Title: Functional Diversification of Gene Duplicates under the Constraint of Protein Structure

Authors: Kangning Guo[1,2]†*, Yuqing Yang[2,4]†, Tingting You[2,4], Kangli Zhu[1,2,3]*

Affiliations:

[1] Key Laboratory of Growth Regulation and Translational Research of Zhejiang Province, School of Life Sciences, Westlake University, Hangzhou 310030, China

[2] Institute of Biology, Westlake Institute for Advanced Study, Hangzhou 310024, China

[3] Westlake Laboratory of Life Sciences and Biomedicine, Hangzhou 310030, China

[4] Fudan University, Shanghai 200433, China

† These authors made equal contribution to this work

*Corresponding author. Email: guokangning@westlake.edu.cn & zhukangli@westlake.edu.cn

Abstract:

Gene differentiation subsequent to duplication plays a pivotal role in the evolutionary process, facilitating the generation of new functional genes. This dynamic process involves alterations in gene characteristics including protein structure, expression patterns, cellular localization, and enzymatic activity, which collectively contribute to the advent of novel functions. However, there are still relatively few reports on how multiple genes differentiate after duplication. Here we show the functional diversification of Glycoside Hydrolase Family 50 (GH50) agarases in the deep-sea bacterium *Agarivorans ablus* JK6, uncovering the evolutionary mechanisms behind the emergence of new enzymatic functions post-duplication. By leveraging AlphaFold for precise protein structure prediction and integrating phylogenetic analysis with experimental methods, we shed light on the structural adaptations and sequence alterations driving these enzymes' functional differentiation. Our findings establish a distinct link between structural changes and the creation of new hydrolysis products, underscoring the significance of gene duplication in broadening the enzymatic repertoire of organisms. Furthermore, the study elucidates how these adaptations contribute to the ecological success of *A.ablus* JK6, highlighting the role of gene duplication in expanding the enzymatic toolkit for environmental adaptation. Through exhaustive biochemical characterization and enzymatic activity assays, we furnish insights into the agarases' adaptive evolution, providing a thorough perspective on how genetic and structural elements contribute to the biochemical diversity and ecological prowess of marine bacteria, and also suggests new pathways for the development of bioengineered enzymes with enhanced functionalities.

Main Text:

Gene duplication events, recorgnized since Ohno's seminal work in 1970[1] as key catalysts for functional innovation and biodiversity, play a crucial role in augmenting genomic complexity and facilitating the emergence of novel gene functions[2-5]. This evolutionary mechanism allows duplicated genes to diverge through functional degradation, subfunctionalization, or neofunctionalization, thereby promoting adaptive evolution in response to environmental pressures[6-9]. Despite the recognized significance of gene duplication in evolutionary biology, numerous questions persist regarding the mechanisms through which these genes attain functional differentiation and achieve regulatory refinement within specific ecological niches and intricate biological networks[10].

Microorganisms, particularly bacteria, exhibit remarkable adaptability and diversity[11], evolving complex enzymatic systems for specialized substrates via gene duplication and subsequent differentiation[12,13]. This adaptive strategy enables them to navigate environmental shifts and competitive pressures. Agarases, a crucial class of glycoside hydrolases[14] that break down agar polysaccharides[15] in seaweed are instrumental for marine bacteria in polysaccharide metabolism and the carbon cycle[16]. Investigating these enzymes sheds light on the evolutionary and functional diversification mechanisms, enriching our understanding of microbial metabolic optimization and survival tactics in distinct ecosystems.

We conducted an in-depth investigation of the JK6 strain from the unique environment of the deep sea to elucidate the evolution and functional differentiation of four agarase proteins (PL3506, PL3511, PL3512, and PL3518), by comparing their activity differences, diversification of cellular localization, and changes in expression patterns. Utilizing AlphaFold for high-accuracy predictions of protein structures, we examined structural changes and constraints during the differentiation process, achieving efficient and accurate predictions and modifications of protein activities. Through a comprehensive analysis of the agarase system in the JK6 strain, we demonstrated how simple adaptive processes evolve into complex networks of ecological interactions[17], unveiling potential pathways for functional diversification following gene duplication. Moreover, our findings reveal bacterial adaptation mechanisms to environmental stresses through gene regulation and enzymatic system refinement, offering new insights into the microbial role in the global carbon cycle and advancing the development of biotechnological applications utilizing agarases and other polysaccharide-degrading enzymes.

Physiological Characterization and Genome Annotation of the Strain JK6

The strain designated JK6, isolated in the present investigation, originates from a deep-sea water sample collected in the vicinity of the Mariana Trench. On L1 agar plates, the JK6 strain exhibited significant hydrolytic activity, forming clear halos indicative of agar degradation (fig. S1). JK6 was identified as a strain of *Agarivorans ablus*. It has a genome size of 4.7 Mb and contains a 4,463 bp plasmid. Annotation identified 4,322 coding genes, 1,152 enzymes, and 280 transport proteins, which map to 282 known metabolic pathways. Positive selection analysis of single-copy orthologous genes identified 53 genes under positive selection (Data S1), mainly

involved in functions such as flagellar movement and ion transport, consistent with characteristics reported for deep-sea bacteria[18].

Gene Duplication and Divergence

We utilized OrthoFinder to analyze the complete protein sequence files of over 600 bacterial strains genomes (Data S2), identifying a set of four agarases belonging to the GH50 family[14], which were classified in the same Orthogroups. Through phylogenetic tree analysis of GH50 family proteins and structural similarity clustering of proteins containing the major glycoside hydrolase domains, we confirmed their evolutionary relationships (Fig. 1a). To establish the relationships among these four proteins, the distant agarase PL2878 was used as a common reference point.

Although gene family expansion in prokaryotes is often considered to be driven by horizontal gene transfer[19], our study has ruled out the possibility for the four agarases. We speculate that these proteins were retained after multiple duplications from a common ancestor and are distributed within a gene cluster spanning approximately 23kb (fig. S4). Specifically, the genes *PL3511* and *PL3512* still retain a tandem duplication pattern adjacent to each other. Considering that tandem duplications are usually highly unstable and without the action of natural selection, the amplified gene arrays would rapidly disappear from the population[2,20,21]. Therefore, natural selection may have occurred during the fixation process of genes *PL3511* and *PL3512*.

After analyzing the complete protein sequences of 628 bacterial strains, we identified variations in the number of homologs for four proteins across different clades, reflecting their evolutionary relationships (fig. S3). Homologous genes of *PL3506* and *PL3512* were predominantly distributed across distinct clades and rarely coexisted within the same species. Phylogenetic analysis indicated that the oldest gene, *PL3511*, had the fewest homologs, while the transitional gene *PL3518* had more but still significantly fewer than *PL3506* and *PL3512*, suggesting potential gene loss. Notably, in the genus *Vibrio*, horizontal gene transfer was observed to be responsible for acquiring this gene group. Functional analysis unveiled distinct agarose degradation patterns: NA2 (neoagarobiose) produced by PL3506, NA4 (neoagarotetraose) produced by PL3512, and various products including NA2, NA4 as well as small amounts of NA6 (neoagarohexaose) and NA8 (neoagarooctaose) produced by PL3511 (Fig. 1B). Over time, the main products of PL3511 and PL3518 shifted primarily to NA2. However, there are significant differences in the rate at which PL3511 and PL3518 break down NA4 (Fig. 1C, Fig. 2C). Structural analysis, favoring Gene3D over Pfam for its accuracy[22] (fig. S2), showed that all proteins contained a glycoside hydrolase domain at the C-terminus (Movie S1) with variations in the number of CBM-like domains at the N-terminus[23] (Movie S2). This structure combination pattern, particularly rare in the GH50 family (fig. S2), indicates an adaptation to oligotrophic environments and underscores both the evolutionary and functional diversification exhibited by these agarases[24].

In summary, these four paralogs have undergone different degrees of functional differentiation in enzyme activity. PL3511 likely retains the most diverse activity closest to the original enzyme,

while PL3506 and PL3512 exhibit subfunctionalization with specific differentiation, and PL3518 is in an intermediate transitional state.

Modifications in Activity within the Constraints of Protein Structural

The interaction between structure and function in the evolution of proteins exhibits a high degree of complexity[25,26]. Typically, the conservation of amino acid sequences is crucial for maintaining the specific three-dimensional structure of a protein, which plays a key role in its biological function[27-30]. As a result, although mutations may occur in the evolutionary process, the conservation at key positions ensures the stability of protein structure and function. This delicate equilibrium reveals that sequence variation drives protein evolution while being constrained by structural and functional limitations[31], reflecting the complexity of biomolecular adaptability and evolutionary dynamics.

In this study, AlphaFold was used to predict the structures of target proteins, revealing that all four agarases displayed mixed folding characteristics. At the enzyme's N-terminus, one to two β-sandwich domains (CBM-like domains) were included, while the C-terminus exhibited a sophisticated 8 (α/β) barrel (glycoside hydrolase domain). Two CBM domains were joined by a linker region of approximately 22 amino acids, with the CBM domain adjacent to the glycoside hydrolase domain located at the (α/β)8 barrel (Fig. 1D).

Unlike the common observation that major mutations occur in the non-active center of enzymes[28], we found that despite of significant divergent variations in DNA and protein sequences, the main differences between the structures of the four agarases were located at the active center, while the overall protein backbone remained highly conserved structurally (Movie S1). As shown in Data S3, pairwise comparisons of the glycoside hydrolase domains of the four proteins showed RMSD values all below 1.9, with average TM scores above 0.9. However, this structural conservatism contrasted sharply with the similarity of protein amino acid sequences (below 50%). Even when comparisons were extended to proteins of different genera, a similar pattern was observed. Pairwise ka/ks values（far below 0.5, Data S4) indicated significant purifying selection on the overall sequences. Considering the multiple influences of genetic code structure and amino acid physicochemical properties on protein evolution, we calculated the ratio of amino acid sites differing by more than 1Å to those below this threshold (named the structural conservation index) to characterize protein structural conservatism, and the SCI values among the four proteins range from -0.132 to 0.349(Data S5), indicating a more conservative structural tendency.

In comparative studies focusing on the active centers of enzymes and the differences in their enzymatic products, targeted modification experiments were conducted. The removal of a specific loop in PL3506, which produces NA2, and PL3512, which yields NA4, led to a significant decrease in the rate of NA4 cleavage by PL3506, yet the agarose cleavage product types remained the same, producing only NA2 as expected. Similarly, eliminating the excess loop in PL3518 also reduced its NA4 cleavage rate, resulting in an accumulation of NA4 during agarose cleavage, suggesting the loop structure may only influence the binding or release of

polysaccharides with agarase, thereby affecting enzyme activity towards different products (Fig. 2). Additionally, removing the CBM domain in PL3506 significantly reduced the hydrolysis rate of NA4 and altered the agarose hydrolysis products, indicating the CBM domain's crucial role in supporting the active center and enhancing polysaccharide binding (fig. S5A). Experiments on completely conserved amino acid sites within the glycoside hydrolase domain showed that mutating Glu-499 to Gly near PL3506's active center (Movie S3) changed the enzyme's activity, highlighting the importance of these conserved sites for protein function. Furthermore, synthetic expression experiments on four homologous proteins from different genera in a *Vibrio astriarenae* strain (HN897)[32] demonstrated NA2 hydrolytic activity without the need for accumulating large amounts of NA2 during agarose hydrolysis. Vsp11, Vsp12, and Vsp18 (named Vsp6, Vsp11, Vsp12 and Vsp18 respectively, based on the homology with the four agarase proteins in JK6) also produced specific monosaccharide products, not derived from hydrolyzing NA2, likely because they cut the outermost recognition units in the agarose chain (fig. S5B).

Our study indicates that within the constraints of structure and function[26], this group of proteins underwent adaptive mutations in smaller active center regions while maintaining the basic protein structural framework.

Changes in Subcellular protein distribution

The retention of duplicated proteins confers significant biological advantages, notably in subcellular localization differentiation, enhancing precise regulatory expression and operational efficiency[33]. This study aimed to elucidate the complex pathway networks formed through physical localization differentiation by speculating on the subcellular localization of four agarases using protein localization prediction, analysis of intracellular and extracellular protein abundance under various culture conditions, and the expression localization of complete proteins in Escherichia coli (Fig. 3A/C, Data S6). Our findings suggest that PL3511 functions as an extracellular secretory protein, PL3506 is primarily intracellular, PL3512 serves as an interstitial protein, and PL3518, despite having extracellular secretion characteristics, predominantly accumulates inside the cell. The intracellular to extracellular accumulation ratio notably varies under different culture conditions, particularly with agarose induction, influencing PL3511 and PL3518's extracellular secretion proportions differently, according to enzyme activity assays, PL3518 has a faster ability to cut NA4 into NA2, such regulation benefits the bacterial cells in rapidly degrading abundant agarose into NA2. The enzyme localization directly impacts the activity differentiation among the four agarases, with extracellular agarases like PL3511 decomposing agar-containing polysaccharides into longer chains outside the cell, which are then transported and hydrolyzed into NA4 by PL3512 in the interstitial space, subsequently broken down into NA2 by intracellular PL3506, and finally metabolized. This study underscores the pivotal role of protein subcellular localization in functional differentiation within complex protein families, offering fresh perspectives on protein precision in biological functions and contributing valuable insights into protein evolutionary dynamics.

Ancestral Sequence Reconstruction Analysis

Reconstructing ancestral enzymes to explore the evolutionary mechanisms and fates of duplicated genes is a viable strategy[34,35]. Our study reconstructed the ancestral sequences of four agarases based on the GH50 gene family tree, analyzing their evolutionary history by comparing tree topological structures (fig. S7). The examination of subcellular localization predictions for ancestral protein sequences at various nodes revealed that ancestors related to PL3511, PL3512, and PL3518 largely retained the same localization as their contemporary counterparts, whereas the ancestors of PL3506 exhibited distinct localization patterns (Data S6). Our analysis suggests that subcellular adaptability is a key mechanism for the functional diversification of duplicated genes, evidenced by the absence of intracellular localization predictions for PL3506 ancestors, indicating that functional diversification may precede localization differentiation.

The study highlights a complex interplay between subcellular localization and functional characteristics of proteins within the GH50 family, pointing out that the evolution of localization and function might not always be synchronous. These insights are crucial for understanding protein adaptation to different subcellular environments and their functional differentiation throughout evolution.

Changes in Gene Expression Regulation

In the process of functional differentiation following protein duplication, the differentiation of gene regulation is a key mechanism[2,36]. Our study reveals distinct responses and expression patterns of these four agarase genes in different sugar environments. This differential expression suggests that bacteria employed a more efficient protein hydrolysis method to acquire nutrients under limited resources, thereby gaining a competitive advantage in survival.

We employed qPCR technology to assess the expression of these four proteins under various sugar and concentration conditions (Fig. 3B). The results demonstrated that these proteins exhibit distinct expression patterns in response to different sugars or sugar concentrations, especially under the induction of NA2, NA4, and oligosaccharides. The PL3511 protein quickly reached a transcriptional peak, while PL3512 and PL3518 showed differences in response to NA2 and NA4. Specifically, PL3512 rapidly reached its peak under NA4 induction but was slower in NA2; conversely, PL3518 responded quickly under NA2 induction and relatively slower under NA4, indicating that in the initial induction, the corresponding product of the enzyme has a negatively feedback function to induce enzyme transcription. Additionally, PL3506 exhibited the fastest transcriptional response under higher concentrations of oligosaccharides.

