# Structural phylogenetics of bacterial flagellum proteins FliP, FliQ, FliR


**THE UNIVERSITY OF AUCKLAND**


## Changhao Li

## Student ID: 408384361

## Supervised by: Nicholas J. Matzke


**A THESIS SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF M ASTER OF BIOTECHNOLOGY,**

**Abstract:**

Bacterial flagella and type III secretion systems (T3SS) share core components, particularly the inner-membrane proteins FliP, FliQ, and FliR, which form part of the flagellar T3SS (F-T3SS) export apparatus; these proteins also have homologs in non-flagellar (injectisome) T3SS (NF-T3SS). Due to extensive sequence divergence among these proteins, their evolutionary relationships have remained unresolved by conventional sequence phylogenetics. This study employs structural phylogenetics to clarify the homology and evolutionary history of FliP, FliQ, and FliR. We used AlphaFold2 structural predictions of FliP, FliQ, and FliR from diverse bacteria (including flagellated species and those with only NF-T3SS), and encoded characters using a 3Di structural alphabet. Multiple alignments were constructed using structure-guided approaches: FoldMason for progressive protein structure alignment, FAMSA3di for fast 3Di-based sequence alignment, and MAFFT for purely AA sequence-based comparison. We then inferred phylogenetic trees using IQ-TREE, employing maximum-likelihood methods with models that integrate amino acid and 3Di characters.

The resulting phylogeny reveals well-supported relationships among FliP, FliQ, and FliR homologs across bacterial phyla, confirming their common ancestry. Integrating structural and amino acid information is essential for correctly aligning the highly divergent FliQ and FliR sequences, thus resolving the misalignment problem commonly seen in methods based solely on AA sequences. Notably, homologous proteins from NF-T3SS cluster within the FliP, FliQ, and FliR clades, indicating that these virulence-associated T3SS components were derived from ancestral flagellar

FliPQR proteins. This suggests that the flagellar exit gate was coopted in the evolution of pathogenic T3SS. Overall, our findings illuminate the evolutionary path from flagella to injectisomes and demonstrate the power of structure-based phylogenetic approaches to reveal deep relationships, providing new insights into the evolution of bacterial secretion systems.

**Contents**

# 1. Background

### 1.1.1Bacterial Flagella: Structure, Function, and Scientific Significance

Bacterial flagella represent one of the most sophisticated nanomachines found in nature, performing the essential function of motility in a wide variety of prokaryotic organisms. The study of flagellar biology has profoundly influenced the fields of microbiology, molecular biology, and structural biology, serving as a model for understanding macromolecular assembly, evolution, and the interplay of structure and function in biological systems (Macnab, 2003; Chevance & Hughes, 2008; Minamino & Imada, 2015). Flagella not only confer the ability to swim and swarm but also underpin processes such as chemotaxis, colonization, biofilm formation, and pathogenesis, contributing to the ecological success and adaptability of bacteria (Wadhams & Armitage, 2004; Morimoto & Minamino, 2014).

### 1.1.2 Architectural Overview of the Flagellum

The bacterial flagellum is classically divided into three principal regions: the filament, the hook, and the basal body (Abrusci et al., 2014; Galán et al., 2014). Each region is composed of distinct proteins that self-assemble into highly ordered supramolecular complexes.

**The filament** is the most prominent part, appearing as a long, thin helical tube that projects outward from the bacterial cell surface. It is composed predominantly of thousands of flagellin (FliC) monomers, which polymerize into a hollow structure with remarkable flexibility and resilience (Samatey et al., 2001; Macnab, 2003). The

flagellar filament acts as a propeller, generating thrust by rotating at high speeds, sometimes exceeding several hundred revolutions per second (Chevance & Hughes, 2008). The biophysical properties of the filament, such as elasticity and the ability to undergo polymorphic transitions, are essential for efficient locomotion and rapid changes in swimming direction, especially during chemotaxis (Turner et al., 2000; Calladine et al., 2013).

**The hook** is a universal joint between the filament and the basal body. Constructed primarily from FlgE subunits, the hook is a curved, flexible structure that transmits torque while permitting the filament to bend and reorient relative to the cell body (Minamino & Imada, 2015; Erhardt et al., 2010). This flexibility is critical for the characteristic "run-and-tumble" motility of many bacteria, enabling them to navigate chemical gradients and escape from hostile environments (Wadhams & Armitage, 2004).

**The basal body** is the most structurally and functionally complex region of the flagellum. Embedded within the bacterial cell envelope, it serves as both an anchor and as the rotary engine of the flagellum (Cornelis, 2006; Galán et al., 2014). The basal body consists of several ring structures, including the L-ring (located in the outer membrane of Gram-negative bacteria), the P-ring (embedded in the peptidoglycan layer), the MS-ring (spanning the cytoplasmic membrane), and the C-ring (situated in the cytoplasm) (Terashima et al., 2008; Chen et al., 2011)(Figure1). These rings not only stabilize the flagellum as it traverses the cell envelope, but also facilitate the assembly and operation of the rotary motor.

