Structural Phylogenetics Reveal Relationships Between Motility and Phage Defence Systems in Bacteria

JIAHE ZHANG
Student ID: 287351052

Supervisors:
Nicholas J. Matzke
Caroline Puente-Lelievre

School of Biological Science,
The University of Auckland,

A thesis submitted in fulfilment of the requirements for the degree of
Master of Science in Biological Science,
the University of Auckland, 2025.
This thesis is for examination purposes only and is confidential to the examination process.

Abstract

Bacterial flagella are powered by membrane-embedded motor complexes such as MotAB, which utilize proton motive force to generate rotational motion. These complexes share structural and functional features with homologous systems like ExbBD and TolQR. In this study, we investigated the evolutionary relationships among these known systems and two lesser-characterized complexes: ZorAB and gliding motility proteins such as GldLM. We employed a combined dataset of amino acid sequences and 3-dimensional interaction(3Di) structural encodings to enhance phylogenetic resolution, which captures residue-level tertiary interaction patterns. Using model-based approaches, including IQ-TREE and BEAST, we reconstructed robust phylogenies that consistently resolve ZorAB as a homologous group to MotAB, forming a well-supported sister clade. In contrast, gliding proteins emerged as a more phylogenetically distant lineage, suggesting a likely independent origin. These findings support the inclusion of ZorAB within the broader family of proton-conducting motor systems, highlighting gliding machinery's structural and evolutionary distinctiveness. This study demonstrates the value of incorporating structural character data such as 3Di into molecular phylogenetics, offering new opportunities to clarify deep evolutionary relationships among diverse bacterial motility and energy transduction systems.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my heartfelt gratitude to my two supervisors, Dr. Nicholas and Dr. Caroline, for their invaluable guidance, patience, and encouragement throughout this research. I would not have been able to complete this thesis without their support.

I am also deeply grateful to my family for their unwavering support, understanding, and belief in me during my academic and research journey. Their love and encouragement have always been my greatest source of strength.

I would also like to thank my friends for their companionship, motivation, and countless moments of inspiration, which made this journey abroad all the more meaningful and memorable.

The author and advisors recognised funding support as follows. CPL and NJM were supported in part by grants HFSP RGY0072/2021 and FoS RDF #3732317, as well as RDF-UOA2104 and 18-UOA-034 to NJM.

The author wish to acknowledge the use of New Zealand eScience Infrastructure (NeSI) high performance computing facilities, consulting support and/or training services as part of this research. New Zealand's national facilities are provided by NeSI and funded jointly by NeSI's collaborator institutions and through the Ministry of Business, Innovation & Employment's Research Infrastructure programme. URL: https://www.nesi.org.nz.

Contents

List of Figures

Chapter 1

1 Chapter1 Bacteria flagella stator complex

1.1 Bacterial flagella and motor complex

To survive, bacteria often must escape find favourable conditions (with nutrients or hosts) and avoid unfavourable ones (conditions that are nutrient-poor, or that are surveilled by the immune system). Therefore, bacteria have evolved various forms of motility. In 1676, Leeuwenhoek first discovered the movement of bacteria through a microscope (van Leewenhoeck, 1677). These tiny organisms can run dozens or even hundreds of times their length in one second, far exceeding the fastest animals on earth (Summers &, 2022). This miraculous motor ability has aroused the research interest of many microbiologists and biophysicists.

The powerful motor ability of bacteria is due their unique motor organ - flagella. Bacterial flagella are a whip-like motor organ mainly found in Gram-negative bacteria. Flagella not only provide power for bacteria, but also help bacteria invade and adhere to host cells through chemotaxis. Flagella are also involved in biofilm formation, and are also essential structures for bacterial pathogenic potential (Chaban et al., 2015; Haiko & Westerlund-Wikström, 2013). Flagella are also related to the bacterial Type III virulence secretion system, sharing approximately 10 homologous proteins with it (Pallen & Matzke, 2006). Although different species of bacteria have different flagella, their core structures are similar.

Flagella are composed of more than 20 core proteins and are a complex supramolecular complex (Berg, 2003). They include the flagellar filament (made of the protein flagellin, which constitutes roughly 99% of the flagellum by mass), as well as the hook, rod, and basal body consisting of several multimeric rings in the inner membrane, cell wall, and outer membrane where present (Macnab, 2003).

The motor complex, consists of a rotor and a stator. The rotor includes the MS ring and the C ring, which is the core component of the flagellar rotation. The MS ring is located in the inner membrane of the bacteria and is assembled by the FliF protein, which plays a role in connecting the flagellar rood/hook and other structures of the basal body. The C ring is located on the inner side of the MS ring and is composed of FliG, FliM and FliN proteins. These proteins not only participate in the assembly of the rotor, but also achieve rotational drive through interaction with the stator (Tan et al., 2024; Yamaguchi et al., 2021). The stator is composed of a transmembrane structure formed by MotA and MotB proteins, thought to assemble into a $MotA_5MotB_2$ complex. A series of motor complexes (approximately 8) aranged around the outside of the MS ring are responsible for converting the energy of proton or sodium ion gradient into mechanical energy to drive the rotation of the rotor. $MotA_5$ forms the channel part of the stator, while $MotB_2$ is a dimeric transmembrane protein that binds to the peptidoglycan layer in the cell wall, thereby stably anchoring the stator to the cell membrane (Rieu et al., 2022). When protons or sodium ions flow through the channel formed by the MotAB stator complex, they interact with the FliG protein in the C ring, and the torque generated acts on the rotor, causing it to rotate around the base, thereby driving the flagella to rotate and achieve bacterial movement (Rieu et al., 2022).

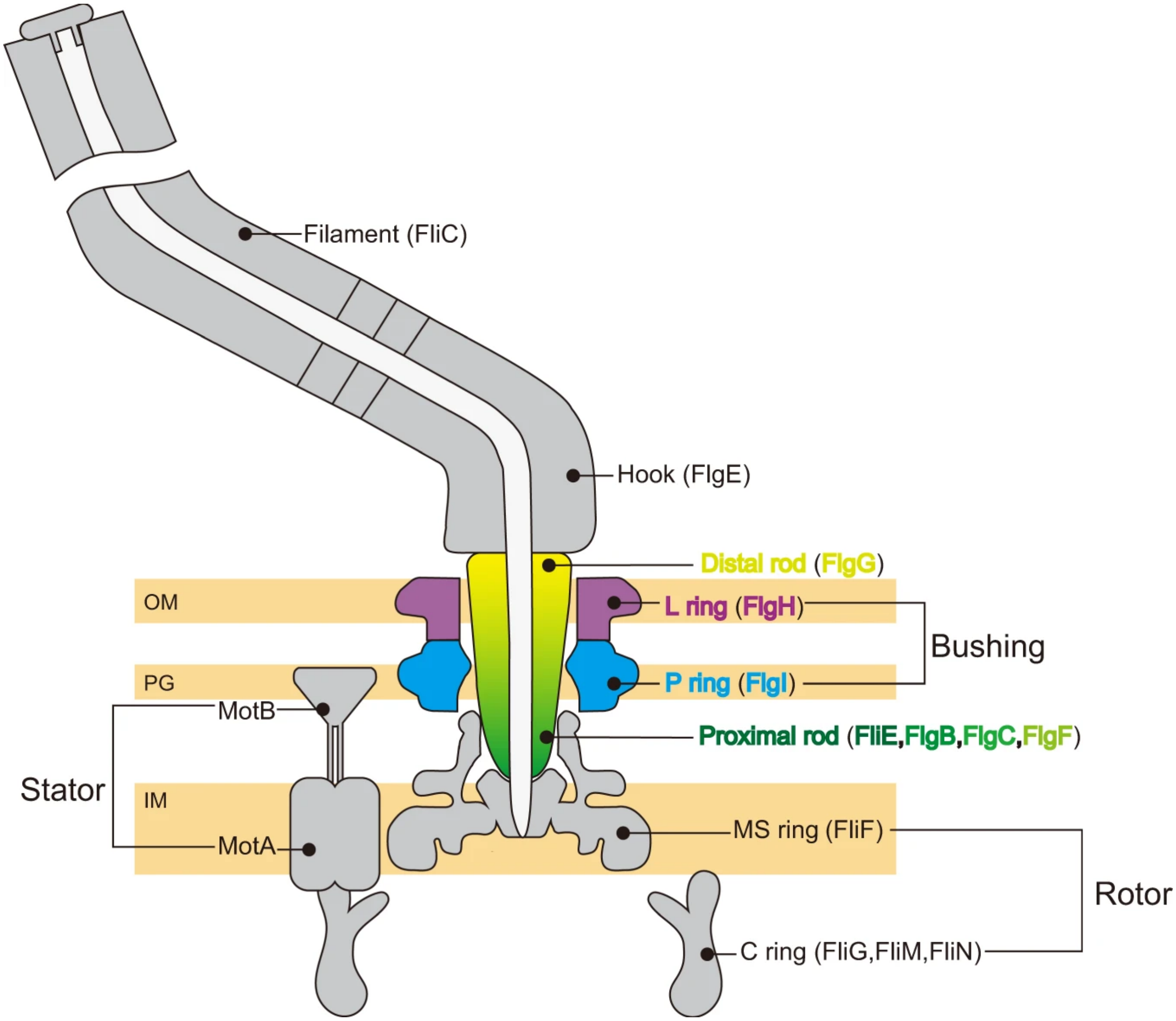

Figure 1: **The main components and structure of bacterial flagella**(Yamaguchi et al., 2021)

The energy-driven mechanism of flagella shows significant diversity in different bacteria, and this difference is closely related to the living environment and ecological adaptability of bacteria. For example, Gram-negative enteric bacteria such as *Escherichia coli* and *Salmonella enterica* mainly rely on Proton Motive Force (PMF) to drive the rotation of flagella (Vilas Boas et al., 2024). The Proton Motive Force is provided by the proton electrochemical gradient across the inner membrane, mainly including membrane potential ($\Delta\Psi$) and proton concentration gradient (ΔpH).

In these bacteria, protons flow through the channel of the stator protein MotA/MotB complex, interact with the charge between the rotor protein FliG, and drive the rotation of flagella. Some bacteria living in high-salinity marine environments, such as *Vibrio cholerae* and *Alteromonas spp.*, rely on sodium ion gradients as the main energy source for flagellar drive (Carroll et al., 2020). The flagellar motors of these bacteria replace the traditional proton-driven MotA/MotB complex with specialized sodium ion-driven stator proteins (such as PomA/PomB or MotX/MotY) (Terashima et al., 2006). The sodium ion gradient is usually maintained by the sodium/proton exchange system or other ion channels on the cell membrane, and the electrochemical energy generated when sodium ions flow through the sodium ion channels is converted into mechanical energy for flagellar rotation (Hu et al., 2023).

1.2 Evolutionary relatives of the flagellar stator complex

Several systems are known to be homologous to the MotAB complex, each thought to share the 5:2 stoichiometry and the function of using ion flow to energise a process conducted by other proteins. We briefly review these here, starting with the two known to be linked to motility.

1.2.1 Flagellar motor stator complex MotAB

The stator complex MotAB was discussed above, so here we focus on mechanism and structure. Its function is to convert the electrochemical potential energy of protons (proton motive force, PMF) into the mechanical rotational force of the flagella (Santiveri et al., 2020). After protons enter the proton channel through the transmembrane region of MotB, they bind to key amino acids in the channel (such as Asp32 in the model organism *E. coli*, strain K12), triggering conformational changes in MotA, resulting in relative displacement of its transmembrane helices (TM3 and TM4) (Mandadapu et al., 2015). This energy is transmitted to the rotor assembly through the cytoplasmic domain of MotA, among which FliG is the core protein that directly interacts with MotA and is responsible for converting the torque of the stator into the rotational motion of the rotor; while FliM and FliN play an important role in the assembly and stability of the rotor and assist in signal transmission during the switching of the direction of flagellar movement (Yamaguchi et al., 2021). MotB anchors the stator complex to the cell wall through its C-terminal peptidoglycan-binding domain (PGBD), homologous to the protein OmpA, ensuring the structural stability of the entire complex during proton flow and preventing proton leakage (Yamaguchi et al., 2021). In addition, the MotAB complex has a highly cooperative dynamic mechanism. When protons pass through the channel, they trigger periodic conformational changes of MotA and MotB. This process is divided into two stages: in the locked state, MotA and MotB form a tight interaction to fix protons and prevent them from leaking; in the power stroke state, proton release triggers the transmembrane helical movement of MotA, transfers energy to the rotor, changes the binding interface with FliG, and finally drives the rotor to rotate (Santiveri et al., 2020).