Covarying Evolutionary Rate Analysis

In studying the co-evolutionary history between genes, we calculated the co-evolutionary rates[37] of gene pairs to reveal their co-variation characteristics (Data S7, Fig. S8). Within the set threshold range, we found that *PL3511* does not exhibit significant correlation with the other three agarase genes (*PL3506*, *PL3512*, and *PL3518*). This may indicate that PL3511 has greater functional and evolutionary independence from the other three enzymes. Considering the evolutionary rates[38] and the average pairwise identity[39] between sequences, there are significant differences between PL3511, PL3512, PL3518, and PL3506. The evolutionary rates from low to

high are PL3506, PL3518, PL3512, PL3511, while the average pairwise identity mean values from high to low are PL3512, PL3518, PL3506, PL3511. These results suggest that PL3511 differentiated earliest from the other three agarase genes, with PL3518 possibly in an intermediate transitional state between PL3506 and PL3512. Although the other differentiation processes among these genes are not entirely clear, it can be inferred that they have undergone complex pathways of functional and evolutionary differentiation. This not only reveals their unique roles in metabolic and physiological processes but also provides important clues for further understanding their functions in specific cellular environments.

Discussion

Divergence Patterns

The most common fate of duplicated genes is the deletion of some members, returning the gene to a single-copy state[8,40]. Besides, duplicated genes may undergo nonfunctionalization (to maintain an appropriate level of function), subfunctionalization (functional partitioning), or neofunctionalization (functional diversification)[8,41]. Since Ohno's proposition, some more detailed models have been accepted, such as duplication-degeneration-complementation (DDC)[3], escape from adaptive conflict (EAC)[42], and innovation-amplification-divergence (IAD)[43]. But overall, the fate of duplicated genes and their impact depend to some extent on the timing of the duplication event. Here, we will briefly compare the divergence of these four proteins following duplication, focusing on aspects such as their evolutionary history, structural differences, functional changes, and expression patterns, to determine which model of divergence they most closely align with.

In analyzing the differentiation types of these four agarase proteins, we observed that PL3506 and PL3512 have markedly changed activities compared to the more primitive version, PL3511. Product specificity suggests subfunctionalization, while enzyme mechanisms imply neofunctionalization. Therefore, we only confirm that functional differentiation has occurred without making overly detailed distinctions. Our analysis indicates PL3511 is closer to the ancestral enzyme, hinting at pre-existing multifunctionality. PL3506 and PL3512 respectively specialize into enzymes with single sub-functions, which can resolve conflicts between different activity transcriptional demands. Additionally, protein localization differentiation likely happened after enzymatic activity differentiation, raising questions about the cause of enzyme diversity. We propose the differentiation aligns more with the EAC model, suggesting adaptation and localization changes post-duplication. Enzyme evolution is complex and understanding it fully requires considering factors like metabolic network integration. Our current models (Fig. 4) are preliminary and will evolve with more data.

Enzyme Specialization Model

Bacterial communities are typically not composed of single cells but possess a certain structure, such as biofilms[44]. Against this backdrop of complexity, including constantly changing environments and heterogeneity among cells and/or species, bacteria have evolved complex enzyme systems for selective expression, thus addressing environmental challenges in an

economical and efficient manner. In the JK6 strain, in addition to this group of homologous agarases, we also identified some agarases with different localizations and hydrolytic activities through genome annotation and similarity comparison of known agarases based on AlphaFold-predicted structures. For example, PL3515[45] involved in the final step of agar metabolism with exo-acting α-1,3-agarase activity into AHG(3,6-anhydro-α-L-galactose) and GAL(D-β-galactose), and PL2878[46] predicted to be located on flagella.

In natural environments, bacteria coexist with other bacteria and organisms, engaging in a variety of interactions. Building on the results of previous studies[47,48], we speculate that the JK6 strain's metabolic pathway for agar (fig. S9). Initially, agar in the environment is hydrolyzed into longer polysaccharide units by exoenzymes and flagellar-located agarases. These two protein types may produce different induction signals for free and fixed agar in the environment, leading to distinct responses. Upon receiving short-chain sugar signals, the cell synthesizes and secretes PL3518 and PL3511 in large quantities, combined with cell motility, to further break down fixed agar into short-chain polysaccharides. Together with polysaccharide transport proteins, the shorter chain polysaccharides are transported to the periplasm, where they are hydrolyzed into NA4-sized units by enzymes located there, such as PL3512, and then transported into the cell via polysaccharide ABC transporters. Intracellular PL3506 hydrolyzes them into disaccharides, which are then hydrolyzed into monosaccharides by PL3515, entering subsequent metabolic pathways like AHG metabolism (PL0686, PL3528, and PL3525) and the Leloir pathway (PL1674, PL1675, and PL1676).

In conclusion, this study reveals the functional differentiation and environmental adaptability strategies of the agarase gene family in the JK6 strain. It provides a new perspective on understanding how bacteria compete for limited resources or form symbiotic relationships in complex microbial communities by specializing their enzyme systems. Evolutionarily, it demonstrates the broad evolutionary significance of gene duplication and functional innovation in the microbial world, especially how gene duplication and subsequent differentiation rapidly adapt to environmental changes, promoting biodiversity.

References and Notes

1 Ohno, S. Evolution by Gene Duplication. (1970).

2 Lan, X. & Pritchard, J. K. Coregulation of tandem duplicate genes slows evolution of subfunctionalization in mammals. Science 352, 1009-1013 (2016). https://doi.org:10.1126/science.aad8411

3 Hittinger, C. T. & Carroll, S. B. Gene duplication and the adaptive evolution of a classic genetic switch. Nature 449, 677-681 (2007). https://doi.org:10.1038/nature06151

4 Kafri, R., Springer, M. & Pilpel, Y. Genetic redundancy: new tricks for old genes. Cell 136, 389-392 (2009). https://doi.org:10.1016/j.cell.2009.01.027

5 Jensen, R. A. ENZYME RECRUITMENT IN EVOLUTION OF NEW FUNCTION. Annual Review of Microbiology 30, 409-425 (1976). https://doi.org:10.1146/annurev.mi.30.100176.002205

6 Innan, H. & Kondrashov, F. The evolution of gene duplications: classifying and distinguishing between models. Nat Rev Genet 11, 97-108 (2010). https://doi.org:10.1038/nrg2689

7 Copley, S. D. Evolution of new enzymes by gene duplication and divergence. FEBS J 287, 1262-1283 (2020). https://doi.org:10.1111/febs.15299

8 Birchler, J. A. & Yang, H. The multiple fates of gene duplications: Deletion, hypofunctionalization, subfunctionalization, neofunctionalization, dosage balance constraints, and neutral variation. The Plant Cell 34, 2466-2474 (2022). https://doi.org:10.1093/plcell/koac076

9 Conant, G. C. & Wolfe, K. H. Turning a hobby into a job: how duplicated genes find new functions. Nat Rev Genet 9, 938-950 (2008). https://doi.org:10.1038/nrg2482

10 Noda-Garcia, L., Liebermeister, W. & Tawfik, D. S. Metabolite-Enzyme Coevolution: From Single Enzymes to Metabolic Pathways and Networks. Annu Rev Biochem 87, 187-216 (2018). https://doi.org:10.1146/annurev-biochem-062917-012023

11 Azam, F. & Malfatti, F. Microbial structuring of marine ecosystems. Nat Rev Microbiol 5, 782-791 (2007). https://doi.org:10.1038/nrmicro1747

12 Despotovic, D. et al. Utilization of diverse organophosphorus pollutants by marine bacteria. Proc Natl Acad Sci U S A 119, e2203604119 (2022). https://doi.org:10.1073/pnas.2203604119

13 Andersson, D. I. & Hughes, D. Gene amplification and adaptive evolution in bacteria. Annu Rev Genet 43, 167-195 (2009). https://doi.org:10.1146/annurev-genet-102108-134805

14 Cantarel, B. L. et al. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. Nucleic Acids Res 37, D233-238 (2009). https://doi.org:10.1093/nar/gkn663

15 Araki, C. STRUCTURE OF THE AGAROSE CONSTITUENT OF AGAR-AGAR. BULLETIN OF THE CHEMICAL SOCIETY OF JAPAN 29, 543-544 (1956). https://doi.org:10.1246/bcsj.29.543

16 Jiang, C., Liu, Z., Cheng, D. & Mao, X. Agarose degradation for utilization: Enzymes, pathways, metabolic engineering methods and products. Biotechnol Adv 45, 107641 (2020). https://doi.org:10.1016/j.biotechadv.2020.107641

17 Borin, J. M. et al. Rapid bacteria-phage coevolution drives the emergence of multiscale networks. Science 382, 674-678 (2023). https://doi.org:10.1126/science.adi5536

18 Campanaro, S., Treu, L. & Valle, G. Protein evolution in deep sea bacteria: an analysis of amino acids substitution rates. BMC Evol Biol 8, 313 (2008). https://doi.org:10.1186/1471-2148-8-313

19 Treangen, T. J. & Rocha, E. P. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. PLoS Genet 7, e1001284 (2011). https://doi.org:10.1371/journal.pgen.1001284

20 Reams, A. B., Kofoid, E., Savageau, M. & Roth, J. R. Duplication frequency in a population of Salmonella enterica rapidly approaches steady state with or without recombination. Genetics 184, 1077-1094 (2010). https://doi.org:10.1534/genetics.109.111963

21 Pettersson, M. E., Sun, S., Andersson, D. I. & Berg, O. G. Evolution of new gene functions: simulation and analysis of the amplification model. Genetica 135, 309-324 (2009). https://doi.org:10.1007/s10709-008-9289-z

22 Blum, M. et al. The InterPro protein families and domains database: 20 years on. Nucleic Acids Research 49, D344-D354 (2021). https://doi.org:10.1093/nar/gkaa977

23 Pluvinage, B., Hehemann, J. H. & Boraston, A. B. Substrate recognition and hydrolysis by a family 50 exo-beta-agarase, Aga50D, from the marine bacterium Saccharophagus degradans. J Biol Chem 288, 28078-28088 (2013). https://doi.org:10.1074/jbc.M113.491068

24 Vajravijayan, S. et al. Structural insights on starch hydrolysis by plant beta-amylase and its evolutionary relationship with bacterial enzymes. Int J Biol Macromol 113, 329-337 (2018). https://doi.org:10.1016/j.ijbiomac.2018.02.138

25 Sadowski, M. I. & Jones, D. T. The sequence-structure relationship and protein function prediction. Curr Opin Struct Biol 19, 357-362 (2009). https://doi.org:10.1016/j.sbi.2009.03.008

26 Lesk, A. M. & Chothia, C. How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. J Mol Biol 136, 225-270 (1980). https://doi.org:10.1016/0022-2836(80)90373-3

27 Grishin, N. V. Fold change in evolution of protein structures. J Struct Biol 134, 167-185 (2001). https://doi.org:10.1006/jsbi.2001.4335

28 Liu, Y. & Bahar, I. Sequence evolution correlates with structural dynamics. Mol Biol Evol 29, 2253-2263 (2012). https://doi.org:10.1093/molbev/mss097

29 Addou, S., Rentzsch, R., Lee, D. & Orengo, C. A. Domain-based and family-specific sequence identity thresholds increase the levels of reliable protein function transfer. J Mol Biol 387, 416-430 (2009). https://doi.org:10.1016/j.jmb.2008.12.045

30 Salemme, F. R., Miller, M. D. & Jordan, S. R. Structural convergence during protein evolution. Proc Natl Acad Sci U S A 74, 2820-2824 (1977). https://doi.org:10.1073/pnas.74.7.2820

31 Parisi, G. & Echave, J. Structural constraints and emergence of sequence patterns in protein evolution. Mol Biol Evol 18, 750-756 (2001). https://doi.org:10.1093/oxfordjournals.molbev.a003857

32 Liu, Y. et al. Genome-Wide Identification and Functional Characterization of β-Agarases in Vibrio astriarenae Strain HN897. Frontiers in Microbiology 11 (2020). https://doi.org:10.3389/fmicb.2020.01404

33 Marques, A. C., Vinckenbosch, N., Brawand, D. & Kaessmann, H. Functional diversification of duplicate genes through subcellular adaptation of encoded proteins. Genome Biol 9, R54 (2008). https://doi.org:10.1186/gb-2008-9-3-r54

34 Voordeckers, K. et al. Reconstruction of ancestral metabolic enzymes reveals molecular mechanisms underlying evolutionary innovation through gene duplication. PLoS Biol 10, e1001446 (2012). https://doi.org:10.1371/journal.pbio.1001446

35 Carroll, S. M., Ortlund, E. A. & Thornton, J. W. Mechanisms for the evolution of a derived function in the ancestral glucocorticoid receptor. PLoS Genet 7, e1002117 (2011). https://doi.org:10.1371/journal.pgen.1002117

36 Huynen, M. A. & Bork, P. Measuring genome evolution. Proc Natl Acad Sci U S A 95, 5849-5856 (1998). https://doi.org:10.1073/pnas.95.11.5849

37 Clark, N. L., Alani, E. & Aquadro, C. F. Evolutionary rate covariation reveals shared functionality and coexpression of genes. Genome Res 22, 714-720 (2012). https://doi.org:10.1101/gr.132647.111

38 Telford, M. J. et al. Phylogenomic analysis of echinoderm class relationships supports Asterozoa. Proceedings of the Royal Society B: Biological Sciences 281 (2014). https://doi.org:10.1098/rspb.2014.0479

39 Chen, M.-Y., Liang, D. & Zhang, P. Phylogenomic Resolution of the Phylogeny of Laurasiatherian Mammals: Exploring Phylogenetic Signals within Coding and Noncoding Sequences. Genome Biology and Evolution 9, 1998-2012 (2017). https://doi.org:10.1093/gbe/evx147

40 Dean, E. J., Davis, J. C., Davis, R. W. & Petrov, D. A. Pervasive and persistent redundancy among duplicated genes in yeast. PLoS Genet 4, e1000113 (2008). https://doi.org:10.1371/journal.pgen.1000113

41 Hahn, M. W. Distinguishing among evolutionary models for the maintenance of gene duplicates. J Hered 100, 605-617 (2009). https://doi.org:10.1093/jhered/esp047

42 Des Marais, D. L. & Rausher, M. D. Escape from adaptive conflict after duplication in an anthocyanin pathway gene. Nature 454, 762-765 (2008). https://doi.org:10.1038/nature07092

43 Näsvall, J., Sun, L., Roth, J. R. & Andersson, D. I. Real-Time Evolution of New Genes by Innovation, Amplification, and Divergence. Science 338, 384-387 (2012). https://doi.org:10.1126/science.1226521

44 Noda-Garcia, L. et al. Chance and pleiotropy dominate genetic diversity in complex bacterial environments. Nat Microbiol 4, 1221-1230 (2019). https://doi.org:10.1038/s41564-019-0412-y

45 Ha, S. C. et al. Crystal structure of a key enzyme in the agarolytic pathway, alpha-neoagarobiose hydrolase from Saccharophagus degradans 2-40. Biochem Biophys Res Commun 412, 238-244 (2011). https://doi.org:10.1016/j.bbrc.2011.07.073

46 Lyu, Q. et al. Structural insights into a novel Ca(2+)-independent PL-6 alginate lyase from Vibrio OU02 identify the possible subsites responsible for product distribution. Biochim Biophys Acta Gen Subj 1863, 1167-1176 (2019). https://doi.org:10.1016/j.bbagen.2019.04.013

47 Arnosti, C. et al. in Annual Review of Marine Science, Vol 13, 2021 Vol. 13 Annual Review of Marine Science (eds C. A. Carlson & S. J. Giovannoni) 81-108 (Annual Reviews, 2021).