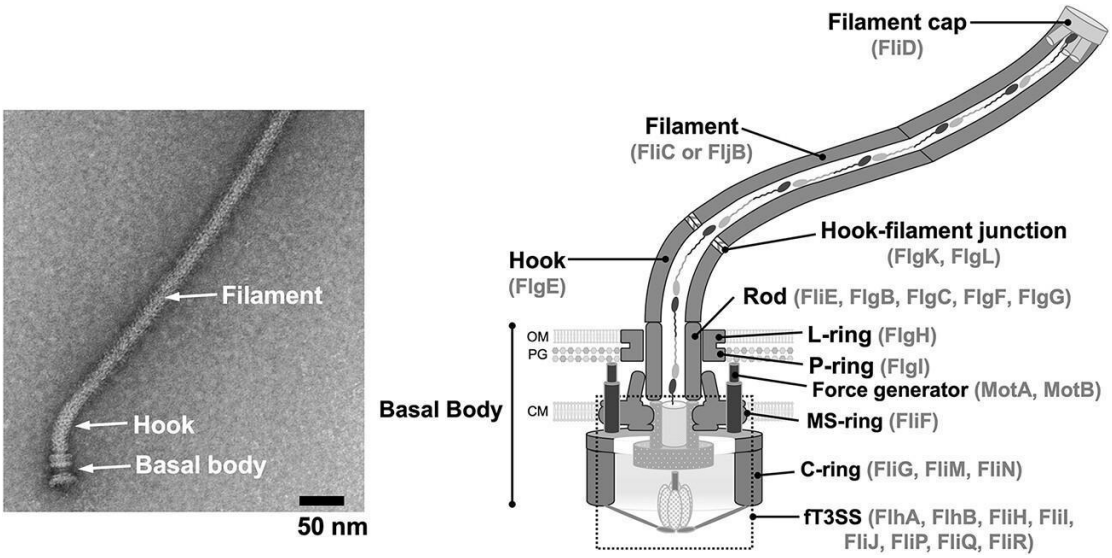



Figure 1 Schematic diagram of bacterial flagella (source: Minamino & Kinoshita, 2023)

### 1.1.3. Rotary Motor Mechanism: Rotor, Stator, and Ion Motive Force

At the core of the flagellar basal body is the rotary motor, which converts ion gradients across the bacterial membrane into mechanical torque. This is accomplished by the rotor—primarily composed of the MS-ring (FliF) and the C-ring (FliG, FliM, FliN)—and the stator, which is typically formed by MotA and MotB proteins in proton-driven systems PomA and PomB in the sodium-driven system (Santiveri et al., 2020). The stator complexes are anchored to the peptidoglycan layer and surround the rotor, forming multiple ion-conducting channels (Kojima & Blair, 2004).

As protons (or sodium ions) flow through the stator channels, they induce conformational changes in the stator proteins, which are coupled to the rotor via electrostatic and hydrophobic interactions (Kojima & Blair, 2004; Morimoto & Minamino, 2014). This process generates rotational torque, allowing the flagellum to spin at high speed. The precise stoichiometry and dynamic assembly of the stator and rotor have been elucidated by high-resolution structural methods, such as cryo-

electron microscopy (Chen et al., 2011).

The efficiency and reversibility of the bacterial flagellar motor underpin critical physiological behaviours. By switching the direction of rotation, the motor enables bacteria to alternate between smooth swimming (counterclockwise rotation) and tumbling (clockwise rotation), the basis for chemotactic movement (Wadhams & Armitage, 2004).

### 1.1.4. Flagellar Protein Export and Assembly: The Type III Secretion System (F-T3SS)

Assembly of the flagellum is a highly regulated, hierarchical process. Central to this is the F-T3SS, a specialized protein export apparatus that translocates structural subunits from the cytoplasm through the nascent flagellar structure to the site of assembly at the filament tip (Diepold & Armitage, 2015; Minamino & Imada, 2015).

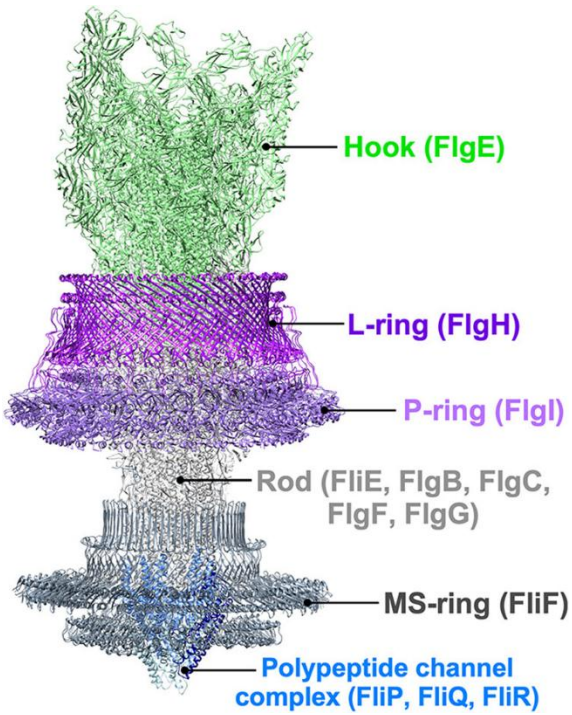



Figure 2. Recent structural model of the basal export apparatus and rod/hook region of the bacterial flagellum (source: Minamino & Kinoshita, 2023).