1.2.2 Gliding motion stator complex GldLM

The gliding motility of Bacteroidetes is independent of flagella. It propels bacteria along solid surfaces through a mechanism that is not completely resolves, but may involve the power of cell surface proteins, the movement of outer membrane microfilaments, or a specific secretion system (Shibata et al., 2023). Although these gliding systems are not related to flagella, they are powered by proteins related to flagellar stator complexes, such as the GldLM system of *Flavobacterium johnsoniae* and the PorLM system of Porphyromonas gingivalis (Hennell James et al., 2021).

GldL has two transmembrane domains in the head and a long enough rod-shaped structure in the cytoplasm at the tail (James et al., 2021). GldM is mainly located in the periplasm, with its N-terminus anchored to the inner membrane through a transmembrane helix. Its periplasmic part consists of four domains and exists as a homodimer. GldLM is a 5:2 motor stator complex consisting of five GldL and two GldM subunits. The truncated version of GldLM (C-terminally truncated GldM) was imaged by cryo-electron microscopy, and a structural model with a resolution of 3.9 Å was obtained. The results showed that the transmembrane helices of GldL form a pentameric cage structure that surrounds and protects the TMH of GldM, while the periplasmic domain of GldM is located at the top of the GldL cage (Hennell James et al., 2021).

1.3 Non-motility-related stator complex

1.3.1 Phage defense system stator complex ZorAB

The Zorya system is a recently discovered new gene family that is widely distributed in bacteria(although found in only a few percent of sequenced genomes) to prevent phage infection. Currently, three subtypes of the Zorya system have been discovered. All Zorya subtypes contain two key membrane proteins, ZorA and ZorB, which are believed to form oligomeric channel complexes and have domains similar to the bacterial flagellar stator unit MotAB (Doron et al., 2018; Payne et al., 2021). As a core component of the Zorya defense system, ZorAB is responsible for sensing and responding to phage invasion. The ZorAB complex consists of five ZorA and two ZorB subunits, forming a unique 5:2 subunit asymmetric complex. ZorAB contains four domains: peptidoglycan binding domain, transmembrane domain, membrane proximal cytoplasmic domain and ZorA C-terminal tail domain. ZorA contains three transmembrane helices (TM1-TM3), of which TM2 and TM3 are close to the transmembrane helices of ZorB, while TM1 is located on the outside and connected to the lipid bilayer to keep the complex stable. The membrane proximal cytoplasmic region of ZorA is composed of three helices (H1-H3) and a β-hairpin structure, and forms a Ca^{2+} binding site between subunits, playing a bridging role. Its massive C-terminal tail structure further extends to the cytoplasmic region for an unknown function, perhaps signal transduction and/or mechanical interference with phage functionality. ZorB is composed of two transmembrane helices and a peptidoglycan binding domain homologus to OmpA. PGBD is located on the periplasmic side and is composed of α helices and β folds. After dimerization, it is anchored to the cell wall to enhance the stability of the complex (Hu et al., 2024). The Zorya system binds to ZorAB The complex senses phage infection as a detector. The transmembrane region of ZorA forms an ion channel, which uses ion gradient to drive its long tail to rotate in the cytoplasm, transmitting the invasion signal to the effector proteins ZorC and ZorD. ZorB anchors the complex to the cell wall through its peptidoglycan-binding domain and locates near the phage DNA injection site. After sensing the phage-induced cell membrane disturbance, the effector proteins ZorC and ZorD are activated, of which ZorC binds to the invading phage DNA, and ZorD acts as a nuclease to degrade the invading DNA, thereby preventing the reproduction of the phage (Hu et al., 2024).

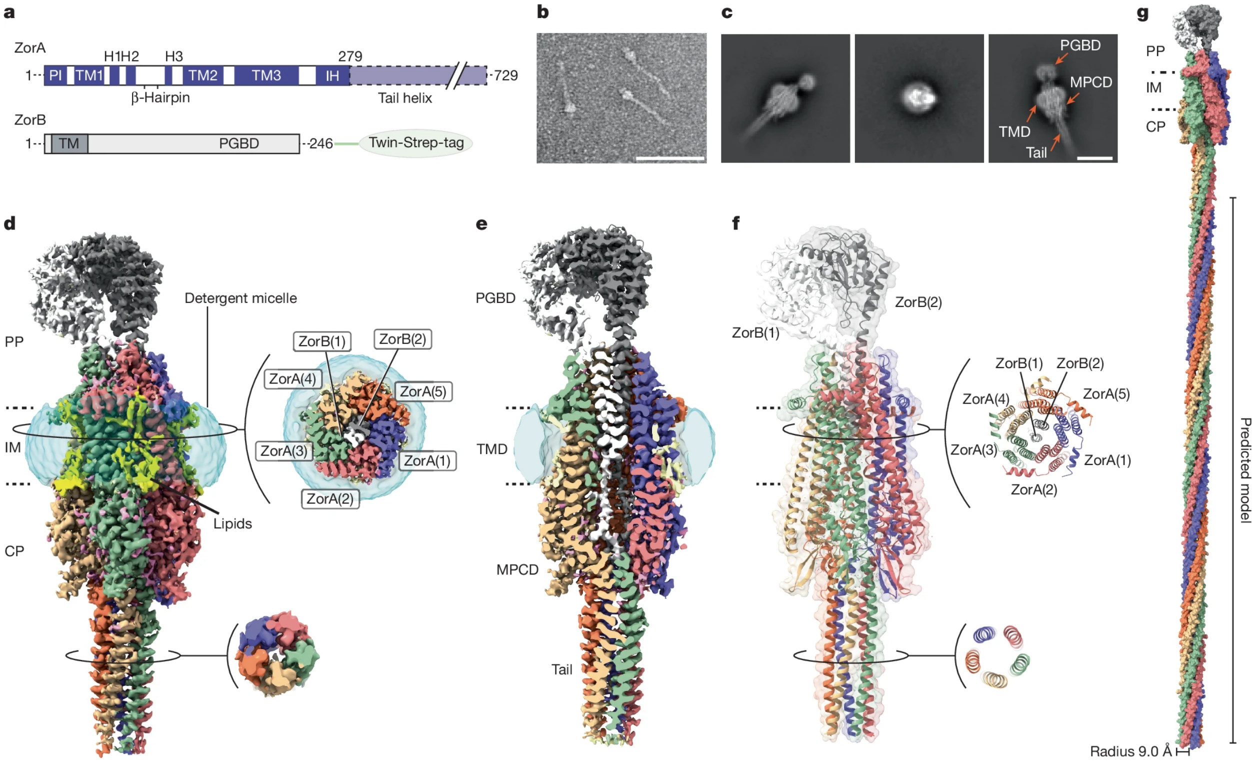

Figure 2: a. Schematic representation of EcZorA and EcZorB.b. Negative-stain EM image of purified EcZorAB particles (scale bar: 1,000Å).c. High-resolution 2D cryo-EM images of EcZorAB and corresponding domain architecture (scale bar: 100Å).d. 3D cryo-EM density map of the EcZorAB complex.e. Cross-sectional view of the density map from the membrane plane.f. Ribbon model of EcZorAB and its cross-sectional views.g. Surface model of the entire EcZorAB complex, with the radius of the ZorA tail indicated.CP: cytoplasm; H: helix; IM: inner membrane; PP: periplasm.

1.3.2 Outer membrane material transport stator complex ExbBD

In Gram-negative bacteria, the outer membrane material transport stator complex ExbBD is a key component of the Ton system, which is mainly responsible for providing energy through the proton motive force (PMF) to promote the transmembrane transport of iron carriers, vitamin B12 and other macromolecules (Marmon, 2013). The Ton system consists of TonB, ExbB and ExbD, among which the ExbBD complex acts as a stator and plays a core role in the energy transfer process. ExbB is a transmembrane protein, whose cytoplasmic domain can interact with TonB, while the transmembrane domains (TMDs) form a proton channel, allowing PMF to act on ExbD through ExbB, thereby driving its conformational changes and completing energy coupling (Ratliff et al., 2022). Cryo-EM studies have shown that the ExbBD complex usually adopts a 5:2 structural arrangement (i.e., 5 ExbBs surround 2 ExbDs), which is similar to the MotAB rotary motor (Celia et al., 2019). ExbB contains 3 transmembrane domains, and its larger cytoplasmic domain helps stabilize the complex and regulates the conformational changes of TonB; while ExbD consists of a single transmembrane domain and a compact periplasmic domain. ExbBD together form a proton channel. Based on the rotation mechanism of MotAB, Santiveri et al. (2020) hypothesized that in the ExbBD system, protons (or hydronium ions) first enter the central pore of the ExbB pentamer from the periplasm and bind to the aspartate residue D25 of ExbD chain Y (white), triggering ExbB to rotate 36° clockwise relative to ExbD. This rotation causes D25 of ExbD chain Z (black) to be exposed to the cytoplasm and release the bound proton, allowing it to enter a new cyclic state. Subsequently, the new proton can bind to D25 of ExbD chain Z, driving ExbB to continue rotating. More data are still needed to support these hypotheses, and further research is needed to explore the mechanism of action of ExbBD.

1.3.3 Membrane-forming stator complex TolQR

The TolQR complex is a core component of the Tol-Pal system, which is widely present in Gram-negative bacteria. It is mainly responsible for maintaining outer membrane stability, participating in the formation of outer membrane vesicles, and assisting in the uptake of plasmid and phage

DNA in some bacteria(Wojdyla et al., 2015). The Tol-Pal system consists of TolQ, TolR, TolA, TolB and Pal. Among them, the TolQR complex plays a key role in the energy transfer process, similar to the TonB-dependent ExbBD complex, which relies on the PMF to drive transmembrane signaling. TolQ is a transmembrane protein with three transmembrane domains. Its cytoplasmic domain interacts with TolR and TolA to mediate energy transfer through PMF(Williams-Jones et al., 2023). As an auxiliary protein, TolR has a short transmembrane region and a cytoplasmic helical structure, which can regulate the conformational changes of TolQ, thereby affecting the activation state of TolA. During cell division, TolQR is involved in energy supply. Petiti et al. (2019) found that TolQ is still localized to the septum even in ΔTolA, ΔTolR,ΔTolB and ΔPal mutants, and TolA maintains a similar localization in the absence of TolQR. However, Pal is only recruited to the division site in the presence of TolA and functional TolQR. Based on this, Petiti et al. (2019) proposed that TolQR is first recruited to the cleavage site to provide energy for TolA to extend to the outer membrane and further recruit TolB and Pal, and that the extension/contraction of TolA may be regulated by the hairpin structure of its periplasmic domain. Like ExbBD, TolQR also needs more experimental data to further prove its operational details.

1.4 Thesis aims

Currently, MotAB, ExbBD, and TolQR are widely considered to be related proteins, collectively referred to as the AQB family. Recent studies have shown that although GldLM has low sequence similarity with AQB family members, it still exhibits a 5:2 stator complex structure similar to that of the AQB family. In addition, ZorAB, a key structure of the recently discovered phage resistance system, also exhibits a 5:2 stator complex structure. These homologies and structural similarities have triggered our in-depth phylogenetic exploration of the structural and functional relationships between the AQB family and proteins such as GldLM and ZorAB.

In this study, we aim to systematically explore the similarities and differences in structure and function between the AQB family and proteins such as GldLM and ZorAB. In particular, we will focus on the structural conservation and variability exhibited by these proteins during evolution. Although we have a relatively in-depth understanding of the structure and function of bacterial flagella, there is still a lack of systematic explanation for the origin and evolution of flagellar-related proteins. Therefore, this study aims to reveal the historical relationship between these protein families through in-depth phylogenetic analysis and explore possible evolutionary scenarios by comparing their structural and functional diversity.

2 Chapter 2 Molecular Phylogenetics—from amino acid to protein

In this chapter, we briefly review traditional molecular phylogenetics with proteins, and then consider how protein structural data, in the form of ”3Di” (3-dimensional) characters from the FoldSeek programme, might be used to enhance traditional amino-acid based phylogenetics.

2.1 Molecular phylogenetics to proteins

Phylogenetics is a science that studies the evolution of life and the relationship between species. It is a key discipline for reconstructing the evolutionary history of life on Earth (Young & Gillung, 2020). Protein phylogenetics focuses on the evolution of protein sequences, aiming to reveal their evolutionary trajectory and functional changes by analyzing the similarities and differences of protein sequences. In the 1960s, with the breakthrough of molecular sequencing technology, scientists gradually realized that protein and nucleic acid sequences can record genetic information and reflect the evolutionary relationship between species. As a result, molecular phylogenetics gradually emerged, and nucleic acids and proteins became essential tools for studying the history of biological evolution. Fitch and Margoliash (1967) pioneering research laid the foundation for studying protein phylogeny by comparing cytochrome c sequences of different species. Since then, bioinformatics has gradually become an essential tool for protein phylogenetics, especially in sequence alignment, homologous sequence detection, and phylogenetic tree reconstruction.