48 McKee, L. S. et al. Polysaccharide degradation by the Bacteroidetes: mechanisms and nomenclature. Environ. Microbiol. Rep., 23 (2021). https://doi.org:10.1111/1758-2229.12980

49 Frank, J. A. et al. Critical Evaluation of Two Primers Commonly Used for Amplification of Bacterial 16S rRNA Genes. Applied and Environmental Microbiology 74, 2461-2470 (2008). https://doi.org:10.1128/aem.02272-07

50 Chun, J. et al. Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes. Int J Syst Evol Microbiol 68, 461-466 (2018). https://doi.org:10.1099/ijsem.0.002516

51 Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34, i884-i890 (2018). https://doi.org:10.1093/bioinformatics/bty560

52 Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A. & Korobeynikov, A. Using SPAdes De Novo Assembler. Current Protocols in Bioinformatics 70 (2020). https://doi.org:10.1002/cpbi.102

53 Ruan, J. & Li, H. Fast and accurate long-read assembly with wtdbg2. Nature Methods 17, 155-158 (2019). https://doi.org:10.1038/s41592-019-0669-3

54 Li, H. & Birol, I. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34, 3094-3100 (2018). https://doi.org:10.1093/bioinformatics/bty191

55 Danecek, P. et al. Twelve years of SAMtools and BCFtools. GigaScience 10 (2021). https://doi.org:10.1093/gigascience/giab008

56 Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol 13, e1005595 (2017). https://doi.org:10.1371/journal.pcbi.1005595

57 Kurtz, S. et al. Versatile and open software for comparing large genomes. GENOME BIOLOGY 5 (2004). https://doi.org:10.1186/gb-2004-5-2-r12

58 Wang, J. et al. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. PLoS ONE 9 (2014). https://doi.org:10.1371/journal.pone.0112963

59 Besemer, J., Lomsadze, A. & Borodovsky, M. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. NUCLEIC ACIDS RESEARCH 29, 2607-2618 (2001). https://doi.org:10.1093/nar/29.12.2607

60 Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11, 119 (2010). https://doi.org:10.1186/1471-2105-11-119

61 Seemann, T. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30, 2068-2069 (2014). https://doi.org:10.1093/bioinformatics/btu153

62 Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nature Methods 18, 366-368 (2021). https://doi.org:10.1038/s41592-021-01101-x

63 Cheng, Y. I. et al. Comparative Genomic Analysis of a Novel Strain of Taiwan Hot-Spring Cyanobacterium Thermosynechococcus sp. CL-1. Front Microbiol 11, 82 (2020). https://doi.org:10.3389/fmicb.2020.00082

64 Sahu, N. et al. Vertical and horizontal gene transfer shaped plant colonization and biomass degradation in the fungal genus Armillaria. Nature Microbiology 8, 1668-1681 (2023). https://doi.org:10.1038/s41564-023-01448-1

65 Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol 20, 238 (2019). https://doi.org:10.1186/s13059-019-1832-y

66 Contreras-Moreira, B. & Vinuesa, P. GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. Appl Environ Microbiol 79, 7696-7701 (2013). https://doi.org:10.1128/AEM.02411-13

67 Zhou, Z. et al. METABOLIC: high-throughput profiling of microbial genomes for functional traits, metabolism, biogeochemistry, and community-scale functional networks. Microbiome 10, 33 (2022). https://doi.org:10.1186/s40168-021-01213-8

68 Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24, 1586-1591 (2007). https://doi.org:10.1093/molbev/msm088

69 Xu, B. & Yang, Z. PAMLX: A Graphical User Interface for PAML. Molecular Biology and Evolution 30, 2723-2724 (2013). https://doi.org:10.1093/molbev/mst179

70 Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583-+ (2021). https://doi.org:10.1038/s41586-021-03819-2
71 Wayment-Steele, H. K. et al. Predicting multiple conformations via sequence clustering and AlphaFold2. Nature 625, 832-839 (2023). https://doi.org:10.1038/s41586-023-06832-9
72 Baek, M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science 373, 871-876 (2021). https://doi.org:10.1126/science.abj8754
73 Lu, W. et al. TANKBind: Trigonometry-Aware Neural NetworKs for Drug-Protein Binding Structure Prediction. (2022). https://doi.org:10.1101/2022.06.06.495043
74 Eberhardt, J., Santos-Martins, D., Tillack, A. F. & Forli, S. AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. J Chem Inf Model 61, 3891-3898 (2021). https://doi.org:10.1021/acs.jcim.1c00203
75 Forli, S. et al. Computational protein-ligand docking and virtual drug screening with the AutoDock suite. Nat Protoc 11, 905-919 (2016). https://doi.org:10.1038/nprot.2016.051
76 Zhang, C., Shine, M., Pyle, A. M. & Zhang, Y. US-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes. Nat Methods 19, 1109-1115 (2022). https://doi.org:10.1038/s41592-022-01585-1
77 Zhang, Z. KaKs_Calculator 3.0: Calculating Selective Pressure on Coding and Non-coding Sequences. Genomics Proteomics Bioinformatics 20, 536-540 (2022). https://doi.org:10.1016/j.gpb.2021.12.002
78 Zhang, Z. et al. ParaAT: A parallel tool for constructing multiple protein-coding DNA alignments. Biochemical and Biophysical Research Communications 419, 779-781 (2012). https://doi.org:10.1016/j.bbrc.2012.02.101
79 Yu, N. Y. et al. PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. Bioinformatics 26, 1608-1615 (2010). https://doi.org:10.1093/bioinformatics/btq249
80 Chou, K.-C. & Shen, H.-B. Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms. Nature Protocols 3, 153-162 (2008). https://doi.org:10.1038/nprot.2007.494
81 Yu, C. S., Lin, C. J. & Hwang, J. K. Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions. Protein Science 13, 1402-1406 (2004). https://doi.org:10.1110/ps.03479604
82 Goldberg, T. et al. LocTree3 prediction of localization. Nucleic Acids Research 42, W350-W355 (2014). https://doi.org:10.1093/nar/gku396
83 Teufel, F. et al. SignalP 6.0 predicts all five types of signal peptides using protein language models. Nat Biotechnol 40, 1023-1025 (2022). https://doi.org:10.1038/s41587-021-01156-3
84 Rozewicki, J., Li, S., Amada, K. M., Standley, D. M. & Katoh, K. MAFFT-DASH: integrated protein sequence and structural alignment. Nucleic Acids Res 47, W5-W10 (2019). https://doi.org:10.1093/nar/gkz342
85 Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Molecular Biology and Evolution 32, 268-274 (2015). https://doi.org:10.1093/molbev/msu300
86 Steenwyk, J. L. et al. PhyKIT: a broadly applicable UNIX shell toolkit for processing and analyzing phylogenomic data. Bioinformatics 37, 2325-2331 (2021). https://doi.org:10.1093/bioinformatics/btab096
87 Hafizah, N. F., Teh, A. H. & Furusawa, G. Biochemical Characterization of Thermostable and Detergent-Tolerant beta-Agarase, PdAgaC, from Persicobacter sp. CCB-QB2. Appl Biochem Biotechnol 187, 770-781 (2019). https://doi.org:10.1007/s12010-018-2849-5
88 Jin, Y., Yu, S. R., Kim, D. H., Yun, E. J. & Kim, K. H. Characterization of Neoagarooligosaccharide Hydrolase BpGH117 from a Human Gut Bacterium Bacteroides plebeius. Mar. Drugs 19, 13 (2021). https://doi.org:10.3390/md19050271

89 Yun, E. J. et al. Production of 3,6-anhydro-l-galactose from agarose by agarolytic enzymes of Saccharophagus degradans 2-40. Process Biochemistry 46, 88-93 (2011). https://doi.org:10.1016/j.procbio.2010.07.019
90 Zhu, Y. et al. Characterization of an alkaline beta-agarase from Stenotrophomonas sp. NTa and the enzymatic hydrolysates. Int J Biol Macromol 86, 525-534 (2016). https://doi.org:10.1016/j.ijbiomac.2016.01.106
91 Yu, S., Yun, E. J., Kim, D. H., Park, S. Y. & Kim, K. H. Dual Agarolytic Pathways in a Marine Bacterium, Vibrio sp. Strain EJY3: Molecular and Enzymatic Verification. Appl Environ Microbiol 86, 15 (2020). https://doi.org:10.1128/AEM.02724-19

Materials and Methods

Abbreviations:
D-Gal, D-β-galactose;
L-AHG, 3,6-anhydro-α-L-galactose;
NAOS, neoagarooligosaccharide;
DP, degree of polymerization;
NABH, α-neoagarobiose hydrolase;
GH, Glycoside hydrolase;
NA2, neoagarobiose;
NA4, neoagarotetraose;
NA6, neoagarohexaose;
NA8, neoagarooctaose;
A3, agarotriose.

All reagents and chemicals were purchased from Sigma-Aldrich (St. Louis, MO) unless otherwise specified.

Field Sample Collection, Laboratory Isolation and Identification, and Microscopic Observation

During the 2020 joint cruise to Mariana Trench of Westlake University and Shanghai Ocean University, water samples were collected at depths of 0-4km at various stops along the route, and sediment samples were collected at a depth of 6km at coordinates 18.5047 132.0094. Microalgae from deep sea were enriched and isolated by spreading 100uL of 1-6km samples (water samples were centrifuged for 10 minutes at 5000g, the supernatant was removed, and 100uL of the bottom liquid was retained; sediment samples were precipitated with sterilized L1 culture medium and the supernatant was collected) onto L1 culture plates. Unexpectedly, colonies that produced agar-hydrolyzing pits were observed and purified by streaking on fresh plates multiple times. Ultimately, strain JK6, which exhibited the largest hydrolysis zone in a unit of time, was selected for subsequent experiments. The strain was preserved long-term at -80°C in 5% glycerol and cultured in L1 medium (see L1 Medium from NCMA) supplemented with 0.2% agar (L1A) or 1% LB Broth (L1B) for liquid culture and 1.5% agar for solid culture.

A near full-length 16S rRNA gene sequence was PCR amplified using universal primers 27F and 1492R to identify the strain[49,50]. The obtained 16S rRNA gene amplicon sequence was deposited in the National Center for Biotechnology Information (NCBI) database and annotated using NCBI's online Basic Local Alignment Search Tool (BLAST) through Sanger sequencing.

After Gram staining, Brightfield images were captured using a Novel DN-107T microscope equipped with an Optronics digital camera. The sample preparation process for scanning electron microscope(SEM) involved collecting the JK6 bacterial aggregates formed in liquid culture, fixing them in a fixation solution (pH=7.2) containing a mixture of 2% paraformaldehyde (PFA) and 2.5% glutaraldehyde(GA) for 1 hour at room temperature, rinsing the samples three times with 0.1 M PB (pH 7.2-7.4) buffer for 10 minutes each time at 4°C, fixing them with 1% osmium tetroxide (in 0.1 M PB) on ice for 1 hour, rinsing the samples three times with ddH2O for 15 minutes each time at room temperature, dehydrating them with 30%, 50%, 70%, 95%, and 100%×3 ethanol for 10 minutes at room temperature, collecting the samples in a paper bag, drying them using a critical-point dryer (Leica EM CPD300), and finally sputter-coating them 5nm thick with Leica EM ACE600. Imaging was conducted using a scanning electron microscope (Zeiss Gemini SEM Crossbeam 550).

Genome Sequencing Assembly and Preliminary Annotation

The genomic DNA of strain JK6 was extracted using the TIANamp Bacteria DNA Kit, and then split into two batches for sequencing using both Illumina next-generation sequencing (PE150, Beijing Novogene Bioinformatics Technology Co., Ltd.) and Oxford Nanopore (PromethION, by Wuhan Benagen Technology Co., Ltd.). All experimental procedures, including sample quality testing, library construction, library quality testing, and library sequencing, followed the standard protocol provided by the sequencer manufacturers.

The genome assembly process consisted of three steps: raw reads filtering, co-assembly of second and third generation sequencing data, and polishing of assembly results. Raw short reads from Illumina sequencing were quality controlled using Fastp (-q 30 -u 10 -n 5 -l 130)[51], and raw long reads from Nanopore sequencing were filtered using the clean reads provided by the company (removing sequences with an average quality value of less than or equal to 7). The clean short reads and long reads were then co-assembled using SPAdes-3.15.2-Linux[52], wtdbg2[53], minimap2[54], SAMtools[55] and Unicycler[56] to generate complete genome sequences. The assembly results of each version were evaluated using Quast.py and MUMmer[57] to examine assembly statistics and collinearity between different versions. The best version was polished twice using pilon-1.24.jar[58]. Gene annotation software was used to compare the annotation results before and after polishing. As there was almost no difference between the two results, the polished genome version was selected as the complete genome sequence. To annotate the selected reference genome, annotation results from different software, including GeneMarkS[59], prodigal[60] and prokka[61], were compared based on the number of annotated genes and consistency of the results obtained by different methods. Prokka was chosen as the final gene annotation file.

We collected datasets such as "uniprot-reviewed_yes" prior to February 25, 2022, and constructed indexes for them. The protein sequence of JK6 was aligned against each dataset using the diamond[62] blastp command, with parameter settings of --evalue 1e-5, --id 30, and --max-target-seqs 1, to annotate homologous proteins for each protein sequence.

Horizontal gene transfer analysis

We used blastp searches against the NCBI non-redundant protein database to identify genes that may have originated from horizontal gene transfer (HGT)[63,64]. To ensure that the hits

represent likely homologs, we only retained hits with high-scoring pairs accounting for at least 90% of the query length and overall amino acid sequence similarity of at least 40%. Genes with the best hit that lacked a genus-level taxonomic assignment or were derived from metagenomics surveys were manually examined for more reliable inference. A gene was considered putatively acquired if more than half of the top-five hits were from other genera.

Comparative Genomics Analysis

Comparative genomics analysis of *Agarivorans* species Using Orthofinder[65] and Get_homologues[66]. In this study, we downloaded 629 genomes (Data S2) containing agarase protein sequences from NCBI using the datasets tool. The genomes were annotated using prokka, and orthofinder.py and get_homologues.pl were used to compute the faa and gbf files for all species. Based on the results, closely related *Agarivorans* species were selected for pan-genomic analysis following the software instructions (partial strain information provided in the attached Data S8). METABOLIC-C.pl[67] was used to annotate the *Agarivorans* genomes, and METABOLIC-G.pl[67] was used to annotate the protein sequences, followed by comparison of the similarities and differences among different genomes.

Positive Selection Analysis and Ancestral Sequence Reconstruction by PAML

We selected the genomes of *Agarivorans* strains collected from shallow sea as background controls (Data S8), and used OrthoFinder to identify the orthologous genes in this group of strains compared to JK6. Only single-copy orthologs with full-length protein sequences longer than 0.7 times the longest sequence in all strains were retained. The corresponding DNA sequence files and tree files were filtered accordingly and they were used for the following selective pressure and convergence analyses.

