presented in Figure 2. Hierarchical clustering with Euclidean distance and three clusters yields the highest silhouette score of 0.1184. The result is visualized below. The clustering analysis shows that GXS013 and GXS015 group closely together, supporting earlier findings of functional and genetic similarity. Interestingly, GXS004

also falls within this cluster and is positioned relatively close to GXS013 and GXS015, which contrasts with prior pathway and gene analysis. This may suggest a weak evolutionary signal or the presence of distant shared ancestry not captured by metabolic function alone. Additionally, GXS006 belongs to a separate cluster and does not share common genes with any of the other genomes, aligning with previous results that indicated a distinct functional profile. GXS005 is quite distinct from the other samples, suggesting it has a significantly different composition or characteristics compared to the rest.

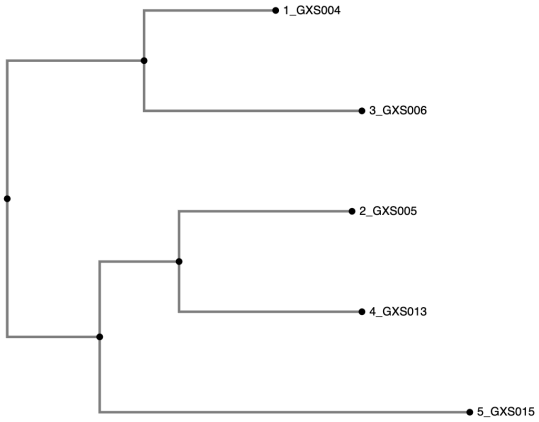

Figure 1: Neighbor Joining Tree constructed using MAFFT.

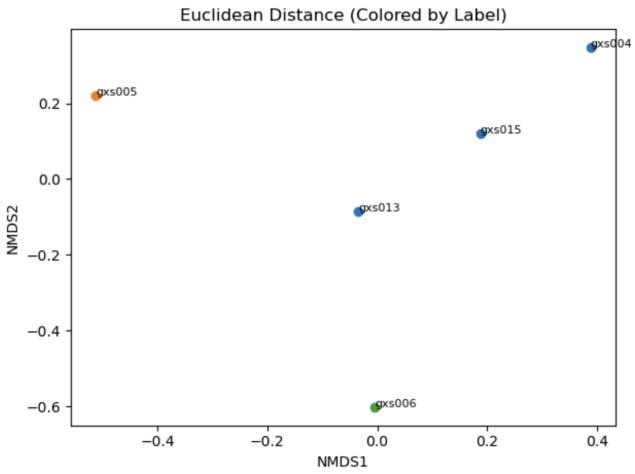

Figure 2: Hierarchical Clustering of Genomes Based on Normalized Similarity Scores (calculated as similarity = (% identity / 100) × alignment length).

Conclusion

Among the five bacterial genomes assigned, GXS013 and GXS015 exhibit the strongest functional overlap, particularly in sulfur-related pathways, suggesting both species may be adapted to sulfur-rich environments. GXS006 showed potential adaptations to osmotic stress, likely linking it to marine environments, though further taxonomic verification is needed. GXS004 and GXS005 appeared functionally distinct, with no clear environmental preference identified. According to the taxonomic analysis and NMDS clustering, GXS013 and GXS015 were shown to be closely related, supporting the hypothesis of evolutionary similarity between them. The neighbor joining tree and hierarchical clustering further reinforced this relationship, as these genomes clustered together based on genomic similarity. Given their functional overlap and genetic relatedness, GXS013 and GXS015 are likely to be host-associated, potentially inhabiting environments with sulfur-rich compounds or specialized niches. Future studies could focus on refining taxonomic assignments and exploring environmental context through additional pathway and gene-specific analyses to confirm or refine these conclusions, particularly with a broader sample size and enhanced functional annotations.

References

Bateman, A., Martin, M.-J., Orchard, S., Magrane, M., Adesina, A., Ahmad, S., Bowler-Barnett,

E.H., Hema Bye-A-Jee, Carpentier, D., Denny, P., Fan, J., Garmiri, P., Jose, L., Hussein, A., Alexandr Ignatchenko, Insana, G., Rizwan Ishtiaq, Joshi, V., Dushyanth Jyothi and Swaathi Kandasaamy (2024). UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids Research*, [online] 53(D1). doi:https://doi.org/10.1093/nar/gkae1010.

Bin, P., Huang, R. and Zhou, X. (2017). Oxidation Resistance of the Sulfur Amino Acids: Methionine and Cysteine. *BioMed Research International*, 2017, pp.1–6. doi:https://doi.org/10.1155/2017/9584932.

Boch, J., Nau-Wagner, G., Kneip, S. and Bremer, E. (1997). Glycine betaine aldehyde dehydrogenase from Bacillus subtilis : characterization of an enzyme required for the synthesis of the osmoprotectant glycine betaine. *Archives of Microbiology*, 168(4), pp.282–289. doi:https://doi.org/10.1007/s002030050500.

Bromberg, Y. (2025) *CourseLecture2_gene_finding*. [PowerPoint presentation]. CS485: Bioinformatics, taught at Emory University, 21 January.

Camacho, C., Boratyn, G.M., Joukov, V., Roberto Vera Alvarez and Madden, T.L. (2023). ElasticBLAST: accelerating sequence search via cloud computing. *BMC Bioinformatics*, 24(1). doi:https://doi.org/10.1186/s12859-023-05245-9.

Janda, J.M. (2014). Shewanella: a Marine Pathogen as an Emerging Cause of Human Disease. *Clinical Microbiology Newsletter*, [online] 36(4), pp.25–29. doi:https://doi.org/10.1016/j.clinmicnews.2014.01.006.

Kiene, R.P. (1998). Uptake of Choline and Its Conversion to Glycine Betaine by Bacteria in Estuarine Waters. *Applied and Environmental Microbiology*, [online] 64(3), pp.1045–1051. doi:https://doi.org/10.1128/aem.64.3.1045-1051.1998.

Minoru Kanehisa, Miho Furumichi, Sato, Y., Matsuura, Y. and Ishiguro-Watanabe, M. (2024). KEGG: biological systems database as a model of the real world. *Nucleic Acids Research*, 53(D1). doi:https://doi.org/10.1093/nar/gkae909.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J. and Higgins, D.G. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics (Oxford, England)*, 23(21), pp.2947–8. doi:https://doi.org/10.1093/bioinformatics/btm404.

Rodríguez Hernáez, J., Cerón Cucchi, M.E., Cravero, S., Martinez, M.C., Gonzalez, S., Puebla, A., Dopazo, J., Farber, M., Paniego, N. and Rivarola, M. (2018). The first complete genomic structure of Butyrivibrio fibrisolvens and its chromid. *Microbial Genomics*, [online] 4(10). doi:https://doi.org/10.1099/mgen.0.000216.

scikit-learn (2019a). *sklearn.cluster.KMeans — scikit-learn 0.21.3 documentation*. [online] Scikit-learn.org. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html.

scikit-learn. (n.d.). *sklearn.cluster.AgglomerativeClustering*. [online] Available at: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html.

Scikit-learn (2019b). *sklearn.metrics.silhouette_score — scikit-learn 0.21.3 documentation*. [online] Scikit-learn.org. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html.

scikit-learn. (2025). MDS. [online] Available at: https://scikit-learn.org/stable/modules/generated/sklearn.manifold.MDS.html.