Multiple sequence alignment (MSA) is the core step of molecular phylogenetic analysis. It reveals the similarities and differences between sequences by aligning multiple protein sequences. Accurate alignment is crucial for subsequent tasks such as evolutionary tree reconstruction, conservative site analysis, and function prediction (Chowdhury & Garai, 2017). Based on these alignment results, mathematical and statistical methods (such as maximum likelihood, Bayesian inference, and neighbour-joining methods) are used to construct phylogenetic trees to display the relationship between target molecules intuitively.

A phylogenetic tree is a diagram used to represent the evolutionary relationship between species or genes. It is usually a tree structure, with each node representing a common ancestor and branches representing the evolutionary path from this ancestor to different species or genes (Gregory, 2008). A phylogenetic tree usually consists of three parts: root, branch, and leaf. The root represents the earliest common ancestor, and the leaf represents the existing species or gene. By analyzing the tree's structure, researchers can infer how different species or genes diverge during evolution, which group of species is most similar, and their relationship with other species (Gregory, 2008).

Another unrooted tree differs from the rooted tree in that it does not contain an explicit common ancestor node. Unrooted trees are used to represent the relative relationship between species or genes without involving time order or the starting point of evolution, so they are suitable for studies that lack root information or cannot determine the earliest common ancestor (Baldauf, 2003; Pearson et al., 2013). The construction methods of phylogenetic trees have also evolved with the advancement of computing technology. Early phylogenetic tree construction mainly relied on simple comparative methods such as the neighbor-joining method. However, with the development of statistical methods, maximum likelihood and Bayesian inference methods have become more commonly used construction methods. These methods maximize the likelihood of observed data by assuming evolutionary models of species or genes, thereby accurately inferring their evolutionary relationships (Mar et al., 2005).

With the advancement of large-scale genome sequencing projects, such as the Human Genome Project and subsequent international projects, the number of protein sequences in the database has increased rapidly, greatly promoting the research progress of protein phylogenetics (Hood &

Rowen, 2013). In 2023, the Ensembl database included genome sequences from thousands of species, including millions of protein sequences, which provided a rich data source for protein phylogenetics (Dyer et al., 2025). At the same time, the optimization of computational algorithms has greatly improved the accuracy and efficiency of phylogenetic tree construction. In recent years, algorithms based on maximum likelihood and Bayesian inference have made significant progress in processing large-scale data sets. For example, updates to software tools such as BEAST2 and IQ-TREE have enabled researchers to process tens of thousands of sequences in a shorter time and construct high-precision phylogenetic trees.

2.2 Homology (Orthologs and Paralogs)

In molecular phylogenetics, homology refers to hypothesis that the sequence or structural similarity between different molecules is due to copying from a common evolutionary ancestor. Homology of genes/proteins is usually divided into orthologs and paralogs (Wagner, 1989; Wagner, 2007). Orthologs refer to genes or proteins inherited from common ancestral genes through species differentiation, and these genes usually perform similar functions in different species. Paralogs refer to homologous genes generated by gene duplication (such as gene rearrangement or gene amplification), which may undergo functional differentiation or evolve new functions within the same species or between different species, and usually present different sequence and structural characteristics (Jensen, 2001).

Orthologous genes usually show high structural and functional consistency during evolution. For example, some important functional genes in bacteria, such as antibiotic resistance genes, retain similar functional patterns despite significant sequence differences between different populations. Paralogs usually evolve new functions due to gene duplication events. Gene duplication is an important mechanism for species adaptation and evolution. Paralogous genes after gene duplication may undergo functional reprogramming and become different biological functions (Koonin, 2005).

In homology analysis, significant sequence similarity is usually the key basis for judging whether proteins are homologous, but this similarity is not the only criterion. In fact, protein pairs with low sequence similarity, especially those with pairwise sequence identity between 20–30%, are often difficult to accurately identify their homology through conventional alignment methods. This area in the sequence is called the ”twilight zone”, which we will mention in the following section. Sequences in the twilight zone are difficult to judge homology using sequence similarity, but in structure, these proteins may have extremely similar structures or highly similar components in the protein core region (Chung & Subbiah, 1996). Therefore, current homology analysis is gradually shifting from amino acid sequence to protein structure.

2.3 Sequence-based protein phylogeny

Traditional protein phylogenetics infers homology by comparing different sequences (Higgins et al., 1996). The most commonly used tool is BLAST, which is one of the earliest work platforms (Altschul et al., 1997). It quickly identifies and queries sequences with similar sequences by local alignment of sequences, and is currently the simplest and fastest tool to operate (Ye et al., 2006). With the differentiation of species and the passage of time, gene sequences will undergo continuous mutations, rearrangements, and deletions, resulting in a gradual decrease in sequence similarity, or even complete disappearance between distant species. Therefore, traditional sequence alignment methods face challenges when dealing with species that have been differentiated for a long time. Tools such as HMMER were developed to increase sensitivity. HMMER is a sequence alignment tool based on the Hidden Markov Model (HMM), which is specifically used to detect and align conserved regions in sequences. As a probabilistic model, HMM can handle local and global relationships in sequences and has good targeting when identifying protein sequences with low similarity (Finn et al., 2015).

Although these technologies have made significant progress, sequence-based phylogenetic methods still face some difficult problems. With the long-term differentiation and gene rearrangement of species, the sequence similarity between distant species will be significantly reduced, resulting in a significant reduction in the reliability of sequence alignment results (Philippe et al., 2011). In this case, although they are considered to be the same protein or homologous proteins, the low similarity often makes accurate alignment and phylogeny estimation difficulty. In addition, evolutionary phenomena such as gene duplication and horizontal gene transfer have further complicated the difficulty of sequence alignment over a long period of time. Proteins with similar morphological structures have similar functions in complex systems, but have very low similarity in sequence (Rost, n.d.). This makes the phylogenetic tree based on sequences unable to reflect the true evolutionary history.

2.4 Twilight zone in amino acid sequences

The ”twilight zone” in amino acid sequences refers to the fact that within the range of low sequence similarity, the structure of proteins may still have similarity, but traditional sequence alignment methods are difficult to effectively reveal this. Studies have shown that when the similarity of protein sequences exceeds 30%, they can usually be considered to have similar three-dimensional structures. This phenomenon is caused by the conservation of protein structure over evolutionary time. However, in proteins with sequence similarity between 20% and 30%, traditional alignment methods often decrease in accuracy, even though structural similarity remains (Chung & Subbiah, 1996). Doolittle (1986) first proposed the concept of ”twilight zone” to describe this sequence similarity interval.

2.5 3Di characters based on FoldSeek reveal structural sequences

From a structural level, proteins have four levels of structure, namely primary structure, secondary structure, tertiary structure and quaternary structure (Alberts et al., 2002). The primary structure is the protein amino acid sequence connected by peptide bonds, which determines the basic composition of the protein and its subsequent structural folding. The secondary structure is a specific folding pattern formed by amino acid chains in local areas, such as α-helix and β-fold, which are stabilized by non-covalent interactions such as hydrogen bonds. The tertiary structure is the three-dimensional folding of the protein as a whole, forming a functional three-dimensional structure and maintaining stability through various molecular interactions). The quaternary structure involves the interaction and combination of multiple protein subunits to form a functional macromolecular complex (Rehman et al., 2025).

In order to improve the speed and accuracy of homology search in the Twilight Zone, and especially to make use of structural databases of AlphaFold predictions, the programme FoldSeek (van Kempen et al., 2024) proposes a structural representation method called 3Di(3-dimensional) alphabet. Different from the traditional representation method that relies on the main chain structure, van Kempen et al. (2024) introduced the concept of virtual center (Vc) to define a precise geometric coordinate point for each excitation residue. The virtual center is located on the plane defined by nitrogen atoms (N), α-carbon (Cα) and β-carbon (Cβ) atoms, and the distance between the center and Cα is set to be a multiple of the distance between Cβ and Cα, and the angle between them is 90°. The introduction of this geometric structure allows the protein structure to be further transformed according to the relative position, geometric angle and spatial relationship between the other residues closest to the candidate residues in space. Through this method, the Steinegger team defined 20 different 3Di states, each corresponding to a specific spatial conformation, which can effectively capture the local pattern of the protein in three-dimensional space (van Kempen et al., 2024).

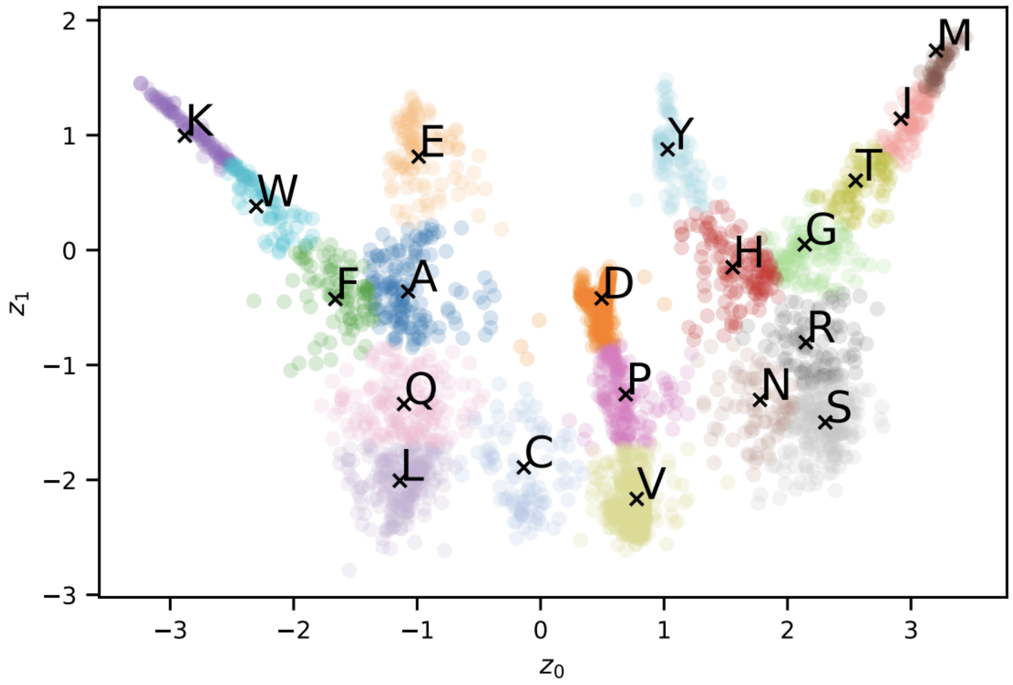

Figure 3:

The distribution of 3Di descriptors compressed into a 2D latent space by the encoder network, with each residue colored according to its assigned 3Di state.(van Kempen et al., 2024)

At the same time, 20 3Di characters correspond to 20 amino acids, which means that 3Di data can be input into homology, alignment, and phylogeny programmes already designed for amino acid data. Foldseek creates a 3Di character for each amino acid position in a protein. For example, K-12 E The amino acid sequence of the MotA protein of coli and the corresponding 3Di sequence:

AA sequence:

MLILLGYLVVLGTVFGGYLMTGGSLGALYQPAELVIIAGAGIGSFIVGNNGKAIKGTLKALPLL

FRRSKYTKAMYMDLLALLYRLMAKSRQMGMFSLERDIENPRESEIFASYPRILADSVMLDFIV

LRLIISGHMNTFEIEALMDEEIETHESEAEVPANSLALVGDSLPAFGIVAAVMGVVHALGSADR

AAELGALIAHAMVGTFLGILLAYGFISPLATVLRQKSAETSKMMQCVKVTLLSNLNGYAPPIA

FGRKTLYSSERPSFIELEEHVRAVKNPQQQTTTEEA

3Di sequence:

DVVLVVLLVVVVCVVVVVCVVVVDDPVLLDDPVLCCQLVVVLVVVLVVVDDPLLVVLLVVC
LVDDALDDLLLLLLVLLLLLVLLVCCVVPNLVSCLCCLVCLCPDPSCVVRVVLVVDVVLSVLLS
LSVVVSVPDDDLVVSLVVVVLVVLLVSLCSSLVSLLVSLVCSLVSLVVQLVVLLVVLVVDPPDD
VVSVVSNSNSVSSNVVSNCCRPVPSNVSSVSSNVSSVSSSLSSVLSSQLSSCVSVPDDSLVSSVS
SLVSDDPSRRDDSVVSVVSSVCSVPVVVVVVVVVVD

Usually, three-dimensional structures are difficult to represent with sequences. 3Di characters can be used to convert the spatial conformation of a specific spiral or folded form into a discrete symbolic representation. Puente-Lelievre, Ridone, et al. (2024) have shown that 3Di traits can be used as standard phylogenetic traits and directly applied to current mainstream phylogenetic inference methods, including the analysis processes of model selection, maximum likelihood tree construction, and bootstrap support assessment. Since these methods were originally designed for conventional sequence data, the natural incorporation of 3Di features indicates their broad applicability. In addition, if the 3Di code does capture structural features that evolve slowly, it may be superior to methods based on conventional unstructured amino acid sequences in terms of the accuracy of phylogenetic inference, especially in distinguishing deeper evolutionary relationships.