The influence of natural selection acting on protein-coding genes was inferred by estimating the ratio of the nonsynonymous and synonymous substitutions (ω). In brief, $\omega > 1$, $= 1$ and < 1 indicate positive selection, neutral evolution and negative selection, respectively. We used program codeml for selective pressure analyses in Phylogenetic Analysis by Maximum Likelihood (PAML)[68]. The p-values were estimated assuming a null-distribution that is a 50:50 mixture of a $\chi 2$ distribution and a point mass at zero. The p-values were corrected for multiple testing, and a false discovery rate (FDR) cutoff of 0.05 was used.

Ancestral sequence reconstruction was performed using PAMLX[69], with the following parameters specified: noisy = 4, verbose = 1, seqtype = 2, aaRatefile = wag.dat, model = 3, fix_alpha = 1, RateAncestor = 1, Small_Diff = 5e-6, fix_blength = 2, method = 0.

AlphaFold Structure Prediction and Docking Comparison

The genome proteins of JK6 were compared to the NCBI database, and those with 100% sequence similarity were downloaded from the AlphaFold Protein Structure Database (https://alphafold.ebi.ac.uk/). For the remaining 1000+ proteins, monomer structures were predicted using locally deployed AlphaFold[70,71], and the agarases were additionally predicted using RoseTTAFold[72] at https://robetta.bakerlab.org/. Polysaccharide ligands were mainly downloaded from PDBe (Data S9), and the full-genome protein docking was performed using TANKBind[73]. The docking of agarase proteins and candidate agarase proteins was carried out using Vina[74,75], with the docking box confirmed by AutoDockTools-1.5.7. Using the structural

domain ranges defined by Gene3D, the glycoside hydrolase domains were extracted and subjected to AutoDock Vina docking experiments with polysaccharide molecules from NA2 to NA8 as substrates. After adjusting parameters and repeating experiments multiple times, amino acids within 5Å of the sugar molecules were counted, with those exceeding a threshold considered as part of the active center region. The docking of glycoside hydrolase domains and related proteins was accomplished in bulk using Python scripts （https://github.com/jkgkn/JK6）.

Structural similarity comparisons between the whole-genome proteins and the obtained homologous glycoside hydrolase domains were conducted using US-align[76]. The magnitude of the spatial vector composed of TM1, TM2, and RMSD was utilized as a measure of distance to compare structural similarity between two proteins. The specific calculation formula is：$\sqrt{(1-Tm_1)^2+(1-Tm_2)^2+(RmSD\ /\ 10)^2}$. USalign was used to pairwise compare the extracted glycoside hydrolase domains of the proteins, with the structural similarity score Vd represented by the spatial vector magnitude of the TM scores and RMSD score between two proteins, and a clustering tree was constructed based on this matrix. The results were generally consistent with the branching relationships of the gene phylogenetic trees, providing valuable insights. It should be noted that due to the potential for large alignment discrepancies caused by the loose structure of multi-domain proteins, this study focused solely on structural similarity analysis of core domains. According to structural similarity clustering, the four agarases were clearly divided into four groups (Fig. S6).

Calculation of Ka/Ks

We utilized KaKs_Calculator3.0[77] to calculate selective pressure. Specifically, we prepared pairwise comparison homolog group files (PLs.homologs), a file containing multiple amino acid sequences (prp.fasta), and a file containing multiple nucleotide sequences (cds.fasta). Then, we executed the ParaAT.pl script[78] with the parameters -h PLs.homologs -n cds.fasta -a prp.fasta to generate the AXT sequence files required for calculating Ka (nonsynonymous substitution rate) and Ks (synonymous substitution rate). For the genetic code, we selected the codes for bacteria, archaea, and plant plastids. Regarding the calculation method, we chose to use all substitution models (ALL) for the computation. Although homologous proteins exhibited overall structural similarity, their significant functional differences indicated the presence of variations. Considering that ka/ks values globally reflect the selection pressure on proteins, there are extensive protein regions under purifying selection to maintain the overall structure and stability of the enzyme, but this overlooks the fact that only a few sites are under positive selection. Hence, we examined the interval changes in ka/ks values of proteins using a sliding window of 120 codon sequences, finding that enzyme active center regions indeed fall within sequences under positive selection.

Considering the multiple influences of genetic code structure and amino acid physicochemical properties on protein evolution, we calculated the ratio of amino acid sites differing by more than 1Å to those below this threshold (named the structural conservation index, SCI) to characterize protein structural conservatism, specifically, align pairwise protein structures using US-align with a threshold set to -d 1Å. Extract the number of amino acids

marked by ":" (denote residue pairs with d < 1 Angstrom), "." (denote other aligned residues), and gaps, labeled as NPR, OAR, NMR, respectively. Calculate SCI as NPR/AR - OAR/AR - 2*NMR/AR, where AR represents the total number of sites. The range of SCI values is from -2 to 1, and the closer the value is to 1, the more similar the two protein structures are.

Prediction of Protein Subcellular Localization

Cellular localization of proteins from the JK6 strain and those reconstructed from ancestors was predicted using a combination of protein localization prediction software, including PSORTb v3.0[79] , Cell-PLoc 2.0[80], CELLO v.2.5[81], and LocTree3[82], in conjunction with SignalP 6.0[83]. Additionally, the distinction was aided by comparing the ratio of supernatant to total protein content in the cell culture fluid.

Evaluate gene-gene covariation, pairwise identity and evolutionary rate

We extracted species from the collected genomic dataset that have homologs of all four target proteins. Then, we constructed a species tree by merging the common genes of all these species aligned using MAFFT[84]. Subsequently, constrained gene trees were built using the species tree as a reference constraint (with iqtree2[85], specifying the species tree using the -keep-ident -te parameter). To calculate gene-gene covariation, we used phykit[86] with the command “cover -r” referencing the species tree. The evolutionary rate was determined using “evolutionary_rate”, and pairwise identity was calculated using “pairwise_identity”.

The expression and purification of recombinant proteins

Protein Expression and Purification. The genes that correspond for agarase were amplified by PCR from genome DNA of JK6 strain. The primer sequences used for cloning were listed in table S1. Each amplified DNA fragment was inserted into the EcoRI-XhoI site of pPEI-His-Sumo. Escherichia coli strain BL21(DE3) cells (Vazyme) carrying each expression plasmid were cultured in 2 L LB medium with 50 μg/mL carbenicillin at 37°C and induced by 1 mM isopropyl β-D-1-thiogalactopyranoside (A100487, Sangon Biotech) at 16°C for 16 h when the OD600 reached 0.8. cells were harvested by centrifugation at 4000 g for 10 min at 4°C. The cell pellet was re-suspended in 80 mL lysis buffer (20 mM Tris-HCl pH 8.0, 500 mM NaCl, 5 mM imidazole) supplemented with 1 mM phenylmethanesulfonyl fluoride (PMSF, A100754, Sangon Biotech). Then, the cells were disturbed by High pressure homogenizer (ATS Engineering Limited, AH1500) for 3 min at 800 bar, and extracted proteins were separated from cell debris by centrifugation at 12000 g for 1 h at 4°C. His-tagged proteins were purified using HisTrap HP Column (5 mL, Cytiva) with Fast Protein Purification System (AKTA Pure 25), and eluted with elution buffer (20 mM HEPES pH 7.5, 300 mM imidazole, 100 mM KCl). The proteins were further enriched using HiTrap Q HP Column (5 mL, Cytiva) and separated by gradient elution with buffer A (20 mM HEPES pH 7.5, 50 mM KCl, 1 mM DTT) and buffer B (20 mM HEPES pH 7.5, 1 M KCl, 1 mM DTT). The peak fractions containing the agarase protein were collected to about 10 mg/mL for further experiments.

The gene sequences of the *Vibrio astriarenae* strain (HN897) strain were synthesized by Genscript, named *Vsp6*, *Vsp11*, *Vsp12* and *Vsp18* respectively, based on the homology with the four agarase proteins in JK6. Following construction through standard methods for sequence replacement and point mutation in plasmids, the procedures were executed consistently with the

previously described conventional protein expression process. The fragment "FADKDKRSGKEVASEV" in PL3512 was replaced by "YDHSKVAARSDDDVTPEDSKEKLEISQEAFDSAYVANQT", which derived from PL3506 fragment. The fragment "STSGADRLPVKQIGSG" in PL3511 was replaced by "IFGDHARISTGND", which derived from PL3512. The following active part "HSKVAARSDDDVTPEDSKEKLEISQEAFDSAY" in PL3506, the "LSEDLSLRPSQISQDSYVDNQ" in PL3518 were deleted in their sequence. The modified proteins were marked as PL3506-M, PL3511-M, PL3512-M and PL3518-M. The point mutant of PL3506 is named 6M, the glycoside hydrolase domain of PL3506 is named 6g, and the point mutation in the glycoside hydrolase domain of PL3506 is named 6gM.The amino acid sequences of each protein and the modified version were shown below.

Assay of agarase activity

The activity of agarases was measured using the 3,5-dinitrosalicylic acid (DNS) method[87]. Briefly, 10 μL of each crude enzyme sample was added to 490 μL of PBS buffer (pH 7.4) containing 0.1% agarose powder, and incubated at 37°C for 1 hour. After centrifugation at 12,000 rpm, 50 μL of the supernatant was added to a 96-well PCR plate containing 100 μL of DNS reagent. As controls, equal volumes of D-galactose solutions with gradient concentrations were added to other wells. The PCR plate was sealed with adhesive film and centrifuged to the bottom. Then, the plate was heated at 99°C for 10 minutes in a PCR machine, and immediately cooled on ice to stop the reaction. After cooling, 100 μL of each well solution was transferred to an enzyme-linked immunosorbent assay (ELISA) plate. The absorbance was measured at 540 nm using an ELISA reader (Thermo Varioskan LUX). The concentration of reducing sugar produced by each enzyme was calculated based on a standard curve.

The enzymatic hydrolysis products analysis using thin-layer chromatography (TLC)

The hydrolysis products of agarose by various agarases were separated using thin-layer chromatography (TLC)[87-89]. Crude enzyme samples, post-expression and purification, amounting to 20 μL, were digested overnight at room temperature in 500 μL of PBS buffer (pH 7.4) containing an excess of agar powder. Following centrifugation at 12,000 rpm, the supernatant was collected and aliquots of 5 μL were spotted onto silica gel 60 plates (Merck Millipore) and dried prior to chromatography. Two different mobile phases and staining solutions were employed for chromatographic separation and visualization, respectively. The first mobile phase consisted of 1-butanol/acetic acid/distilled water (2:1:1, v/v/v)[90], and the gels were visualized with a 3 mL/40 mL sulfuric acid/ethanol solution, followed by heating in a microwave oven for 30 seconds. The second mobile phase was composed of 1-butanol/acetic acid/distilled water (6:3:2, v/v/v), with visualization achieved using a 3 mL/40 mL sulfuric acid/ethanol solution containing 0.2% 1,3-dihydroxynaphthalene[91], followed by a similar heating step in a microwave oven for 30 seconds. As standards, NA2, A3, NA4, agarose oligosaccharides were purchased from Qingdao Bozhi Huli Biotechnology Co., Ltd., Gal was purchased from Sangon Biotech Co., Ltd.

Mass Spectrometry Techniques for Polysaccharide Analysis

To identify the molecular weights of sugars from different bands obtained via Thin Layer Chromatography (TLC), the major bands were scraped, combined, and dissolved in deionized water (ddH2O) before being centrifuged to collect the supernatant. The hydrolysis products from different proteins were filtered directly through a 5kD ultrafiltration tube to measure the filtrate. All samples were sent to The Mass Spectrometry & Metabolomics Core Facility at Westlake University for analysis using Quadrupole Time-of-Flight Mass Spectrometry (Q/TOF-MS). The specific process involves: The Mass Spectrometry & Metabolomics Core Facility at Westlake University was provided with the combined major bands obtained from TLC, dissolved in ddH2O and then centrifuged to obtain the supernatant. The sample was subjected to Q/TOF-MS mass spectrometry detection. The specific process was as follows: Chromatographic elution was performed on an Agilent 1290 infinity system (Agilent technologies, USA) equipped with a sample manager coupled to a mass spectrometry with an electrospray ion source in both positive and negative ion modes. The separation was carried out on the BEH amide column (100 mm × 2.1 mm, 1.7 μm) at 40°C. The mobile phase A was Ultra-pure water containing 0.3% ammonia and 15 mM ammonium acetate, and the mobile phase B was acetonitrile/water (9:1) containing 0.3% ammonia and 15 mM ammonium acetate. The flow rate was 0.3mL/min, and the gradient of mobile phase A was held at 5% for 1 min, 5% to 50% in 8 min, held at 50% for 3 min, then 50% to 5% in 0.5 min, and finally held at 5% for 6.5 min. The sample volume injected was 5 μL.

An Agilent 6545 Q/TOF-MS system both in the ESI+ and ESI- modes was used for the detection. The parameters used were as follows: capillary voltage, 3500 V; Nozzle voltage, 120 V; nebulizer gas, 35 psi; drying gas flow rate, 8 L/min; and gas temperature, 350°C. A full scan was run with a mass range from m/z 50 to 1200 using the high-resolution mode (Extended Dynamic Range 2GHz). Also, to produce accurate mass correction, an automated calibration delivery system was conducted to calibrate the MS and MS/MS results automatically by prepared calibration solutions before and during the analysis. Data processing was using Qualitative Workflow.

Proteomic analysis of the JK6 strain.

In this study, protein samples were prepared for mass spectrometry analysis in two groups. For the first group, bacterial cells were cultured in L1A medium for 24 hours, collected by centrifugation, and washed twice before being inoculated into equal volumes of L1A, L1B, and L1 media, respectively. After another 24 hours of cultivation, the cells were lysed under high pressure, and the supernatant containing proteins was collected for analysis. The second group followed a similar procedure, starting with cultivation in L1B medium. However, after the second 24-hour cultivation in L1A, L1B, and L1 media, both the cells and the culture supernatant were collected. The culture supernatant was concentrated using a 10kD concentration tube, and the cells were lysed under high pressure to collect and concentrate the proteins in the supernatant. This approach enabled the comparison of protein expression profiles under different culture conditions, providing valuable insights into the impact of these conditions on cellular protein production and secretion. Then followed by mass spectrometry detection.

The SDS-PAGE was used to separate the protein, and stained with Coomassie Blue G-250. The gel bands of interest were cut into pieces. Sample was digested by trypsin with prior

reduction and alkylation in 50 mM ammonium bicarbonate at 37ºC overnight. The digested products were extracted twice with 1% formic acid in 50% acetonitrile aqueous solution and then dried using a speedvac.

Extracted protein was quantified using BCA assays (Thermo Scientific). After reduction, alkylation, and precipitation with acetone, 50 ug of protein was digested using trypsin at 37 ºC overnight. The digested products were desalted using Solid-phase extraction cartridges and then subjected for LC-MS analysis.