2.6 Perspective

Given that traditional methods for comparing amino acid sequences and constructing phylogenetic trees often lead to reduced inference accuracy when dealing with distantly related species due to low sequence similarity, and since structural conservation is higher than AAs sequence, this study proposes to introduce structural information encoded 3Di sequences into the phylogenetic analysis process to improve the resolution and credibility of phylogenetic inference. Specifically, we will attempt to integrate 3Di sequences into traditional amino acid sequence analysis packages and apply them to key steps such as sequence trimming, alignment, and phylogenetic tree construction. The aim is to more clearly resolve the phylogenetic relationship between the ZorAB and AQB-BRD (MotA/TolQ/ExbB-MotB/TolR/ExbD) protein families and further explore the possible evolutionary associations with distantly related GldLM protein systems. In addition, this study will further evaluate the potential of the 3Di method as a novel structural information representation for phylogenetic analysis. By comparing it with the traditional amino acid sequence method, we will explore whether phylogenetic trees constructed based on 3Di sequences have higher accuracy and stability in distinguishing evolutionary relationships, and further verify whether the use of 3Di in combination with amino acid sequences can improve the inference results.

3 Method

3.1 Data accession

To conduct a comprehensive and in-depth analysis of bacterial flagellar motor complexes and their homologous proteins, the dataset includes a substantial number of MotAB proteins, as well as their homologs and potential homologs, encompassing PomAB, MotCD, ExbBD, TolQR, ZorAB, GldLM, among others. Given that MotCD and PomAB are considered evolutionary variants of MotAB present in different species all of these proteins were uniformly categorized under the broader MotAB classification. Puente-Lelievre, Ridone, et al. (2024) selected 193 fully annotated bacterial genome samples from 27 distinct phyla and utilized Escherichia coli K12 MotA as a reference sequence to perform a homology search. This analysis identified many MotA proteins and their homologous sequences across these genomes. Since the MotB protein is typically located downstream of MotA within the same operon, the corresponding MotB proteins and their homologs were co-extracted. Consequently, a preliminary database of 379 MotA and MotB protein sequences was systematically constructed.

The ZorAB protein data originated from a dataset compiled by Hu et al. (2024), which included thousands of ZorAB proteins that had been systematically screened from an extensive collection of bacterial genomes. From this comprehensive dataset, 90 ZorA and ZorB proteins were randomly selected. Based on the classification of Zorya protein types detailed in the file, 30 proteins each from Type 1, Type2, and Type3 Zorya A and Zorya B categories were extracted. The corresponding protein accession numbers were identified, and their respective amino acid sequences were subsequently downloaded from the National Center for Biotechnology Information (NCBI) database (https://www.ncbi.nlm.nih.gov) to ensure reliable and accurate sequence retrieval.

The sequences for the Gld protein dataset were derived from multiple independent and curated datasets (Deme et al., 2020; Hennell James et al., 2022; James et al., 2021; Santiveri et al., 2020; Shrivastava et al., 2013). Homology searches for five representative GldL proteins were conducted using the Foldseek algorithm through its online platform, the Foldseek Search Serverthe Foldseek Search Server the Foldseek Search Server(van Kempen et al. et al., 2024; https://search.foldseek.com). The detailed procedure is outlined as follows:

Each of the five GldL protein sequences was uploaded individually to the Foldseek server. The default databases, including AlphaFold / proteome, AlphaFold / Swiss-Prot, AlphaFold / UniProt50, CATH50, GMGCL, MGnify-ESM30 and PDB100, were selected to maximize the search scope. The search mode was configured to 3Di/AA for optimal identification of homologous structures and sequences.The five protein targets with the lowest e-value scores were selected for each search result as the most statistically significant hits. To ensure that the identified proteins belonged to the Gld family, their functional annotations and detailed descriptions were cross-verified using the UniProt database(https://www.uniprot.org). The corresponding protein sequences were then downloaded from GenBank(https://www.ncbi.nlm.nih.gov/genbank/). Due to the high degree of similarity between homologous proteins, the search results often included duplicate entries. To maintain dataset diversity and uniqueness, duplicate sequences were excluded, and additional targets were selected in descending order of e-value until five unique protein sequences were obtained for each GldL query.Similar to the methodology used for MotB protein selection, the corresponding GldM proteins, which are typically encoded immediately downstream of GldL within the genome, were identified. GenBank's genome data retrieved the GldM sequences corresponding to the GldL proteins. This process led to constructing a dataset comprising 30 unique GldLM protein sequences.

Finally, based on a combination of metadata, including species taxonomy, genome sources, and detailed protein labeling annotations, protein files were systematically compiled and organized. The final dataset encompassed a total of 499 target proteins, consisting of 379 MotAB proteins, 90 ZorAB proteins, and 30 GldLM proteins. All protein sequences were saved as amino acid sequence collections in FASTA format, ensuring compatibility for downstream computational and experimental analyses.

3.2 Acquisition of protein structure data

To obtain structural data for these motor proteins and their homologs, I utilized R Studio(4.3.1) to combine various bioinformatics packages for processing and converting information from the FASTA files. The ape package, the ape::read.FASTA function was used to read the specified FASTA file, extract amino acid sequences, and count the number of sequences. Since incomplete or misannotated sequences can lead to anomalies in time estimation for phylogenetic trees, I used an automated approach to process each sequence individually within a for-loop framework to prevent errors caused by damaged data.

Within the loop, a custom R function, *fasta_to_alphafold_cifs*, was employed to convert amino acid sequences into structural data. This was achieved by using R to call functions from ChimeraX 1.7.1 (Meng et al., 2023) that BLAST the input sequence against a database of stored AlphaFold models, and retrieve the pre-calculated predicted structure of the closest match (usually a 100% sequence match) The function also downloads the resulting CIF files and extracted key information, including sequence similarity, coverage, and confidence scores. During this process, certain unrecognisable sequences or others issues (e.g., format inconsistencies or failure in model prediction) would terminate the loop prematurely. To address this, we manually logged the sequence numbers, verified their completeness and correctness, and removed defective sequences during the subsequent iterations of the loop.

The AlphaFold models was visualised locally using molecular visualization software ChimeraX and the AlphaFold2 framework(Jumper et al., 2021), which generated detailed HTML files containing comprehensive protein structure prediction information. After processing, all HTML files were organized and stored in a single dataframe. This dataframe included predicted information for each sequence, file paths, and other relevant metadata. The data was exported in txt format for manual inspection and subsequent re-import into the code.

A custom function, *alphafold_htmls_to_df*, was used to parse the AlphaFold-generated HTML files. Additionally, another function, *structures_to_3Dis*, was used to convert CIF files generated by AlphaFold into the 3Di format. After completing the conversion, the code created an information table documenting metadata for each CIF file and its corresponding 3Di file.

Finally, all the processed data were integrated into a comprehensive dataframe containing sequence identifiers, AlphaFold prediction results, 3Di file details, and other auxiliary data. The data were standardized to ensure consistency in column names and formats. The final output was saved as a txt file to facilitate further analysis. This pipeline not only streamlined the processing of structural data but also ensured the integrity and usability of the resulting datasets for downstream applications.

3.3 Optimal sequence alignment

During the data processing, FASTA sequence files obtained for various proteins, including MotA, MotC, PomA, ZorA, and GldL, were subjected to multiple alignment and refinement steps using a combination of advanced tools. The detailed procedure is as follows:

3.3.1 Alignment of amino acid sequences with USalign

Amino acid sequences were aligned using Universal Structural Alignment (USalign; Zhang et al., 2022) in the macOS Terminal environment to identify sequences most suitable for constructing phylogenetic trees. The alignment was performed with the parameter -mm 4, focusing on each protein chain individually. After obtaining alignment results, a custom R function, *align_3Di*, was utilized to map the alignment results from amino acid sequences to their corresponding 3Di sequences. This step ensured a seamless integration of sequence and structural data.

3.3.2 Comparison using FAMSA3Di

Different alignment methods often yield significantly varying results. To handle sequences composed solely of 3Di characters, we utilized FAMSA3Di, a modified version of Fast and Accurate Multiple Sequence Alignment(FAMSA; Deorowicz et al., 2016) recompiled with Foldseek's substitution cost matrix for the 3Di alphabet. This approach addresses the significant variations often observed between different alignment methods. Subsequently, we employed a custom R function, *align_3Dis_to_AAs*, to map the alignment results from 3Di sequences back to their corresponding amino acid (AA) sequences. This mapping provides a comparative perspective between alignment methods, enhancing our understanding of their relative performances.

3.3.3 Compression and alignment of protein structure data with FoldMason

The CIF structure files for all A and B subunits were tar-compressed into a single archive. This tar.gz file, containing all protein structure information, was processed using FoldMason (Gilchrist et al., 2024). FoldMason performed easy-multiple-sequence alignment (easy-msa) on the structural data, generating aligned sequences in both AAs-only and 3Di-only formats. This approach streamlined the alignment of complex structural information.

3.3.4 Refinement with MAFFT

Following the initial alignment, further refinement was carried out due to the presence of structural elements such as alpha-helices and beta-sheets, which resulted in extended 3Di strings. These strings were individually aligned using MAFFT (Katoh et al., 2002). This operation was performed on macOS Terminal by invoking MAFFT through Ruby scripts. The parameters were configured to retain amino acids in positions 1–249 and 311–476, while aligning the region spanning positions 250–310, which consisted of a 61-character string. The alignment in this region was optimized using the –maxiterate 1000 parameter to achieve a highly refined result.

Comparing alignments between USalign, FAMSA3Di, and FoldMasonwe identify FAMSA as the best alignment scheme. Then with MAFFT refinement, helped produce a plausible alignment incorporating high-quality sequence and structural alignments. Visual inspection of the AA and corresponding 3Di alignment side-by-side suggests that they provide a reasonable basis for constructing accurate phylogenetic trees and conducting subsequent evolutionary and functional analyses.

3.4 Trim

Certain proteins in the data set contain long, nonhomologous regions. For example, ZorA proteins possess an extended tail structure, and GldL proteins similarly feature elongated tails. These tail structures are not part of the motor complex components. To address this, the sequences of ZorA and GldL were individually aligned, and the elongated tail regions were manually trimmed. This ensured that, during global sequence alignment, the presence of obvious non-homologous segments was minimized, improving the overall alignment quality.

To further enhance the quality of the multiple sequence alignment, trimAl 1.4 (Capella-Gutiérrez et al., 2009) was utilized to refine the alignment data. Both AA-only and 3Di-only alignment datasets were processed. An initial gap threshold of 35% was applied, removing columns in the alignment where the proportion of non-gap residues was below 35%. This approach effectively eliminated low-confidence regions and areas with low information content, optimizing the alignment for downstream analyses.

To evaluate the impact of varying trimming thresholds on subsequent analyses, additional thresholds of 10% and 5% were applied to the original alignment data. This iterative trimming process was performed to generate multiple alignment datasets with differing levels of stringency. The trimmed alignments were used for phylogenetic analysis and alignment quality assessment, with a particular focus on examining the retention of critical structural regions under varying thresholds.

This methodical approach ensured that non-homologous regions were excluded, while maximizing the retention of meaningful and high-confidence alignment columns, ultimately improving the reliability of the phylogenetic and structural analyses.

3.5 Phylogenetic analysis

The alignment and phylogenetic analysis were conducted based on two types of datasets: amino acid sequence datasets without structural information (AAs-only) and structural sequence datasets predicted using AlphaFold (3Di-only). Using these data types, three distinct datasets were constructed:

- **AAs-only Dataset:** Containing only amino acid sequences.
- **3Di-only Dataset:** Comprising only structural sequence information.
- **AAs+3Di Dataset:** A combination of amino acid sequences and structural sequences,add the 3Di sequence after the last amino acid of the AA sequence.

Each dataset was used to construct maximum likelihood (ML) phylogenetic trees using the IQ-TREE (Minh et al., 2020). The entire workflow was executed on the integrated server New Zealand eScience Infrastructure(NESI;https://www.nesi.org.nz.

3.5.1 Model selection

IQ-TREE provides the capability to efficiently evaluate various substitution models and automatically select the model that best fits the sequence dataset. The custom models used for this study included the following:

- **Amino Acid Substitution Models:** Blosum62, Dayhoff, DCMut, JTT, JTTDCMut, LG, Poisson, PMB, and WAG.
- **3Di-Specific Model:** A substitution rate matrix derived from Foldseek 3Di substitution cost matrices (Puente-Lelievre, Malik, et al., 2024).

For model selection, the -m MFP+MERGE option was employed to invoke Model Finder Plus(MFP;Kalyaanamoo et al., 2017), which automatically merges similar models based on statistical criteria. Base frequency options were set using -mfreq FU,F, considering both uniform amino acid frequency and empirical frequency distributions to adapt to the characteristics of the alignment data. Rate heterogeneity among sites was modeled using -mrate E,G,R, corresponding to:

- **Equal Rates:** Uniform substitution rates across sites.
- **Gamma Rates:** Gamma-distributed rate heterogeneity.
- **Free Rate Model:** Site-specific substitution rate flexibility.

MFP scores models based on their statistical fit, considering amino acid frequency weighting and gamma-distributed substitution rates with four rate categories.

3.5.2 Partition files

To account for differing evolutionary rates and patterns between amino acid and 3Di features, partitioning was applied (Chernomor et al., 2016). The datasets were divided as follows:

- **part1_SubunitA_AA:** Amino acid portion of subunit A.
- **part2_SubunitB_AA:** Amino acid portion of subunit B.
- **part3_SubunitA_3Di:** 3Di sequence portion of subunit A.
- **part4_SubunitB_3Di:** 3Di sequence portion of subunit B.

Partition analysis was performed to allow shared tree topology and relative branch lengths across partitions, while accommodating differences in substitution rates and patterns between the partitions.

3.5.3 Phylogenetic inference - Maximum Likelihood

For each dataset, the following IQ-TREE parameters were employed:

- **Bootstrapping:** 1,000 ultrafast bootstrap (UFBoot;Hoang et al., 2018) replicates were used to assess branch support.

- **Approximate Likelihood Ratio Test (aLRT):** Conducted 1,000 replicates to compute branch statistical support.
- **Tree Optimization:** Enabled branch topology optimization using nearest-neighbor interchange (NNI) with the -bnni flag.

The full command used for running IQ-TREE is as follows:

```
iqtree -s input.fasta -spp BOTHp.raxml -m MFP+MERGE -madd 3Di
-mdef 3Di.nexus -mfreq FU,F -mrate E,G,R --ufboot 1000 -alrt 1000
-bnni --redo | tee output.txt &
```

This systematic approach allowed for robust phylogenetic tree construction, ensuring that the results accounted for both sequence and structural variability while optimizing computational efficiency.

3.6 Phylogenetic inference - Bayesian

Using BEAST 2.7.4 (R. Bouckaert et al., 2019) to analyze the datasets, employing distinct models to account for the unique characteristics of amino acid (AAs) and structural sequence (3Di) datasets. The OBAMA 1.1.1 (R. R. Bouckaert, 2020) Bayesian Aminoacid Model Averaging model was applied to the AAs dataset, while the Fold Beast 3Di Model Averaging model was utilized for the 3Di dataset. Both datasets were analyzed under the Optimized Relaxed Clock molecular clock framework (Drummond et al., 2006), which accommodates variation in evolutionary rates across branches while ensuring an optimal fit to the data. The Yule skyline tree prior was chosen to model speciation processes, and the Markov chain Monte Carlo (MCMC) analysis was run for 1 billion iterations to sample the posterior distribution comprehensively (Drummond et al., 2005; Drummond et al., 2002). The Maximum Clade Credibility (MCC) tree was extracted from this posterior distribution, representing the most likely tree topology, with posterior probability values computed to quantify statistical support for each node. Entropy was measured to evaluate the diversity of phylogenetic information within the posterior distribution, enabling a detailed comparison of phylogenetic consistency and differences between the AAs and 3Di datasets. Convergence of the MCMC chains was assessed using Tracer 1.7.2 (Rambaut et al., 2018), ensuring all effective sample size (ESS) values exceeded 200, while the evolutionary rate distribution and its confidence intervals, as estimated by the Optimized Relaxed Clock model, were validated to confirm model robustness and reliability.

3.7 Phylogenetic analyse

Figtree 1.4.4(http://tree.bio.ed.ac.uk/software/figtree/) was utilized for detailed visualization of the phylogenetic trees generated by IQ-TREE. A custom annotation file was incorporated to represent different branches and sequence types using colour coding, facilitating the intuitive display of classification features and enhancing the interpretability and clarity of the phylogenetic data.

For the phylogenetic tree samples generated by BEAST2, TreeDensitree 2.7.7 (R. R. Bouckaert & Heled, 2014) was initially employed to evaluate the tree files, providing an assessment of posterior consistency by analyzing overall density distributions and topological variation within the tree samples. Subsequently, the Tree Annotator (R. Bouckaert et al., 2019) was applied to summarize the tree samples and generate a summary tree (MCC tree). During this process, default parameter configurations were used, and the low memory mode was activated to optimize computational resource utilization. Tree Annotator integrated information from the posterior distribution, calculated statistical support values for each node, and constructed a representative summary tree that reflected the most probable topology derived from the posterior distribution.

4 Result

4.1 Phylogenetic relationships

We performed a phylogenetic analysis using IQ-TREE based on the amino acid sequences of 470 related proteins, including MotAB, ZorAB, and GldLM, and the 3Di sequence. The data were divided into four partitions: part1_SubunitA_AA, part2_SubunitB_AA, part3_SubunitA_3Di and part4_SubunitB_3Di, corresponding to the amino acid sequences and structural features of the different subunits. Model selection was based on the Bayesian information criterion (BIC), and the best alternative model was determined for each partition: LG+F+R10 (LogL = –179962.370, AIC = 360000.740, AICc = 360009.043, BIC = 360152.034) was selected for part1_SubunitA_AA, part2_SubunitB_AA selected WAG+F+R9 (LogL = –145428.827, BIC = 291080.061), part1_SubunitA_3Di and part1_SubunitB_3Di both selected 3Di+F+R8 (LogL = –73313.569 and –72774.812, BIC = 146830.505 and 145759.674). All models had the lowest BIC value in each partition, and the w-BIC weights were close to 100%, indicating the best-fitting effect.

Based on the above model, the partitioned maximum likelihood inference was performed on the combined amino acid and three-dimensional structure encoding (AA+3Di) data, and the obtained phylogenetic tree showed high resolution and stable topology. Most of the main branches and internal nodes have high support (UFBoot 95%, SH-aLRT 80%), while some key nodes do not have high support. In the position where type 3 Zorya and type 1 Zorya are classified as sister branches, both UFBoot and SH-aLRT values are below 80, indicating that this branch relationship is not statistically robust. In addition, the support for some internal subdivisions is relatively low, which may reflect insufficient sequence or structural information in local regions. In contrast, at the node where type 1 Zorya diverged internally to type 2 Zorya, UFBoot and SH-aLRT were both 100, indicating that this branch has strong statistical support, further supporting the stability of type 2 Zorya as an independent evolutionary unit.

The system tree clearly divides the four major clades into non-flagellar proteins, flagellar proteins, Zorya proteins, and gliding proteins, which is consistent with the functional annotations during database searches, indicating that the results of phylogenetic inference have good biological correspondence. The non-flagellar protein group includes CCD2 and CCD3, which are concentrated cytoplasmic domains composed of two or three short helices, respectively. The flagellar protein group includes TGI4 and TGI5, which contain torque-generating interface structures with four or five short helices, respectively, and were first proposed in studies by Puente-Lelievre, Ridone, et al. (2024). The Zorya protein group appears as a sister branch of the flagellar protein group, which includes three subtypes: type 1, type 2 and type 3. Type 3 and type 1 form a sister branch, and type 2 diverged from type 1. The Gliding protein group, as the earliest differentiated lineage, is significantly separated from other groups in the phylogenetic tree, showing the uniqueness of its evolutionary path.

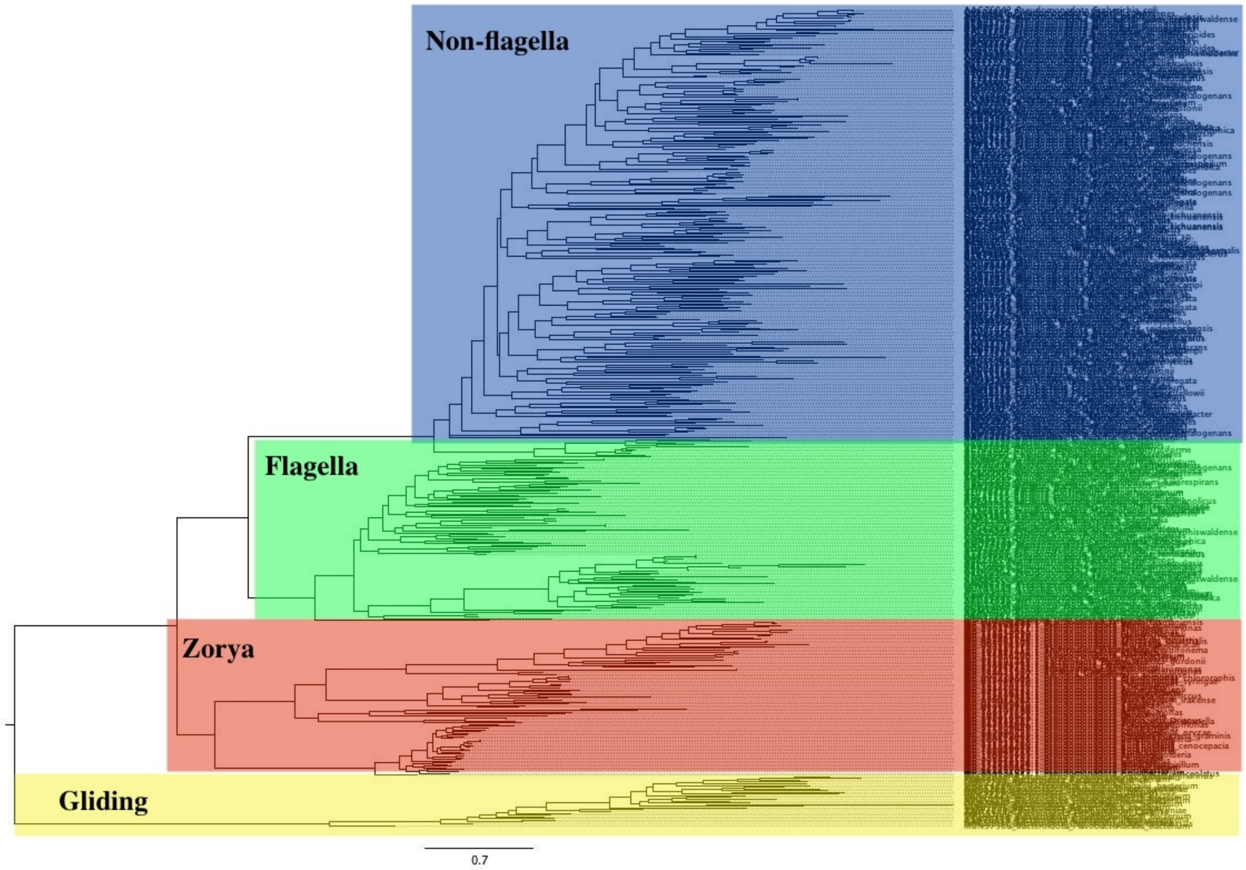

Figure 4: Phylogenetic tree constructed using IQtree

To reduce computational cost and time, a subset of 320 proteins, including MotAB, ZorAB, and GldLM, was analyzed using BEAST. Phylogenetic inference was performed under an uncorrelated lognormal relaxed clock model and the Yule speciation process. The maximum MCMC chain length was set to 100 million generations, with sampling every 1,000 generations and a burn-in of 10%. Convergence was assessed using Tracer based on effective sample size (ESS) values and likelihood stability.

Among the three datasets (AAs-only, 3Di-only, and AA+3Di), only the AAs-only run achieved sufficient convergence, with most ESS values exceeding 200. This run terminated at generation 63,915,000, with a posterior of –239555.05, a prior of –239162.65, and a posterior–prior difference of –392.40. The average runtime per million samples was approximately 7 hours and 53 minutes. In contrast, the 3Di-only run stopped at generation 60,760,000 (posterior = –115678.88, prior = –115485.73, difference = –193.15), and the AA+3Di run terminated early at generation 29,203,000 (posterior = –358264.43, prior = –358171.69, difference = –92.74). Both runs showed low ESS values (typically ¡ 50), indicating poor convergence and preventing reliable posterior probability estimation. The average runtime per million samples was 8 hours and 17 minutes for the 3Di-only dataset and 17 hours and 15 minutes for the AA+3Di dataset. Due to computational limitations, these analyses could not be completed to full convergence.