For LC-MS/MS analysis, the peptides were separated by a 65 min gradient elution at a flow rate 0.300 µL/min with the Thermo EASY-nLC1200 integrated nano-HPLC system which is directly interfaced with the Thermo Q Exactive HF-X mass spectrometer. The analytical column was a home-made fused silica capillary column (75 µm ID, 150 mm length; Upchurch, Oak Harbor, WA) packed with C-18 resin (300 A, 3 µm, Varian, Lexington, MA). Mobile phase A consisted of 0.1% formic acid, and mobile phase B consisted of 100% acetonitrile and 0.1% formic acid. The mass spectrometer was operated in the data-dependent acquisition mode using the Xcalibur 4.1 software and there is a single full-scan mass spectrum in the Orbitrap (400-1800 m/z, 60,000 resolution) followed by 20 data-dependent MS/MS scans at 30% normalized collision energy. Each mass spectrum was analyzed using the Thermo Xcalibur Qual Browser and Proteome Discovery for the database searching.

Reverse transcription and quantitative real-time PCR (RT-qPCR)

The sample processing involved culturing bacterial cells in L1B medium for 24 hours, followed by transfer to L1 medium for an additional 36 hours. After centrifugation and resuspension, samples were taken at 0 hours and then transferred into various modified L1 media: agar-L1 medium containing 5% agar pieces, GAL-L1 medium supplemented with 0.1 g/L of galactose, NA2-L1 medium with 0.1 g/L of NA2, NA4-L1 medium with 0.1 g/L of NA4, S-L1 medium with 0.1 g/L of agar oligosaccharides, and S10-L1 medium with 1 g/L of agar oligosaccharides, three bottles per cultivation condition for each repetition. The agar oligosaccharides, as shown by Thin Layer Chromatography (TLC), had a chain length greater than NA8 but were uniformly soluble in the culture medium. Timing began upon transfer to the corresponding media, with samples collected at 5 minutes, 30 minutes, 1 hour, 5 hours, and 20 hours post-transfer, followed by rapid freezing in liquid nitrogen.

Based on previous RNA sequencing results, the internal reference genes PL0252 and PL1865 were selected for qPCR analysis. A single pair of amplification primers was designed for each target gene, with three pairs of primers designed for certain genes of special interest. According to the manufacturer's instructions, RNA from the samples was extracted using the FastPure Cell/Tissue Total RNA Isolation Kit V2 produced by Nanjing Vazyme Biotech Co., Ltd. Subsequently, the HiScript III RT SuperMix for qPCR (+gDNA wiper) R323 (Vazyme), was employed for the reverse transcription. ChamQ Universal SYBR qPCR Master Mix Q711 (Vazyme) was employed in each qPCR reaction.

Acknowledgments: We thank Prof. Weicheng Cui and his technical team members of the 2020 joint cruise to Mariana Trench of Westlake University and Shanghai Ocean University. We appreciate Dr. Xiaobo Li, Dr. Ying Zhen and Dr. Yanlei Feng for his advice and assistance in

writing the article. We thank Dr. Mingzhu Fan, Xiaoyan Xu and Jinheng Pan from the Biomedical Research Core Facilities at Westlake University for assistance and discussion during the experiment. Thank the Westlake University HPC Center for computation support. We thank Lei Gao, Yingping Gao and Fang Xiao at the Westlake University Microscopy Core Facility for advice and assistance in microscopy sample preparation and data collection. We Thank Huican Li from Westlake University for his help in the cultivation and collection of Escherichia coli, and thanks to all colleagues in the outbound team, especially Xinyu Huang, for their help in providing the experimental materials needed for sample collection.

Author contributions:

Conceptualization: KNG

Methodology: KNG, KLZ, YYQ

Investigation: KNG, YYQ

Visualization: YYQ, TTY, KNG

Funding acquisition: KLZ

Project administration: KLZ, KNG

Supervision: KLZ

Writing – original draft: KNG

Writing – review & editing: KNG, YYQ, TTY, KLZ

KNG and YYQ made equal contribution to this work.

Competing interests: Authors declare that they have no competing interests.

Data and materials availability: The bacterial strain JK6 in this study will be shared upon request. The specific code and Alphafold predicted structures mentioned in the article are available on the GitHub link https://github.com/jkgkn/JK6. The raw sequencing data can be obtained from SRR27869121, SRR27869122, SRR27869123. Upon request, the corresponding author can provide any further information needed for the re-analysis of the data presented in this paper.

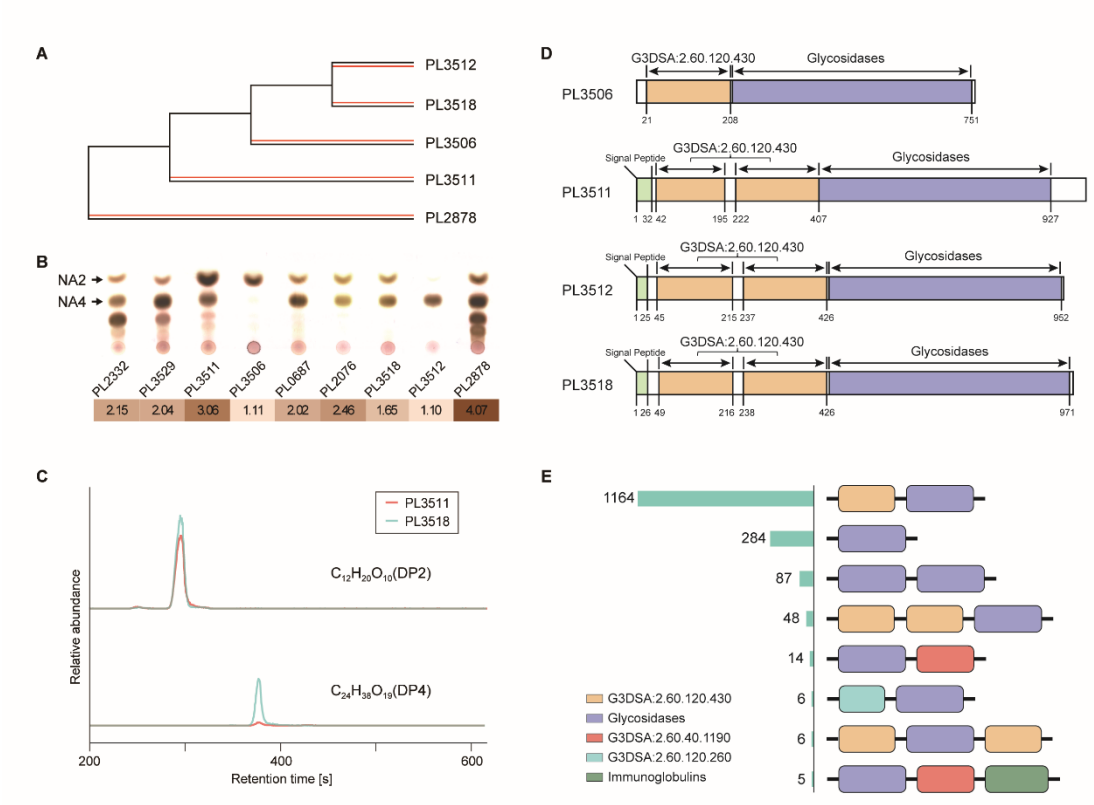


Fig. 1. Gene Structure, Evolutionary Relationships, and Enzymatic Activity of Four Amylases. (A) Evolutionary topological relationships of four amylases and outgroup agarases, supported by phylogenetic trees based on protein sequences and clustering based on structural similarity. **(B)** Thin layer chromatography (TLC) results of the hydrolysis products of agarose by some agarases in the JK6 Strain, with the mobile phase consisting of 1-butanol/acetic acid/distilled water (2:1:1, v/v/v); with bottom numbers indicating the relative values of reducing sugars concentration detected in corresponding samples. **(C)** Polysaccharide content mass spectrometry abundance after 5 hours of reaction with excess agarose for proteins PL3511 and PL3518, showing significantly less accumulation of NA4 product in PL3511 compared to PL3518, indicating divergence in hydrolytic activity between the two enzymes. **(D)** Gene3D annotation of the four proteins' structural domains, with the G3DSA:2.60.120.430 domain marked as a CBM-like (carbohydrate-binding module-like) domain in Pfam annotations. **(E)** Structural domain composition types and corresponding quantities of GH50 family proteins in the CAZy (carbohydrate-active enzymes) database.

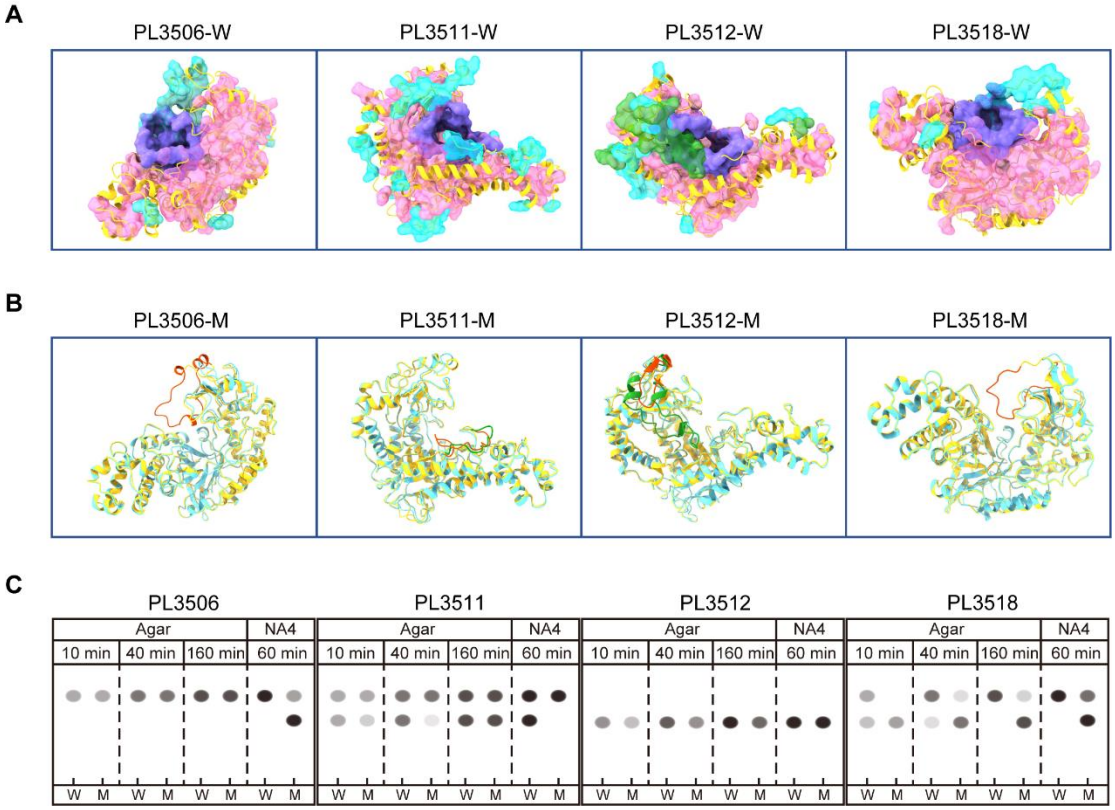
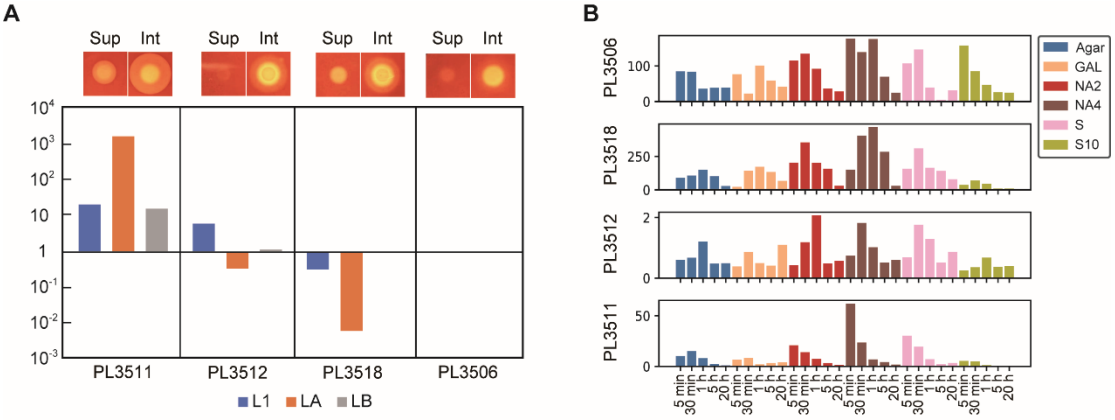


Fig. 2. Structural Comparison of Four Proteins and Changes in Enzymatic Activity After Directed Modification. (A) Differences (light blue), conserved regions (pink), and active sites (purple) in the structural comparison of each enzyme with the other three enzymes; **(B)** Structural regions of directedly modified enzymes (modification sites marked in red); **(C)** Changes in hydrolytic activity towards two substrates for the four enzymes and their directed mutants, with "W" representing the wild-type enzyme and "M" denoting the modified mutant, the grayscale spots in the image represent hydrolysis products, with the depth of grayscale indicating the concentration of the products. In PL3506, which produces NA2 as a product, the differing loop from PL3512, which produces NA4, was removed. According to previous analyses of GH50 family protein structures, this loop structure is highly variable in sequence and its presence or absence does not clearly correlate with whether the enzyme produces NA2 or NA4 as hydrolysis products. After modification, PL3506's rate of cutting NA4 significantly decreased, but the types of agarose cleavage products remained unchanged, only producing NA2, as predicted. Similarly, after removing the excess loop in PL3518, its rate of cutting NA4 also significantly decreased, and more NA4 accumulated when cleaving agarose, suggesting the loop structure may only affect the binding or release of polysaccharides with agarase, thus influencing enzyme activity against different products. Additionally, experiments removing the CBM domain in PL3506 found a significant reduction in hydrolysis rate of NA4, and a change in agarose hydrolysis products, indicating the CBM domain adjacent to the glycoside hydrolase domain plays an important role in supporting the active center and enriching polysaccharide binding.



C

SeqID	Localization(PSORTb Score)	Prediction(SignalP-6.0)	Predicted location(Cell-PLoc 2.0)	CELLO Prediction	Localization Class (LocTree3)
PL3506	Cytoplasmic (8.96)	NO SP	Cell inner membrane. Extracell.	Extracellular/ Cytoplasmic	cytoplasm
PL3511	Extracellular (9.65)	signal peptide	Extracell.	Extracellular	secreted
PL3512	Unknown (2)	lipoprotein signal peptide	Extracell.	Periplasmic/ Extracellular	secreted
PL3518	Unknown (2.5)	signal peptide	Cell inner membrane.	Extracellular	secreted

Fig. 3. Localization and Expression Changes of Four Proteins (A) The top section shows the differences in agarolytic activity on agar plates of supernatant concentrated proteins (Sup) and cell lysates (Int) during the expression of the four enzymes in Escherichia coli. The bottom section compares the mass spectrometry ratios of supernatant to intracellular proteins under three different culture conditions (L1 medium, L1A medium, L1B medium). No mass spectrometry detection in the supernatant for 3518 under L1B culture and for 3506 under all culture conditions, hence blank. Notably, the ratio of intracellular to extracellular accumulation shows significant changes under different culture conditions. Especially under agarose induction, PL3511 and PL3518 respectively show a higher and lower proportion of extracellular secretion. **(B)** qPCR results of the four proteins cultured under five different sugar conditions (agar-L1 medium with 5% agar blocks, GAL-L1 medium with 0.1g/L galactose, NA2-L1 medium with 0.1g/L NA2, NA4-L1 medium with 0.1g/L NA4, S-L1 medium with 0.1g/L agar oligosaccharides, S10-L1 medium with 1g/L agar oligosaccharides). **(C)** Localization predictions for the four proteins using different software.