The BEAST tree inferred from the AAs-only dataset produced a topology consistent with the IQ-TREE results, recovering four distinct clades: non-flagellar proteins, flagellar proteins, Zorya proteins, and gliding proteins. In this tree, flagellar and Zorya proteins formed a sister group with a posterior probability of 53.28%, while gliding and non-flagellar proteins formed another sister group with a posterior probability of 57.8%. Within the Zorya clade, type 3 and type 1 Zorya proteins clustered together as sister branches, with type 1 Zorya proteins exhibiting high internal diversity (average posterior probability ¿ 95%). Consistent with the IQ-TREE topology, type 2 Zorya proteins emerged as a distinct lineage branching from within the type 1 group.

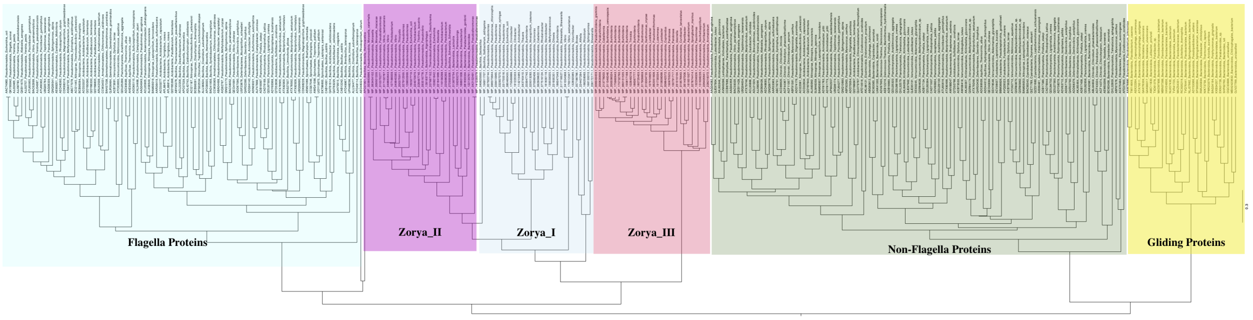

Figure 5: Phylogenetic tree constructed using BEAST2.

In the 3Di-only summary tree, flagellar and non-flagellar proteins formed a sister group, while Zorya and gliding proteins formed another. Within the Zorya clade, type 3 and type 1 Zorya proteins again clustered together, but type 2 Zorya proteins appeared as a separate lineage distinct from both. A similar pattern was observed in the AA+3Di summary tree, in which gliding proteins formed the outermost clade, flagellar and non-flagellar proteins grouped together, and Zorya proteins appeared as an independent clade. Within the Zorya group, type 3 and type 1 Zorya proteins formed a sister group, while type 2 Zorya proteins remained as a distinct lineage. Due to low ESS values, posterior probabilities are not reported for the 3Di-only and AA+3Di trees.

4.2 High conservation of zorB with Flagellar Proteins

Sequence alignment results indicate that the B subunits of Zorya proteins exhibit high sequence conservation with the B subunits of the flagellar motor system. Specifically, multiple sequence alignments of the B subunit regions revealed that, in addition to the conserved regions shared by flagellar and non-flagellar B subunits, the B subunits of Zorya proteins also share an Outer Membrane Protein A (OmpA) domain with flagellar B subunits (e.g., MotB). This OmpA domain is particularly highly conserved in 3Di sequence alignments. In non-flagellar proteins, the OmpA domain is replaced by another structural region. In GldM proteins, due to their extremely low conservation, a significant portion of sequence fragments was removed during alignment. Even in the limited regions that could be aligned, the similarity between GldM and B subunits such as MotB and ZorB was too low to provide strong evidence that gliding proteins are homologous to the AQB family or Zorya proteins.

4.3 3Di sequences exhibit higher alignment consistency than AA sequences

To assess the comparative alignment consistency between AA and 3Di sequences, we examined the AA+3Di concatenated alignment using AliView, in which AA and 3Di sequences were placed on the left and right halves of each sequence, respectively. The "highlight consensus characters" function in AliView was used to visualize conserved positions(Figure 6).

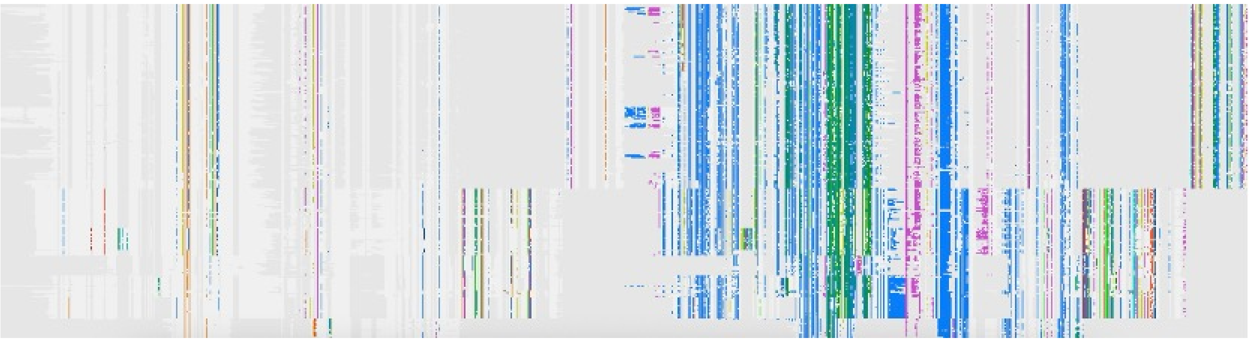

Figure 6: Alignment of AAs sequence(on left) and 3Di sequence(on right)

Quantitative analysis revealed that only 42 columns in the AA region showed complete consensus across all 470 sequences, 411 consensus columns, representing 46.8% of its alignment length. These results clearly demonstrate that 3Di sequences are considerably more conserved than primary amino acid sequences across homologous proteins, likely reflecting structural constraints that remain stable despite substantial sequence divergence.

5 Discussion

This study demonstrates that Zorya proteins and flagellar proteins are sister clades, with internal differentiation of Zorya into Type 1/2/3 sister clades. The conserved OmpA domain shared by

ZoryaB and MotB indicates possible similarities in their mechanochemical mechanisms, while the paraphyletic origin of Type 2 Zorya reflects undefined functional specialization. These results clarify the potential coordinated evolutionary relationship between bacterial motility and defense systems.

5.1 Structural and genomic features of the zorya system

At the sequence level, type III Zorya proteins exhibit a clear N-terminal deletion in the A subunit compared to type I and II. Additionally, the C-terminal region of the B subunit in type III shows distinct structural features in its 3Di sequence relative to the other two types. As the most recently identified Zorya type, the type III system contains ZorF and ZorG upstream and downstream of the core genes, whereas type I and II systems contain only ZorCD and ZorE downstream(Doron et al., 2018; Payne et al., 2021). These differences in genomic context and structural composition may together account for the pronounced divergence observed in the sequence and structure of type III Zorya proteins.

Flagellar protein MotB and Zorya protein ZorB exhibit significant homology through their shared OmpA domain. In contrast, members of the ExbBD/TolQR family lack this domain but still retain the core proton-conducting motif, such as the conserved aspartate residue, suggesting that these systems may have originated from a common ancestral mechanochemical coupling complex. Structural comparisons between the Zorya and flagellar systems further demonstrate that the proton-conducting core of ZorB is highly conserved with that of MotB, including the essential aspartate residue and surrounding Ser/Thr ring. Mariano et al. (2025) pointed that type I Zorya may be Na-driven, whereas type II resembles H-driven systems, indicating potential diversification in ion selectivity. Moreover, unlike the flexible peptidoglycan-binding (PG) domain observed in MotB, the C-terminal domain of ZorB forms a rigid dimer with a conserved interface, suggesting a structurally stabilized anchoring role unique to the Zorya lineage. As the most recently identified member, the structural features of type III Zorya proteins remain to be further validated through experimental analysis.

5.2 OmpA-Like domains in non-flagellar and flagella system

Several studies have suggested that the structure of MotB primarily consists of a TolR-like N-terminal region and a conserved OmpA domain(Boags et al., 2019). The C-terminal sequence conservation observed between OmpA-related outer membrane proteins and MotB implies a shared functional role in both Gram-positive and Gram-negative bacteria, possibly involving interactions between these domains and the peptidoglycan layer.

In this study, we initially hypothesized that B subunits in non-flagellar systems, such as ExbD and TolR, lack OmpA-related domains. However, structural comparisons between the peptidoglycan-binding (PGB) domain of TolR and the OmpA domain of MotB revealed notable structural similarities despite significant sequence divergence.

Further analysis of our alignment data showed that the C-terminal regions of non-flagellar proteins display a degree of sequence similarity to the OmpA domain located at the C-terminus of flagellar proteins such as MotB. Although these regions are unlikely to be annotated as canonical OmpA domains by domain prediction tools or manual inspection, their potential structural resemblance suggests that they may fulfill analogous functions. These findings imply that OmpA-like domains may have undergone structural convergence across different systems, or alternatively, may have retained partial homology during evolution—reflecting a conserved role in mechanochemical coupling.

5.3 why place gliding protein as root

Gliding proteins are components of the Type IX Secretion System that is found primarily in the phylum Bacteroidetes(Hennell James et al., 2022). This system drives cells to slide on solid surfaces, enabling movement without flagella or pili. This mechanism of movement is completely different from the typical flagellar rotation mechanism and depends on the transfer of energy (such as proton motive force) from the intracellular membrane to the extracellular mechanical movement. In our phylogenetic analysis, gliding proteins consistently form a separate clade that is significantly different from other sequences and is the most distant phylogenetically. This difference may be due to its independent evolutionary origin and functional specialization. The GldLM complex does not contain the OmpA domain commonly found in MotB or ZorB, and its overall structural layout is also highly specific, resulting in a large number of regions being truncated due to poor conservation during alignment.

Since the gliding phylum always appears as an external branch in the rootless tree, we have deliberately set it as the root node of the phylogenetic tree in this study and used it as a working hypothesis for subsequent topological interpretation. Although this hypothesis needs to be further verified, judging from the uniqueness of its function and system structure, this phylum may represent an early-appearing non-flagellar power mechanism, and its position as the root has a certain biological rationality.

5.4 challenges in BEAST analyses

In this study, certain analyses did not yield ideal results due to limitations in computational resources. For example, in the BEAST analysis, even after allocating 10 CPU cores on New Zealand's national high-performance computing platform (NeSI) and reducing the AA+3Di dataset by 30%, the job still ran for over 504 hours (approximately 21 days), completing only around 36% of the total MCMC iterations. Despite parameter tuning and data size reduction, the BEAST runs failed to reach convergence.

This outcome highlights the computational complexity involved in modeling structurally encoded data such as 3Di, and reveals limitations in the scalability of current MCMC algorithms when applied to large-scale phylogenetic reconstruction involving hundreds of sequences, multiple partitions, and complex substitution models. Furthermore, it underscores both the potential and the challenge of incorporating 3Di: while they provide higher alignment conservation and phylogenetic resolution, they also significantly increase computational load.

Future analyses may benefit from alternative strategies to improve efficiency and convergence. These could include the use of more efficient sampling algorithms (such as adaptive operators in BEAST2), simplified model settings, or the use of replicated IQ-TREE runs to obtain a stable tree topology, which can then be fixed during time-calibrated Bayesian inference to reduce model search space.

5.5 rooting strategies-Rootstrap and Root-Test

The phylogenetic trees generated by IQ-TREE are unrooted, meaning they represent the relative relationships among sequences without specifying the direction of evolution or the position of the common ancestor. For instance, in our analysis, flagellar proteins and Zorya proteins consistently appeared as sister groups, but their order of emergence and shared ancestral lineage remain ambiguous. Such unrooted topologies limit the ability to perform ancestral state reconstruction or to infer evolutionary trajectories with confidence.

To address this limitation, we aim to incorporate recent IQ-TREE features—namely, Rootstrap and the root-test function—in future analyses. Rootstrap provides statistical support for potential root positions by evaluating the proportion of bootstrap replicates in which each branch appears as the root, thereby identifying the most likely evolutionary origin. The root-test allows for formal comparison of alternative rooting scenarios based on likelihood scores, enabling hypothesis-driven inference of root positions.

These tools and functions will help test our current working hypothesis, in which gliding proteins are positioned as the root of the tree, and offer a more robust statistical framework for future investigations into the evolutionary history of proton-conducting motor systems.

5.6 Application and prospects of 3Di in molecular phylogenetics

As a phylogenetic tool still in its infancy, the introduction of 3Di features provides a new structural dimension to traditional sequence alignment. It is encoded based on the spatial relationship between residues and can provide additional conservation information in lineages where sequence similarity has decayed severely. In this study, we found that using 3Di features together with amino acid sequences for phylogenetic reconstruction not only enhances the topological consistency of the tree, but also improves the resolution of some deep branches, especially at nodes where traditional sequence methods are unclear.