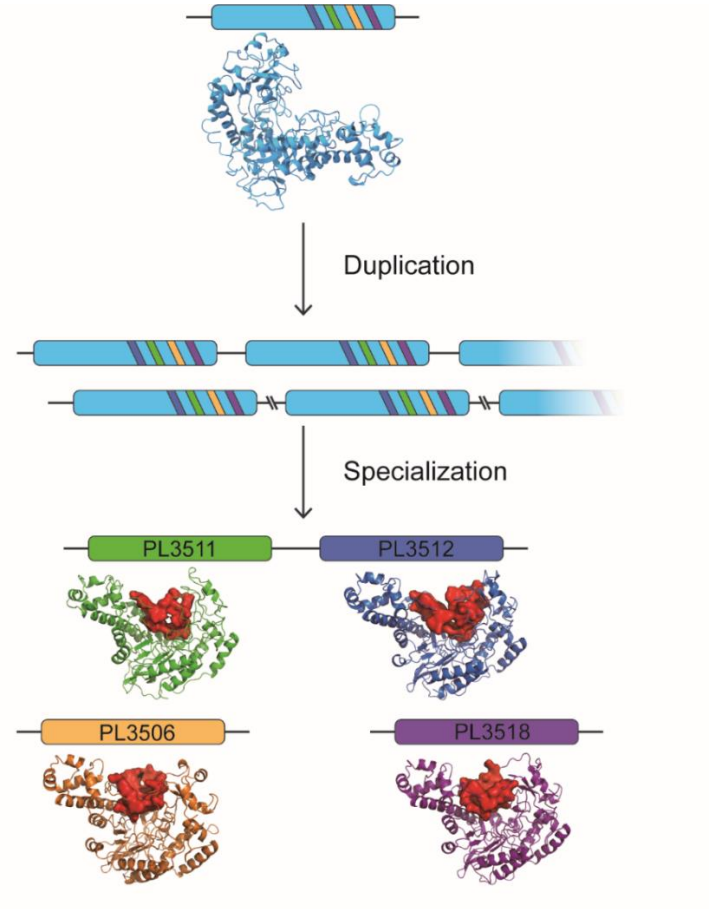


Fig. 4. Protein Duplication and Divergence Model. An ancestral gene with multifunctional activity may be duplicated and retained under selective pressures such as dosage requirements, and subsequently each duplicated gene copy diverges through mutations. From the perspective of the final products, some genes tend towards specialization, led to the resolution of the conflict between the transcriptional regulation of the two functions, and gaining advantages in the precise regulation of enzyme activity and expression levels, thus gradually forming an enzyme control network for hydrolyzing the substrate. During the process of mutation, due to the similarity in enzyme functions, constraints are observed in the protein's 3D structure, whereby although there are significant changes in sequence similarity, the main structure of the protein remains consistent, while major variations in structure occur at the enrichment of the enzyme active center. As we explore further into the intricacies of enzyme evolution, particularly within the GH50 family of glycoside hydrolases, find that they maintain a fundamentally consistent spatial structure similarity, only showing local structural differences. This pattern is akin to bird nests built on different trees in a forest: while the basic form of the trees is consistent, the position and type of nests vary among trees due to the differences in birds. And may be because facing structurally diverse polysaccharide substrates, highly specific enzymes are needed to function effectively, and given the relatively small variety of basic units and linkages in polysaccharides, different glycoside hydrolases only need slight adjustments to tackle various polysaccharide structures.

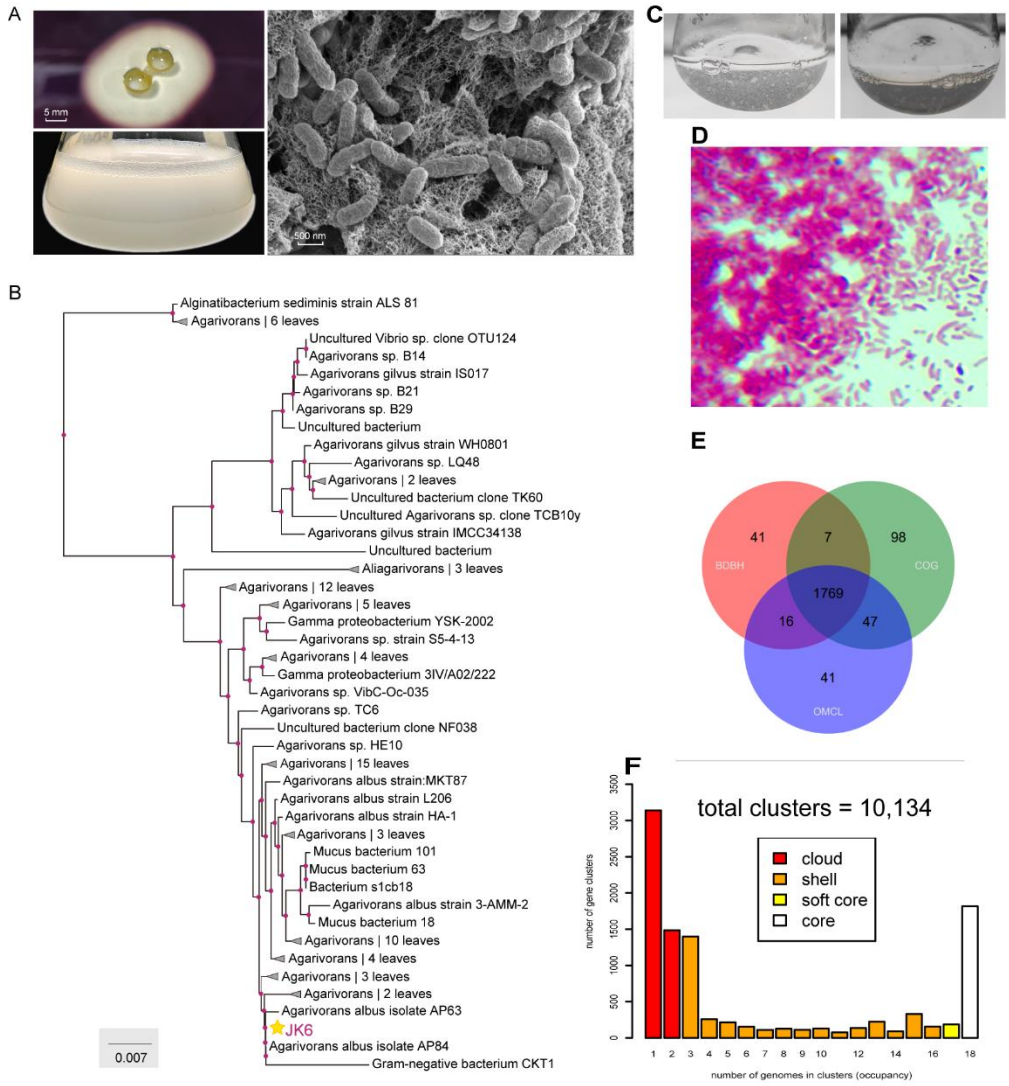


Fig. S1. Physiological Culturing Characteristics and Pangenome Analysis of the JK6 Strain. (A) The strain forms hydrolysis circles on L1 medium plates, with transparent staining, and can grow rapidly and in large quantities in L1B medium, reaching concentrations up to the order of 1×10^9. Scanning electron microscopy shows the bacteria to be short rod-shaped, with a wrinkled surface, closely adhering to the agar surface.**(B)** Amplification and sequencing of the 16S rRNA sequence of the JK6 strain followed by tree construction using NCBI BLAST identified it as a member of the genus *Agarivorans*, and together with the growth phenotype, confirmed it to be *Agarivorans albus*.**(C)** Culturing overnight in L1A liquid medium with added 0.2% agar shows bacterial clumping into white patches, with the solution becoming clear and transparent after a week.**(D)** The bacteria show a negative result with Gram staining.**(E, F)** Pangenomic comparison detailing the shared genes among 17 closely related strains within the same genus and the JK6 strain.

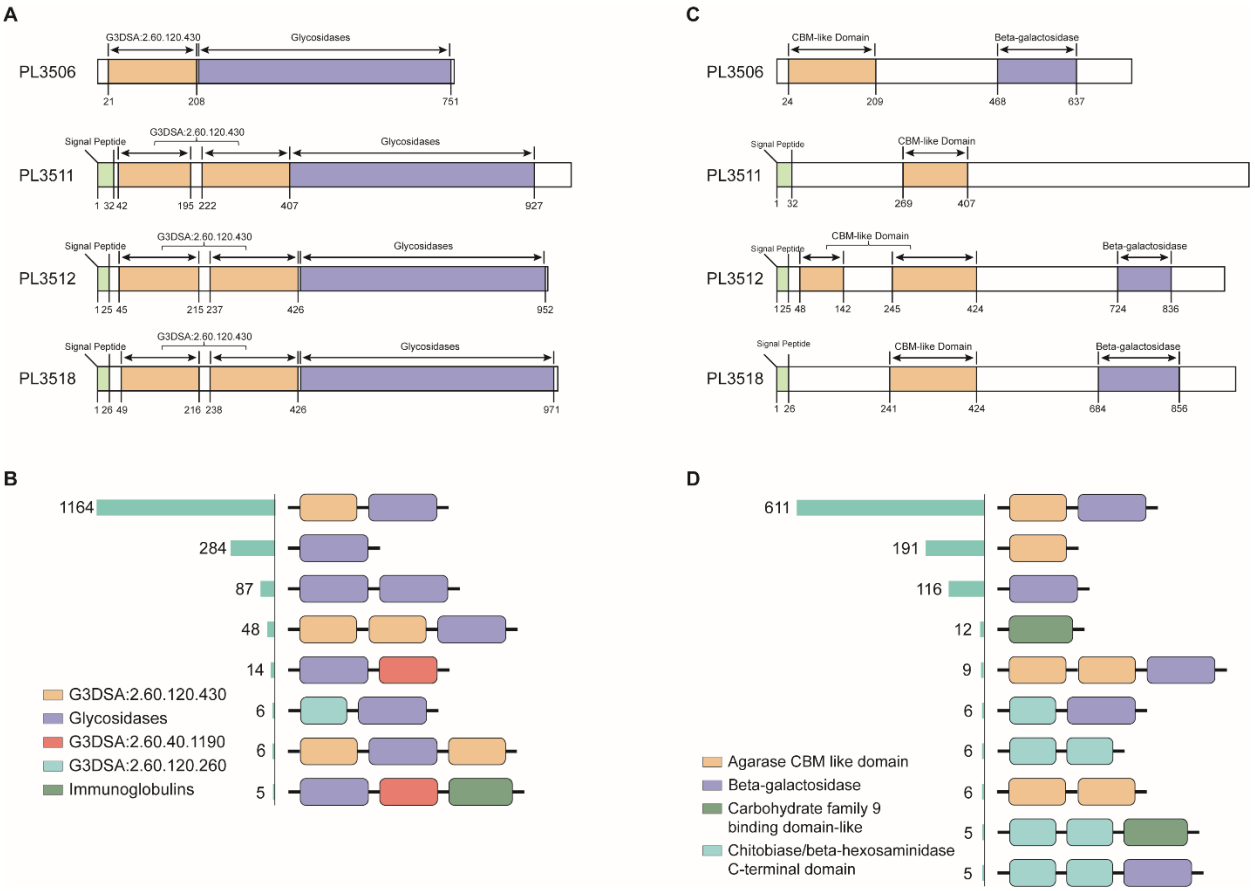


Fig. S2. Gene3D and Pfam annotations reveal the structural domain scenarios for the four proteins and genes of the GH50 family. (A) Gene3D annotation of the structural domains for four genes. **(B)** Gene3D annotation of the structural domains for the GH50 family. **(C)** Pfam annotation of the structural domains for four genes. **(D)** Pfam annotation of the structural domains for the GH50 family genes. In terms of gene structure, despite differences in annotations between Gene3D and Pfam, the AlphaFold predicted structures suggest that Gene3D's predictions are more accurate. According to the annotation statistics for GH50 family proteins, the combination of dual CBM-like domains(G3DSA:2.60.120.430) with a glycoside hydrolase domain is relatively rare in the GH50 family.

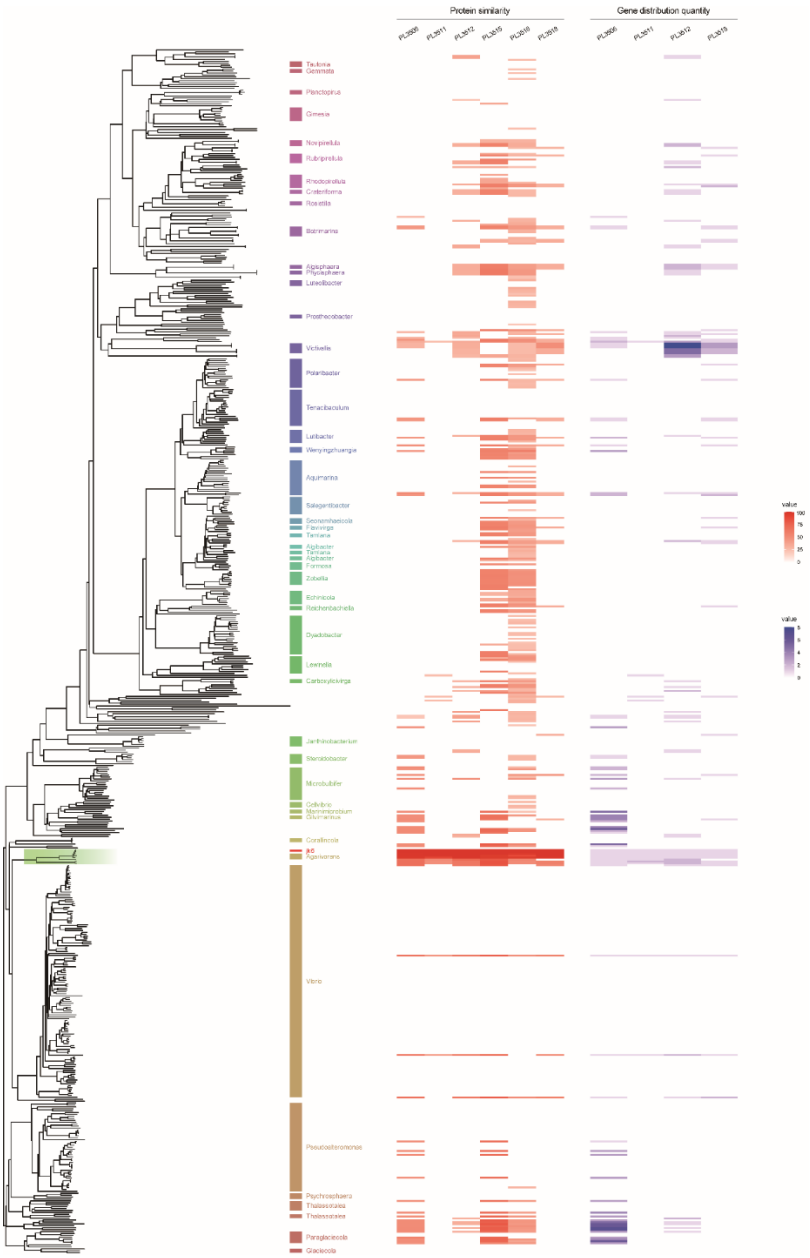


Fig. S3. The number of homologous genes and the highest homology of four proteins in 629 bacterial strains. On the left is the phylogenetic tree, with text annotations indicating the topological structure showing three or more consecutive genus names. In the middle are the highest protein similarities for the four proteins and PL3515/PL3516 within the gene clusters of the four proteins. On the right is the number of occurrences of the four proteins in each species. The homologs of these four proteins vary in quantity across different clades according to evolutionary relationships. The homologous genes of proteins PL3506 and PL3512 are primarily distributed in different clades and often do not coexist in the same species. Based on the inferred phylogenetic topology of the four genes, we observed that the earliest diverged PL3511 has the least number of homologs across species, while the transitional PL3518 has slightly more, but still significantly fewer than PL3506 and PL3512.

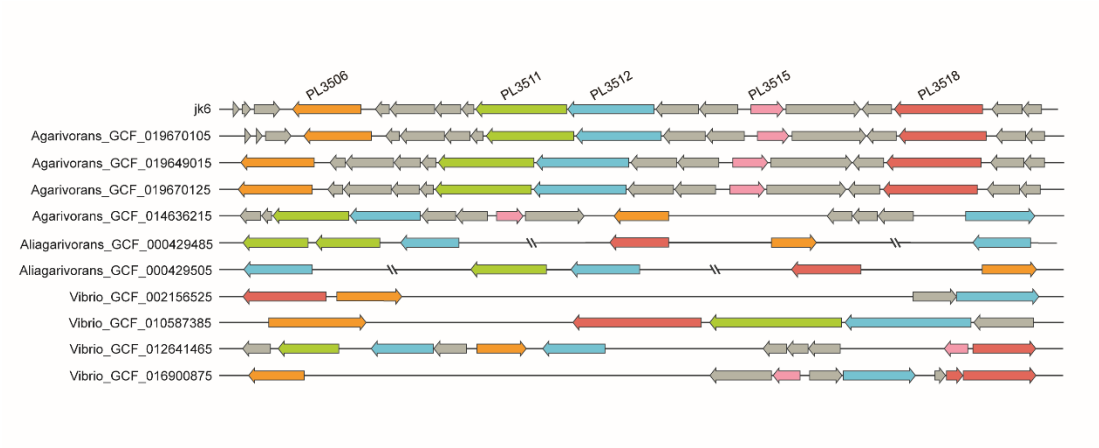


Fig. S4. The gene cluster structures in several representative species containing the four proteins. The four proteins and PL3515 are marked with different colors, and double slashes indicate discontinuity in the genome. It is observed that the gene structures are largely consistent within the genus *Agarivorans*, and the genes PL3511 and PL3512 still retain a tandem duplication pattern adjacent to each other, although structural changes have occurred in one strain. In strains from other genera, significant structural variations are evident.

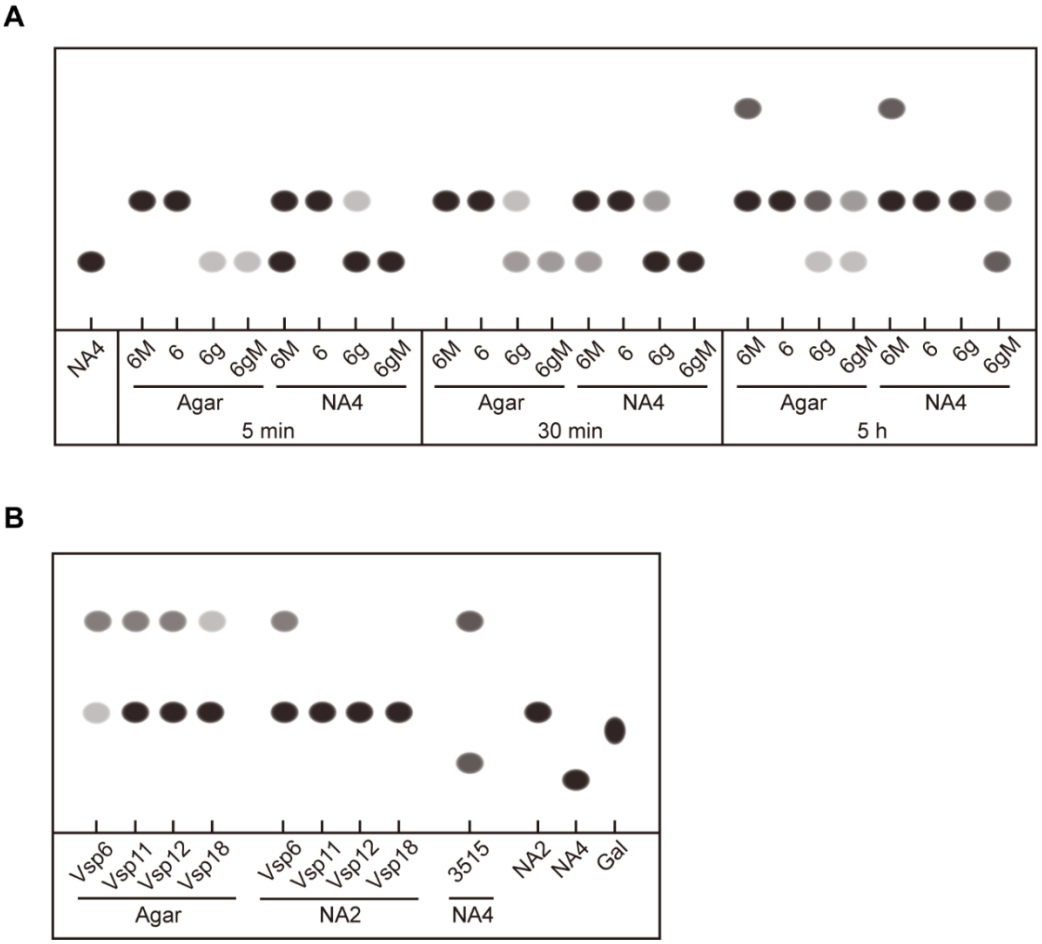


Fig. S5. Hydrolytic activity of mutants of the PL3506 protein and the four homologous proteins in the *Vibrio astriarenae* strain. (A) 6M represents a mutant protein obtained by point mutation of the complete PL3506 protein, 6 represents the complete PL3506 protein, 6g represents the protein expressed by the glycoside hydrolase domain of PL3506, and 6gM represents the protein of the same site point mutation in the glycoside hydrolase domain of PL3506. It shows that 6M can hydrolyze both agar and NA4 into monosaccharide products over time, but the hydrolysis rate for NA4 is significantly slower than that of the original protein. 6g has a slower rate of hydrolyzing NA4 and produces NA4 bands during the hydrolysis of agar, which were not present in the original protein. The hydrolytic activity of 6gM is similar to 6g, but the hydrolysis rate for NA4 is significantly slower than 6g. Within the observed time, NA4 in the substrate is not completely consumed, leaving it uncertain whether it also produces monosaccharide hydrolysis products similar to 6M.**(B)** Vsp6 to Vsp18 represent the homologous proteins in *Vibrio astriarenae* (HN897) corresponding to PL3506-PL3518. Among them, all four proteins in HN897 produce weak monosaccharide bands, but in comparison to the hydrolysis of NA2, except for Vsp3506 which can hydrolyze NA2, the other three proteins do not have this activity, indicating their monosaccharide production is due to other activities. The grayscale spots in the image represent hydrolysis products, with the depth of grayscale indicating the concentration of the products.

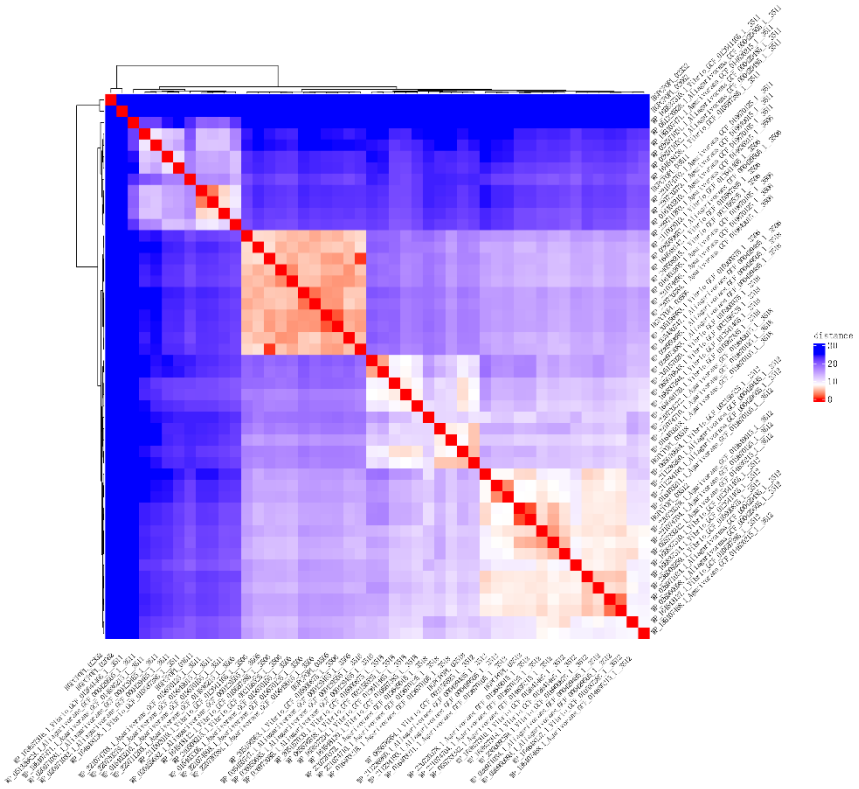


Fig. S6. Glycoside hydrolase domain structural similarity clustering heatmap. Construct a heatmap using the Vd values calculated from the eigenvalues obtained through the alignment of the glycoside hydrolase domains of the compared proteins using USalign. The heatmap shows that the four proteins are distinctly divided into four units.

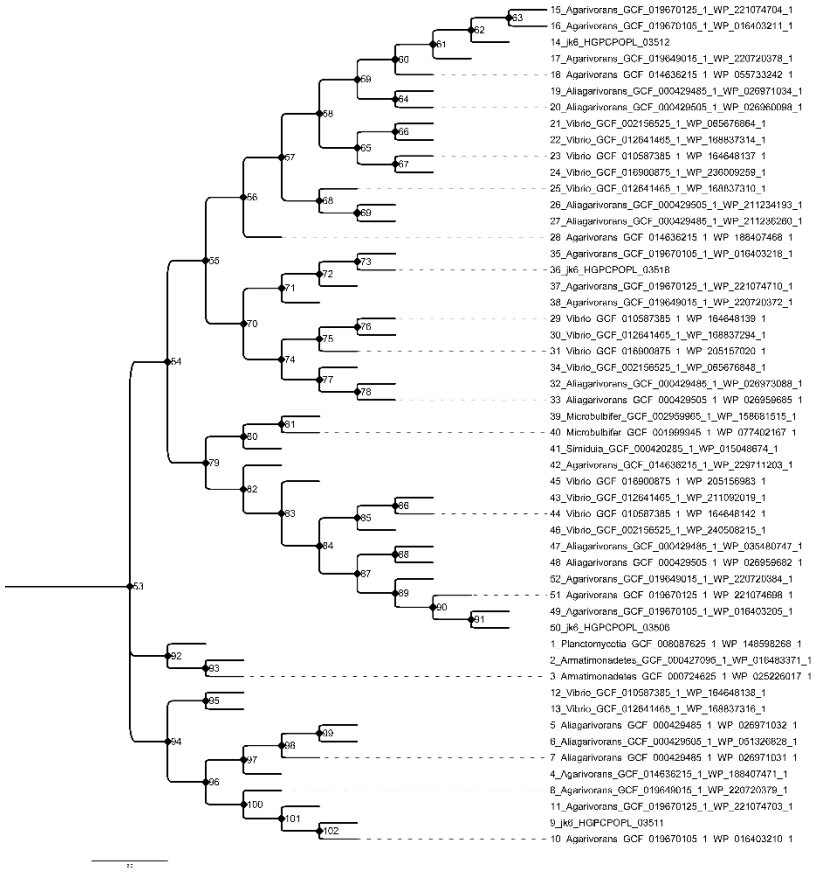


Fig. S7. PAML-reconstructed ancestral tree of the four proteins. It was found that the ancestors related to PL3511, PL3512, and PL3518 essentially maintained the same protein localization as their corresponding contemporary proteins. However, the ancestral proteins of PL3506 showed different localization patterns. The numbers marked at each node represent the ancestral protein numbers, consistent with the numbering in the ancestral protein localization table (Data S6).

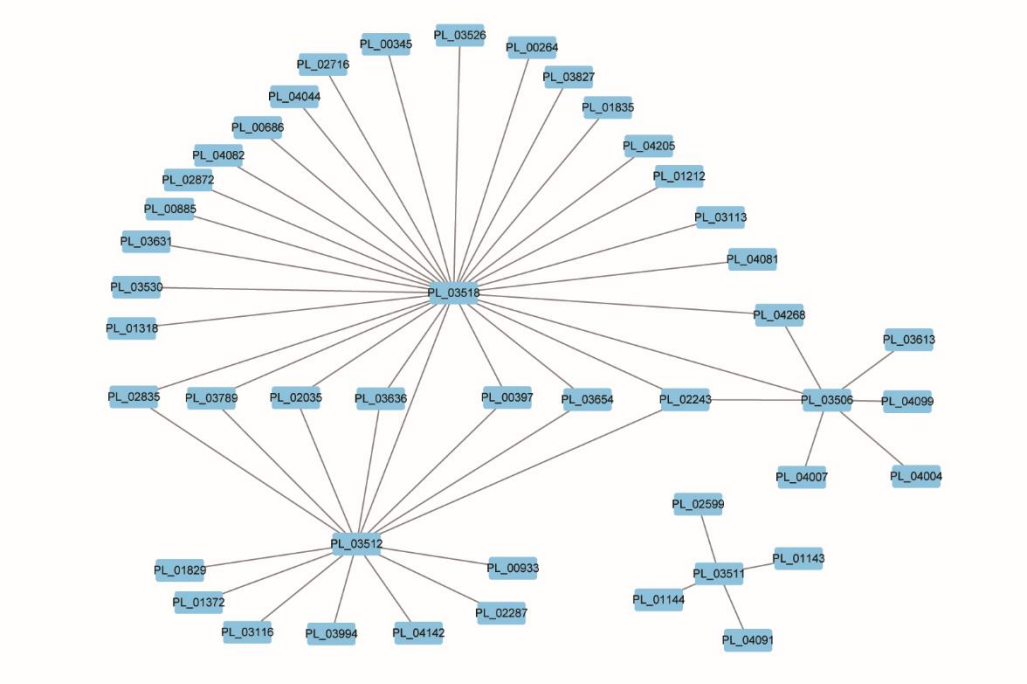


Fig. S8. Co-evolutionary history between the four proteins and the shared proteins among the collected species. Covarying evolutionary rates between pairs of proteins, calculated using PhyKit, with a threshold of 8.0 used to filter protein pairs related to the four proteins. Within the set threshold range, we found that PL3511 does not exhibit significant correlation with the other three agarase genes (PL3506, PL3512, and PL3518). However, PL3511 shows a co-variation relationship with enzymes related to sugar metabolism, such as carbonic anhydrase and glycosyltransferase. Among PL3506, PL3512, and PL3518, we discovered that they are related to a common associated gene, the glycoside/cation symporter (PL2243). Additionally, the highest association of PL3506 is with the Type IV pilin protein (PL3613), which is related to host adhesion, and it is associated with the greatest number of proteins related to phosphate metabolism processes. For PL3512, the protein with the highest association is the periplasmic transport-related protein PL1829 (Tol-Pal system protein TolQ), which matches its cell localization, and it is associated with the greatest number of proteins related to serine hydrolase activity and peptidase activity. Meanwhile, PL3518 has the highest association with PL3512, followed by the protein PL2243, which is related to all three enzymes; the highest independent association is with Probable peptidoglycan glycosyltransferase (PL3631), associated with the greatest number of proteins related to catalytic activity and hexosyltransferase activity.The network diagram is constructed using Cytoscape to illustrate the associations.