Despite its promising future, the 3Di method still faces a number of technical and methodological challenges. The reliability of the results is highly dependent on the accuracy of the structure prediction tools (such as AlphaFold). In particular, 3Di data may be highly noisy when dealing with membrane proteins, disordered regions, or low-confidence structures. In addition, since 3Di features are derived from the global topological relationships of 3D structures, it remains to be further explored whether they satisfy the site independence assumption commonly used in phylogenetic modeling. Ignoring potential dependencies may introduce bias in support assessment. On the other hand, 3Di features currently focus mainly on the spatial layout between the main chains, and have limited response to side chain conformation changes or skeletal geometric details.

In the future, further standardization of 3Di data, optimization of matrix models, and differentiated treatment of feature states in different types of structural regions (such as -helices and -folds) may significantly expand their scope of application. Moreover, with the possibility of integrating Bayesian phylogenetic methods with structural feature data, 3Di is also expected to be used for time-scale inference, complex model integration, and comprehensive modeling of phylogenetic uncertainty. We anticipate that structural features will become key variables for breaking through bottlenecks in deep phylogeny and drive protein phylogeny to a new stage with higher resolution and a wider range of data sources as related tools and standards gradually mature.

6 Conclusion

This study demonstrates the potential of combining amino acid sequences with 3Di sequences for phylogenetic reconstruction of proton-motor complexes such as MotAB, ZorAB and gliding proteins. We constructed a phylogeny with high support by incorporating AlphaFold-based structure prediction and 3Di sequences in a model-driven framework, which successfully divided the tree into four major groups, especially excelling in the resolution of deep branches. Zorya proteins, which are homologous to flagellar proteins, especially type III Zorya proteins, exhibit uniqueness within the Zorya protein family. Although the computational pressure of BEAST analysis and the accuracy of structure prediction still pose challenges, our results support the feasibility and research value of incorporating structure-derived traits into phylogenetic inference. Future work should focus on the further optimization of the 3Di alternative model, the application and verification of the rooting method, and its application in a wider range of phylogenetic problems and more protein families of interest (e.g., lateral hair motor protein LafTU, sliding motility protein AlgRS, etc.).

References

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2002). The Shape and Structure of Proteins. In *Molecular Biology of the Cell. 4th edition*. Garland Science. Retrieved March 19, 2025, from https://www.ncbi.nlm.nih.gov/books/NBK26830/

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–3402. https://doi.org/10.1093/nar/25.17.3389

Baldauf, S. L. (2003). The deep roots of eukaryotes. *Science (New York, N.Y.)*, *300*(5626), 1703–1706. https://doi.org/10.1126/science.1085544

Berg, H. C. (2003). The Rotary Motor of Bacterial Flagella [Publisher: Annual Reviews]. *Annual Review of Biochemistry*, *72*(Volume 72, 2003), 19–54. https://doi.org/10.1146/annurev.biochem.72.121801.161737

Boags, A. T., Samsudin, F., & Khalid, S. (2019). Binding from Both Sides: TolR and Full-Length OmpA Bind and Maintain the Local Structure of the E. coli Cell Wall [Publisher: Elsevier]. *Structure*, *27*(4), 713–724.e2. https://doi.org/10.1016/j.str.2019.01.001

Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., Heled, J., Jones, G., Kühnert, D., Maio, N. D., Matschiner, M., Mendes, F. K., Müller, N. F., Ogilvie, H. A., Plessis, L. d., Popinga, A., Rambaut, A., Rasmussen, D., Siveroni, I., … Drummond, A. J. (2019). BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis [Publisher: Public Library of Science]. *PLOS Computational Biology*, *15*(4), e1006650. https://doi.org/10.1371/journal.pcbi.1006650

Bouckaert, R. R. (2020). OBAMA: OBAMA for Bayesian amino-acid model averaging. *PeerJ*, *8*, e9460. https://doi.org/10.7717/peerj.9460

Bouckaert, R. R., & Heled, J. (2014, December). DensiTree 2: Seeing Trees Through the Forest [Pages: 012401 Section: New Results]. https://doi.org/10.1101/012401

Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldón, T. (2009). trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, *25*(15), 1972–1973. https://doi.org/10.1093/bioinformatics/btp348

Carroll, B. L., Nishikino, T., Guo, W., Zhu, S., Kojima, S., Homma, M., & Liu, J. (2020). The flagellar motor of Vibrio alginolyticus undergoes major structural remodeling during rotational switching (E. H. Egelman, J. Kuriyan, E. H. Egelman, C. Aizawa, & S. M. Lea, Eds.) [Publisher: eLife Sciences Publications, Ltd]. *eLife*, *9*, e61446. https://doi.org/10.7554/eLife.61446

Celia, H., Botos, I., Ni, X., Fox, T., De Val, N., Lloubes, R., Jiang, J., & Buchanan, S. K. (2019). Cryo-EM structure of the bacterial Ton motor subcomplex ExbB–ExbD provides information on structure and stoichiometry [Publisher: Nature Publishing Group]. *Communications Biology*, *2*(1), 1–6. https://doi.org/10.1038/s42003-019-0604-2

Chaban, B., Hughes, H. V., & Beeby, M. (2015). The flagellum in bacterial pathogens: For motility and a whole lot more. *Seminars in Cell & Developmental Biology*, *46*, 91–103. https://doi.org/10.1016/j.semcdb.2015.10.032

Chernomor, O., von Haeseler, A., & Minh, B. Q. (2016). Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices. *Systematic Biology*, *65*(6), 997–1008. https://doi.org/10.1093/sysbio/syw037

Chowdhury, B., & Garai, G. (2017). A review on multiple sequence alignment from the perspective of genetic algorithm. *Genomics*, *109*(5), 419–431. https://doi.org/10.1016/j.ygeno.2017.06.007

Chung, S. Y., & Subbiah, S. (1996). A structural explanation for the twilight zone of protein sequence homology. *Structure (London, England: 1993)*, *4*(10), 1123–1127. https://doi.org/10.1016/s0969-2126(96)00119-0

Deme, J. C., Johnson, S., Vickery, O., Aron, A., Monkhouse, H., Griffiths, T., James, R. H., Berks, B. C., Coulton, J. W., Stansfeld, P. J., & Lea, S. M. (2020). Structures of the stator complex that drives rotation of the bacterial flagellum [Number: 12 Publisher: Nature Publishing

Group]. *Nature Microbiology*, *5*(12), 1553–1564. https://doi.org/10.1038/s41564-020-0788-8

Deorowicz, S., Debudaj-Grabysz, A., & Gudyś, A. (2016). FAMSA: Fast and accurate multiple sequence alignment of huge protein families [Publisher: Nature Publishing Group]. *Scientific Reports*, *6*(1), 33964. https://doi.org/10.1038/srep33964

Doolittle, R. F. (1986). *Of Urfs And Orfs: A Primer on How to Analyze Derived Amino Acid Sequences*. University Science Books.

Doron, S., Melamed, S., Ofir, G., Leavitt, A., Lopatina, A., Keren, M., Amitai, G., & Sorek, R. (2018). Systematic discovery of antiphage defense systems in the microbial pangenome [Publisher: American Association for the Advancement of Science]. *Science*, *359*(6379), eaar4120. https://doi.org/10.1126/science.aar4120

Drummond, A. J., Rambaut, A., Shapiro, B., & Pybus, O. G. (2005). Bayesian Coalescent Inference of Past Population Dynamics from Molecular Sequences. *Molecular Biology and Evolution*, *22*(5), 1185–1192. https://doi.org/10.1093/molbev/msi103

Drummond, A. J., Ho, S. Y. W., Phillips, M. J., & Rambaut, A. (2006). Relaxed Phylogenetics and Dating with Confidence [Publisher: Public Library of Science]. *PLOS Biology*, *4*(5), e88. https://doi.org/10.1371/journal.pbio.0040088

Drummond, A. J., Nicholls, G. K., Rodrigo, A. G., & Solomon, W. (2002). Estimating Mutation Parameters, Population History and Genealogy Simultaneously From Temporally Spaced Sequence Data. *Genetics*, *161*(3), 1307–1320. https://doi.org/10.1093/genetics/161.3.1307

Dyer, S. C., Austine-Orimoloye, O., Azov, A. G., Barba, M., Barnes, I., Barrera-Enriquez, V. P., Becker, A., Bennett, R., Beracochea, M., Berry, A., Bhai, J., Bhurji, S. K., Boddu, S., Branco Lins, P. R., Brooks, L., Ramaraju, S. B., Campbell, L. I., Martinez, M. C., Charkhchi, M., … Yates, A. D. (2025). Ensembl 2025. *Nucleic Acids Research*, *53*(D1), D948–D957. https://doi.org/10.1093/nar/gkae1071

Finn, R. D., Clements, J., Arndt, W., Miller, B. L., Wheeler, T. J., Schreiber, F., Bateman, A., & Eddy, S. R. (2015). HMMER web server: 2015 update. *Nucleic Acids Research*, *43*(W1), W30–W38. https://doi.org/10.1093/nar/gkv397

Fitch, W. M., & Margoliash, E. (1967). Construction of Phylogenetic Trees [Publisher: American Association for the Advancement of Science]. *Science*, *155*(3760), 279–284. https://doi.org/10.1126/science.155.3760.279

Gilchrist, C. L. M., Mirdita, M., & Steinegger, M. (2024, August). Multiple Protein Structure Alignment at Scale with FoldMason [Pages: 2024.08.01.606130 Section: New Results]. https://doi.org/10.1101/2024.08.01.606130

Gregory, T. R. (2008). Understanding Evolutionary Trees [Number: 2 Publisher: BioMed Central]. *Evolution: Education and Outreach*, *1*(2), 121–137. https://doi.org/10.1007/s12052-008-0035-x

Haiko, J., & Westerlund-Wikström, B. (2013). The Role of the Bacterial Flagellum in Adhesion and Virulence. *Biology*, *2*(4), 1242–1267. https://doi.org/10.3390/biology2041242

Hennell James, R., Deme, J. C., Kjr, A., Alcock, F., Silale, A., Lauber, F., Johnson, S., Berks, B. C., & Lea, S. M. (2021). Structure and mechanism of the proton-driven motor that powers type 9 secretion and gliding motility. *Nature Microbiology*, *6*(2), 221–233. https://doi.org/10.1038/s41564-020-00823-6

Hennell James, R., ·, &, (2022). Structures of the Type IX Secretion/Gliding Motility Motor from across the Phylum Bacteroidetes [Publisher: American Society for Microbiology]. *mBio*, *13*(3), e00267–22. https://doi.org/10.1128/mbio.00267-22

Higgins, D. G., Thompson, J. D., & Gibson, T. J. (1996). Using CLUSTAL for multiple sequence alignments. *Methods in Enzymology*, *266*, 383–402. https://doi.org/10.1016/s0076-6879(96)66024-8

Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q., & Vinh, L. S. (2018). UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Molecular Biology and Evolution*, *35*(2), 518–522. https://doi.org/10.1093/molbev/msx281

Hood, L., & Rowen, L. (2013). The Human Genome Project: Big science transforms biology and medicine. *Genome Medicine*, *5*(9), 79. https://doi.org/10.1186/gm483

Hu, H., Popp, P. F., Hughes, T. C. D., Roa-Eguiara, A., Rutbeek, N. R., Martin, F. J. O., Hendriks, I. A., Payne, L. J., Yan, Y., Humolli, D., Klein-Sousa, V., Songailiene, I., Wang, Y., Nielsen, M. L., Berry, R. M., Harms, A., Erhardt, M., Jackson, S. A., & Taylor, N. M. I. (2024). Structure and mechanism of the Zorya anti-phage defence system [Publisher: Nature Publishing Group]. *Nature*, 1–9. https://doi.org/10.1038/s41586-024-08493-8

Hu, H., Popp, P. F., Santiveri, M., Roa-Eguiara, A., Yan, Y., Martin, F. J. O., Liu, Z., Wadhwa, N., Wang, Y., Erhardt, M., & Taylor, N. M. I. (2023). Ion selectivity and rotor coupling of the Vibrio flagellar sodium-driven stator unit [Publisher: Nature Publishing Group]. *Nature Communications*, *14*(1), 4411. https://doi.org/10.1038/s41467-023-39899-z

James, R. H., Deme, J. C., Kjær, A., Alcock, F., Silale, A., Lauber, F., Johnson, S., Berks, B. C., & Lea, S. M. (2021). Structure and mechanism of the proton-driven motor that powers Type 9 secretion and gliding motility. *Nature microbiology*, *6*(2), 221–233. https://doi.org/10.1038/s41564-020-00823-6

Jensen, R. A. (2001). Orthologs and paralogs - we need to get it right [Number: 8 Publisher: BioMed Central]. *Genome Biology*, *2*(8), 1–3. https://doi.org/10.1186/gb-2001-2-8-interactions1002

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold [Number: 7873 Publisher: Nature Publishing Group]. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., & Jermiin, L. S. (2017). ModelFinder: Fast model selection for accurate phylogenetic estimates [Publisher: Nature Publishing Group]. *Nature Methods*, *14*(6), 587–589. https://doi.org/10.1038/nmeth.4285

Katoh, K., Misawa, K., Kuma, K.-i., & Miyata, T. (2002). MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, *30*(14), 3059–3066. Retrieved March 23, 2025, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC135756/

Koonin, E. V. (2005). Orthologs, Paralogs, and Evolutionary Genomics1 [Publisher: Annual Reviews]. *Annual Review of Genetics*, *39*(Volume 39, 2005), 309–338. https://doi.org/10.1146/annurev.genet.39.073003.114725

Macnab, R. M. (2003). How Bacteria Assemble Flagella [Publisher: Annual Reviews]. *Annual Review of Microbiology*, *57*(Volume 57, 2003), 77–100. https://doi.org/10.1146/annurev.micro.57.030502.090832

Mandadapu, K. K., Nirody, J. A., Berry, R. M., & Oster, G. (2015). Mechanics of torque generation in the bacterial flagellar motor. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(32), E4381–E4389. https://doi.org/10.1073/pnas.1501734112

Mar, J. C., Harlow, T. J., & Ragan, M. A. (2005). Bayesian and maximum likelihood phylogenetic analyses of protein sequence data under relative branch-length differences and model violation. *BMC Evolutionary Biology*, *5*(1), 8. https://doi.org/10.1186/1471-2148-5-8

Mariano, G., Deme, J. C., Readshaw, J. J., Grobbelaar, M. J., Keenan, M., El-Masri, Y., Bamford, L., Songra, S., Blower, T. R., Palmer, T., & Lea, S. M. (2025). Modularity of Zorya defense systems during phage inhibition [Publisher: Nature Publishing Group]. *Nature Communications*, *16*(1), 2344. https://doi.org/10.1038/s41467-025-57397-2

Marmon, L. (2013). Elucidating the origin of the ExbBD components of the TonB system through Bayesian inference and maximum-likelihood phylogenies. *Molecular Phylogenetics and Evolution*, *69*(3), 674–686. https://doi.org/10.1016/j.ympev.2013.07.010

Meng, E. C., Goddard, T. D., Pettersen, E. F., Couch, G. S., Pearson, Z. J., Morris, J. H., & Ferrin, T. E. (2023). UCSF ChimeraX: Tools for structure building and analysis [_eprint:

https://onlinelibrary.wiley.com/doi/pdf/10.1002/pro.4792]. *Protein Science*, *32*(11), e4792. https://doi.org/10.1002/pro.4792

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, *37*(5), 1530–1534. https://doi.org/10.1093/molbev/msaa015

Pallen, M. J., & Matzke, N. J. (2006). From The Origin of Species to the origin of bacterial flagella [Number: 10 Publisher: Nature Publishing Group]. *Nature Reviews Microbiology*, *4*(10), 784–790. https://doi.org/10.1038/nrmicro1493

Payne, L. J., Todeschini, T. C., Wu, Y., Perry, B. J., Ronson, C. W., Fineran, P. C., Nobrega, F. L., & Jackson, S. A. (2021). Identification and classification of antiviral defence systems in bacteria and archaea with PADLOC reveals new system types. *Nucleic Acids Research*, *49*(19), 10868–10878. https://doi.org/10.1093/nar/gkab883

Pearson, T., Hornstra, H. M., Sahl, J. W., Schaack, S., Schupp, J. M., Beckstrom-Sternberg, S. M., O'Neill, M. W., Priestley, R. A., Champion, M. D., Beckstrom-Sternberg, J. S., Kersh, G. J., Samuel, J. E., Massung, R. F., & Keim, P. (2013). When outgroups fail; phylogenomics of rooting the emerging pathogen, Coxiella burnetii. *Systematic Biology*, *62*(5), 752–762. https://doi.org/10.1093/sysbio/syt038

Petiti, M., Serrano, B., Faure, L., Lloubes, R., Mignot, T., & Duché, D. (2019). Tol Energy-Driven Localization of Pal and Anchoring to the Peptidoglycan Promote Outer-Membrane Constriction. *Journal of Molecular Biology*, *431*(17), 3275–3288. https://doi.org/10.1016/j.jmb.2019.05.039

Philippe, H., Brinkmann, H., Lavrov, D. V., Littlewood, D. T. J., Manuel, M., Wörheide, G., & Baurain, D. (2011). Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough [Publisher: Public Library of Science]. *PLOS Biology*, *9*(3), e1000602. https://doi.org/10.1371/journal.pbio.1000602

Puente-Lelievre, C., Malik, A. J., Douglas, J., Ascher, D., Baker, M., Allison, J., Poole, A., Lundin, D., Fullmer, M., Bouckert, R., Kim, H., Steinegger, M., & Matzke, N. (2024, January). Tertiary-interaction characters enable fast, model-based structural phylogenetics beyond the twilight zone [Pages: 2023.12.12.571181 Section: New Results]. https://doi.org/10.1101/2023.12.12.571181

Puente-Lelievre, C., Ridone, P., Douglas, J., Amritkar, K., Kaçar, B., Baker, M., & Matzke, N. (2024, July). Molecular and structural innovations of the stator motor complex at the dawn of flagellar motility [Pages: 2024.07.22.604496 Section: New Results]. https://doi.org/10.1101/2024.07.22.604496

Rambaut, A., Drummond, A. J., Xie, D., Baele, G., & Suchard, M. A. (2018). Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Systematic Biology*, *67*(5), 901–904. https://doi.org/10.1093/sysbio/syy032

Ratliff, A. C., Buchanan, S. K., & Celia, H. (2022). The Ton Motor [Publisher: Frontiers]. *Frontiers in Microbiology*, *13*. https://doi.org/10.3389/fmicb.2022.852955

Rehman, I., Kerndt, C. C., & Botelho, S. (2025). Biochemistry, Tertiary Protein Structure. In *StatPearls*. StatPearls Publishing. Retrieved March 19, 2025, from http://www.ncbi.nlm.nih.gov/books/NBK470269/

Rieu, M., Krutyholowa, R., Taylor, N. M. I., & Berry, R. M. (2022). A new class of biological ion-driven rotary molecular motors with 5:2 symmetry [Publisher: Frontiers]. *Frontiers in Microbiology*, *13*. https://doi.org/10.3389/fmicb.2022.948383

Rost, B. (n.d.). Twilight zone of protein sequence alignments. Retrieved March 19, 2025, from https://dx.doi.org/10.1093/protein/12.2.85

Santiveri, M., Roa-Eguiara, A., Kühne, C., Wadhwa, N., Hu, H., Berg, H. C., Erhardt, M., & Taylor, N. M. I. (2020). Structure and Function of Stator Units of the Bacterial Flagellar Motor. *Cell*, *183*(1), 244–257.e16. https://doi.org/10.1016/j.cell.2020.08.016

Shibata, S., Tahara, Y. O., Katayama, E., Kawamoto, A., Kato, T., Zhu, Y., Nakane, D., Namba, K., Miyata, M., McBride, M. J., & Nakayama, K. (2023). Filamentous structures in the cell envelope are associated with bacteroidetes gliding machinery [Publisher: Nature Publishing Group]. *Communications Biology*, *6*(1), 1–11. https://doi.org/10.1038/s42003-023-04472-3

Shrivastava, A., Johnston, J. J., van Baaren, J. M., & McBride, M. J. (2013). Flavobacterium johnsoniae GldK, GldL, GldM, and SprA Are Required for Secretion of the Cell Surface Gliding Motility Adhesins SprB and RemA [Publisher: American Society for Microbiology]. *Journal of Bacteriology*, *195*(14), 3201–3212. https://doi.org/10.1128/jb.00333-13

Summers, J. K., &, .-. (2022). Predation Strategies of the Bacterium Bdellovibrio bacteriovorus Result in Overexploitation and Bottlenecks [Publisher: American Society for Microbiology]. *Applied and Environmental Microbiology*, *88*(1), e01082–21. https://doi.org/10.1128/AEM.01082-21

Tan, J., Zhang, L., Zhou, X., Han, S., Zhou, Y., & Zhu, Y. (2024). Structural basis of the bacterial flagellar motor rotational switching [Publisher: Nature Publishing Group]. *Cell Research*, *34*(11), 788–801. https://doi.org/10.1038/s41422-024-01017-z

Terashima, H., Fukuoka, H., Yakushi, T., Kojima, S., & Homma, M. (2006). The Vibrio motor proteins, MotX and MotY, are associated with the basal body of Na-driven flagella and required for stator formation. *Molecular Microbiology*, *62*(4), 1170–1180. https://doi.org/10.1111/j.1365-2958.2006.05435.x

van Kempen, M., Kim, S. S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C. L. M., Söding, J., & Steinegger, M. (2024). Fast and accurate protein structure search with Foldseek [Publisher: Nature Publishing Group]. *Nature Biotechnology*, *42*(2), 243–246. https://doi.org/10.1038/s41587-023-01773-0

van Leewenhoeck, A. (1677). Observations, Communicated to the Publisher by Mr. Antony van Leewenhoeck, in a Dutch Letter of the 9th of Octob. 1676. Here English'd: Concerning Little Animals by Him Observed in Rain-Well-Sea. and Snow Water; as Also in Water Wherein Pepper Had Lain Infused [Publisher: The Royal Society]. *Philosophical Transac-*

tions (1665-1678), *12*, 821–831. Retrieved March 18, 2025, from https://www.jstor.org/stable/101758

Vilas Boas, D., Castro, J., Araújo, D., Nóbrega, F. L., Keevil, C. W., Azevedo, N. F., Vieira, M. J., & Almeida, C. (2024). The Role of Flagellum and Flagellum-Based Motility on Salmonella Enteritidis and Escherichia coli Biofilm Formation [Number: 2 Publisher: Multidisciplinary Digital Publishing Institute]. *Microorganisms*, *12*(2), 232. https://doi.org/10.3390/microorganisms12020232

Wagner, G. P. (1989). The Biological Homology Concept [Publisher: Annual Reviews]. *Annual Review of Ecology and Systematics*, *20*, 51–69. Retrieved October 15, 2023, from https://www.jstor.org/stable/2097084

Wagner, G. P. (2007). The developmental genetics of homology [Publisher: Nature Publishing Group]. *Nature Reviews Genetics*, *8*(6), 473–479. https://doi.org/10.1038/nrg2099

Williams-Jones, D. P., Webby, M. N., Press, C. E., Gradon, J. M., Armstrong, S. R., Szczepaniak, J., & Kleanthous, C. (2023). Tunable force transduction through the *Escherichia coli* cell envelope [Company: National Academy of Sciences Distributor: National Academy of Sciences Institution: National Academy of Sciences Label: National Academy of Sciences Publisher: Proceedings of the National Academy of Sciences]. *Proceedings of the National Academy of Sciences*, *120*(47), e2306707120. https://doi.org/10.1073/pnas.2306707120

Wojdyla, J. A., Cutts, E., Kaminska, R., Papadakos, G., Hopper, J. T. S., Stansfeld, P. J., Staunton, D., Robinson, C. V., & Kleanthous, C. (2015). Structure and Function of the Escherichia coli Tol-Pal Stator Protein TolR * [Publisher: Elsevier]. *Journal of Biological Chemistry*, *290*(44), 26675–26687. https://doi.org/10.1074/jbc.M115.671586

Yamaguchi, T., Makino, F., Miyata, T., Minamino, T., Kato, T., & Namba, K. (2021). Structure of the molecular bushing of the bacterial flagellar motor [Publisher: Nature Publishing Group]. *Nature Communications*, *12*(1), 4469. https://doi.org/10.1038/s41467-021-24715-3

Ye, J., McGinnis, S., & Madden, T. L. (2006). BLAST: Improvements for better sequence analysis. *Nucleic Acids Research*, *34*(Web Server issue), W6–9. https://doi.org/10.1093/nar/gkl164

Young, A. D., & Gillung, J. P. (2020). Phylogenomics — principles, opportunities and pitfalls of big-data phylogenetics [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/syen.12406]. *Systematic Entomology*, *45*(2), 225–247. https://doi.org/10.1111/syen.12406

Zhang, C., Shine, M., Pyle, A. M., & Zhang, Y. (2022, April). US-align: Universal Structure Alignments of Proteins, Nucleic Acids, and Macromolecular Complexes [Pages: 2022.04.18.488565 Section: New Results]. https://doi.org/10.1101/2022.04.18.488565