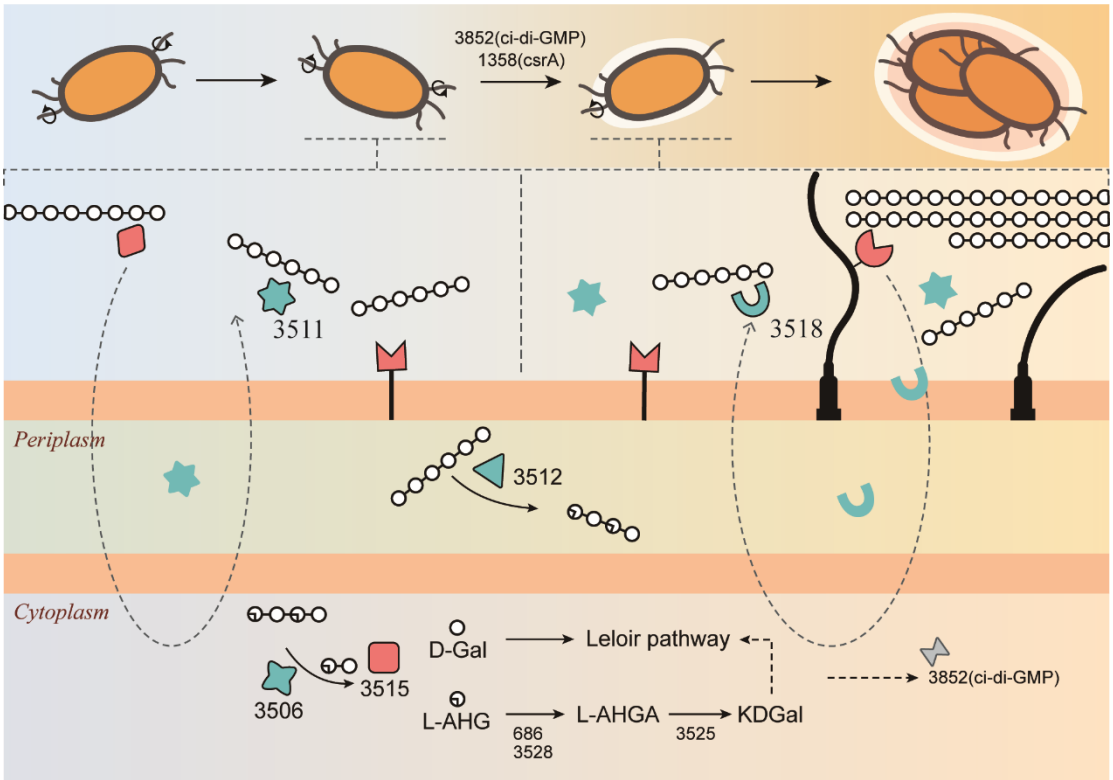


Fig. S9. Regulatory model of JK6 strain's perception and response to agar induction. Macroscopically, upon sensing the presence of agar and its abundance in the environment, the bacteria transition from individual swimming states to multicellular aggregation, forming biofilms that adhere to the surface of agar bodies. On a molecular level, in the free-swimming state, certain agarases are secreted extracellularly to explore and sense the presence of agar in the environment. When encountering low levels of agar, certain agarases are predominantly induced to express in specific ratios for the degradation and absorption of agar. Conversely, in the presence of high concentrations of agar, other agarases take precedence or are expressed in different ratios to regulate the absorption and utilization of agar. Intracellular PL3506 hydrolyzes them into disaccharides, which are then hydrolyzed into monosaccharides by PL3515, entering subsequent metabolic pathways like AHG metabolism and the Leloir pathway. The process of agar absorption and utilization is accompanied by the induction of signals such as ci-di-GMP, leading to the transition of bacteria from a free-swimming to an attached state.

Table S1.

List of Primers Involved in the Experiment.

ID #	Assay	Description	Sequences (5'-3')
PL2332.FOR	Recombinant protein expression in E.coli	Forward primer for PL2332 cloning into pPEI-His-Sumo	GCACCCACACGCGCGAATTCATGACTATCAAGGCAAAGAAAATGGC
PL2332.REV	Recombinant protein expression in E.coli	Reverse primer for PL2332 cloning into pPEI-His-Sumo	TGGTGGTGGTGGTGCTCGAGTTAGCGTTGTGGCGGGGC
PL3529.FOR	Recombinant protein expression in E.coli	Forward primer for PL3529 cloning into pPEI-His-Sumo	GCACCCACACGCGCGAATTCATGAAAGGATTCACTAAGCACCC
PL3529.REV	Recombinant protein expression in E.coli	Reverse primer for PL3529 cloning into pPEI-His-Sumo	TGGTGGTGGTGGTGCTCGAGCTACTGGAATTTAAAACGTTGGTTGG
PL3506.FOR	Recombinant protein expression in E.coli	Forward primer for PL3506 cloning into pPEI-His-Sumo	GCACCCACACGCGCGAATTCATGGAAACGCGCTTAAGGGC
PL3506.REV	Recombinant protein expression in E.coli	Reverse primer for PL3506 cloning into pPEI-His-Sumo	TGGTGGTGGTGCTCGAGCTACTTTTTCGCCTTATTAAAACGG
PL0687.FOR	Recombinant protein expression in E.coli	Forward primer for PL0687 cloning into pPEI-His-Sumo	GCACCCACACGCGCGAATTCATGCTGTTTAGAAACACTCTATTACTTG
PL0687.REV	Recombinant protein expression in E.coli	Reverse primer for PL0687 cloning into pPEI-His-Sumo	TGGTGGTGGTGGTGCTCGAGTTAATTTACGAATTTAAAGCTTTGA
PL3512.FOR	Recombinant protein expression in E.coli	Forward primer for PL3512 cloning into pPEI-His-Sumo	GCACCCACACGCGCGAATTCATGGTAGAAGTTATGAAATTTACAAAAAAT
PL3512.REV	Recombinant protein expression in E.coli	Reverse primer for PL3512 cloning into pPEI-His-Sumo	TGGTGGTGGTGGTGCTCGAGTTATTTTTTGTAACGAAGCATGTAAACG
PL2878.FOR	Recombinant protein expression in E.coli	Forward primer for PL2878 cloning into pPEI-His-Sumo	GCACCCACACGCGCGAATTCATGCAGTTCATGTTAAAGCGCC
PL2878.REV	Recombinant protein expression in E.coli	Reverse primer for PL2878 cloning into pPEI-His-Sumo	TGGTGGTGGTGGTGCTCGAGCTATTGGCAAGTATAACCTGATACAAC

PL3511.FOR	Recombinant protein expression in E.coli	Forward primer for PL3511 cloning into pPEI-His-Sumo	GCACCCACACGCGCGAATTCATGAAGATTAAATTTTTATCTGCAGC
PL3511.REV	Recombinant protein expression in E.coli	Reverse primer for PL3511 cloning into pPEI-His-Sumo	TGGTGGTGGTGGTGCTCGAGTTACACTTTACGACGTCTTAGTAAAA
PL2076.FOR	Recombinant protein expression in E.coli	Forward primer for PL2076 cloning into pPEI-His-Sumo	CACCCACACGCGCGAATTCATGGAATATATGATGATTCGTAAAAATAATC
PL2076.REV	Recombinant protein expression in E.coli	Reverse primer for PL2076 cloning into pPEI-His-Sumo	TGGTGGTGGTGGTGCTCGAGTTATTCGTCAACGAGTTTCCAAG
PL3518.FOR	Recombinant protein expression in E.coli	Forward primer for PL3518 cloning into pPEI-His-Sumo	GCACCCACACGCGCGAATTCTTGCTTAGTGGTTGTCCAAATATTG
PL3518.REV	Recombinant protein expression in E.coli	Reverse primer for PL3518 cloning into pPEI-His-Sumo	TGGTGGTGGTGGTGCTCGAGCTACTTCACCAGCTCTGGGTAAC
PL3515.FOR	Recombinant protein expression in E.coli	Forward primer for PL3515 cloning into pPEI-His-Sumo	GCACCCACACGCGCGAATTCATGTCAGGGACAGGTAGCAAG
PL3515.REV	Recombinant protein expression in E.coli	Reverse primer for PL3515 cloning into pPEI-His-Sumo	TGGTGGTGGTGGTGGTGCTCGAGTTAATTTGAATTTTGATAAGTTGAAGC
PL3512M.R1	Recombinant protein expression in E.coli	primer for modified protein PL3512-M expression construction	GTGGTCGTAGTCGTGGCCAGTAATGGTTAC
PL3512M(+06).F	Recombinant protein expression in E.coli	primer for modified protein PL3512-M expression construction	CTGGCCACGACTACGACCACTCAAAAGTGGC
PL3512M(+06).R	Recombinant protein expression in E.coli	primer for modified protein PL3512-M expression construction	TTGAACGGCGAGTTTGGTTAGCAACATAGG
PL3512M.F2	Recombinant protein expression in E.coli	primer for modified protein PL3512-M expression construction	TAACCAAACTCGCCGTTCAATGTTTACTTG
PL3506M.R1	Recombinant protein expression in E.coli	primer for modified protein PL3506-M expression construction	GGTTAGCAACGTCGTAATCAATACCGGTAA

PL3506M.F2	Recombinant protein expression in E.coli	primer for modified protein PL3506-M expression construction	TGATTACGACGTTGCTAACCAAACTCGCCG
PL3518M.R1	Recombinant protein expression in E.coli	primer for modified protein PL3518-M expression construction	AACGAGCATCCACCAAATCGGTAAAGCCAA
PL3518M.F2	Recombinant protein expression in E.coli	primer for modified protein PL3518-M expression construction	CGATTTGGTGGATGCTCGTTCAGTGGCTTC
PL3511M.R1	Recombinant protein expression in E.coli	primer for modified protein PL3511-M expression construction	AGTGCTAATGCGTGCGTGGTCTCCAAAAATCCAACCATTTGCCACGTAAG
PL3511M.F2	Recombinant protein expression in E.coli	primer for modified protein PL3511-M expression construction	ACCACGCACGCATTAGCACTGGTAACGACTACTGGGGACCACTTCCTGA
PL3506g.FOR	Recombinant protein expression in E.coli	Forward primer for PL3506 glycoside hydrolase domain cloning into pPEI-His-Sumo	GCACCCACACGCGCGAATTCGGTTGTTTAGATAAGTTTGG
PL3506g.REV	Recombinant protein expression in E.coli	Reverse primer for PL3506 glycoside hydrolase domain cloning into pPEI-His-Sumo	GGTGGTGGTGGTGCTCGAGTTAACGGCGCTGATACATGCC
PL3506gM.F	Recombinant protein expression in E.coli	Primer for point mutant of PL3506 glycoside hydrolase domain	TTGATAATGGGAAAAGCTGGGGCCGAATGG
PL3506gM.R	Recombinant protein expression in E.coli	Primer for point mutant of PL3506 glycoside hydrolase domain	CAGCTTTTCCCATTATCAATGAAAATACCC
0252qF	qPCR	qPCR primer of PL0252	TGGTCACTCAACAAGCCTCT
0252qR	qPCR	qPCR primer of PL0252	AGCAAAGGCATCAAGGCAAA
1865qF	qPCR	qPCR primer of PL1865	TTGTTGGCGGAATACGTCAC
1865qR	qPCR	qPCR primer of PL1865	ACTGCCAGTACAGCAGTGAT
3506qF	qPCR	qPCR primer of PL3506	CGATTACTGGGCGTTCGTTT

3506qR	qPCR	qPCR primer of PL3506	TACATGCCGCCATTTAACCG
3506qF2	qPCR	qPCR primer of PL3506	CAGATCCTTTCGACCCGGTA
3506qR2	qPCR	qPCR primer of PL3506	TGGTGAACCTTGCATTTCGG
3511qF	qPCR	qPCR primer of PL3511	TATGCAAACGGTGACGTTCC
3511qR	qPCR	qPCR primer of PL3511	TAGCCGCTGCCAATTTGTTT
3511qF2	qPCR	qPCR primer of PL3511	GATGGCGAAACAGTCACTCC
3511qR2	qPCR	qPCR primer of PL3511	CCCACCATCAGGTCGTAAGT
3511qF3	qPCR	qPCR primer of PL3511	ATTATGCGTCTTGGGATGCG
3511qR3	qPCR	qPCR primer of PL3511	AGCAAATTGCCAAGCAAGGA
3512qF	qPCR	qPCR primer of PL3512	ACCCACAAGGCAACTTGTTC
3512qR	qPCR	qPCR primer of PL3512	GCGAACTTCAGAAGCCACTT
3518qF	qPCR	qPCR primer of PL3518	CGACGCTCTTAAAGCCGATT
3518qR	qPCR	qPCR primer of PL3518	CGCGCACCTAGGTAAAGATG

Movie S1. Structural Alignment of Glycoside Hydrolase Domains of the Four Proteins. Extract the glycoside hydrolase domains of the four proteins based on Gene3D annotation results, load them into PyMOL, and display the aligned structures.

Movie S2. Structural Alignment of CBM-like Domains of the Four Proteins. Extract the CBM-like (Carbohydrate-Binding Module-like) domains from the four proteins based on Gene3D annotation results. Except for PL3506, which contains one, the other three proteins each contain two of these domains. Load and display the aligned structures in PyMOL.

Movie S3. Conserved Sites and Active Center of the Glycoside Hydrolase Domain of PL3506. Within the structures, the active sites predicted by docking are marked in red, the sites that are both structurally and sequentially conserved within the glycoside hydrolase domains of PL3506 and a few other homologous proteins are marked in green, and the conserved sites mutated in the PL3506m mutant are marked in blue. Load and display these marked structures in PyMOL.

Data S1.

Whole genome annotation information and proteome abundance results.

Data S2.

Genome accession numbers of 629 bacterial strains downloaded from NCBI.

Data S3.

Structural and sequence similarities among the glycoside hydrolase domains of the four proteins, some of their homologous proteins, and reconstructed ancestral proteins.

Data S4.

The Ka and Ks values between the four proteins and their respective homologous proteins.

Data S5.

The structural conservation index values between the four proteins and the homologous proteins.

Data S6.

Localization prediction information for the four proteins, their homologous proteins, and reconstructed ancestral proteins.

Data S7.

The gene-gene covariation 、pairwise identity and evolutionary rate.

Data S8.

Partial geographical information of *Agarivorans* genus strain samples

Data S9.

Information on carbohydrate small molecules used for molecular docking.