The F-T3SS consists of a membrane-embedded export gate, including proteins such as FliP, FliQ, FliR, FlhA, and FlhB, as well as a cytoplasmic ATPase complex (FliH, FliI, FliJ) (Kuhlen et al., 2018). The FliPQR complex, in particular, forms a helical export gate with a defined stoichiometry (5:4:1) (Figure 3), serving as the primary channel for substrate translocation (Abrusci et al., 2014; Kuhlen et al., 2018). These export gate proteins coordinate with chaperones and regulatory factors to ensure precise timing and order of assembly, allowing the flagellum to self-assemble from the inside out (Minamino et al., 2017).

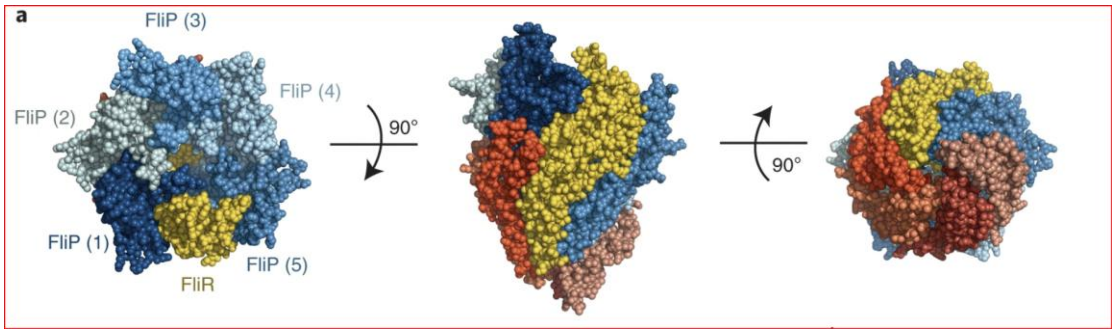



Figure 3. Image showing how the FliPQR protein forms a complete export gate structure (source: Kuhlen et al., 2018).

Importantly, the F-T3SS is homologous to the virulence-associated Type III secretion systems (NF-T3SS) found in many pathogenic Gram-negative bacteria, known as injectisomes (Blocker et al., 2003; Abby & Rocha, 2012). Comparative genomics and structural biology have shown that the core export machinery, including the FliPQR complex and its non-flagellar homologs (SctR, SctS, SctT, and similar names such as YscRST, EscRST, HrpRST, etc.; some NF-T3SS proteins described earlier have other naming schemes), reflects an evolutionary relationship in which the injectisome T3SS originated from an ancestral flagellar system (Kuhlen et al., 2018; Diepold & Armitage, 2015).

### 1.1.5. Scientific Importance and Research Relevance

Flagellar motility is not only fundamental for the behaviour and ecology of many bacteria but also has direct implications for human health, environmental microbiology, and biotechnology. Motility is a key virulence determinant in many pathogens, enabling colonization and invasion of host tissues (Chevance & Hughes, 2008). Disrupting flagellar assembly or function is, therefore, a promising strategy for antimicrobial development (Morimoto & Minamino, 2014).

From a scientific perspective, the bacterial flagellum exemplifies core principles of molecular evolution, including modularity, self-assembly, and the adaptive reuse of protein complexes. Detailed study of its structure and biogenesis provides insights into the origins of complex cellular machinery and the mechanisms by which evolutionary innovations are achieved (Abby & Rocha, 2012; Blocker et al., 2003).

### 1.2.1 Molecular Phylogenetics: Theory, Challenges, and Modern Solutions

Molecular phylogenetics is a cornerstone of modern biology, providing the framework for reconstructing the evolutionary history of genes, proteins, and organisms. By comparing biological sequences, scientists can infer patterns of descent, functional relationships, and the origins of complex molecular machines such as the bacterial flagellum and its associated T3SS (Gabaldón & Koonin, 2013; Felsenstein, 2004). In the study of bacterial motility and secretion systems, molecular phylogenetics is used for tracing the evolutionary relationships between the flagellar export apparatus and virulence-associated T3SS injectisomes (Blocker et al., 2003; Abby & Rocha, 2012).

### 1.2.2 The Purpose and Significance of Phylogenetic Trees

A phylogenetic tree represents a hypothesis of evolutionary relationships among

biological entities—be they species, genes, or proteins (Felsenstein, 2004; Eisen, 1998). In the context of protein families, such trees reveal how gene duplications, losses, and horizontal transfers have shaped the complexity and diversity of life. Phylogenetic analyses inform not only our understanding of evolutionary history but also practical aspects such as protein function prediction, identification of orthologs and paralogs, and the discovery of new therapeutic targets (Gabaldón & Koonin, 2013).

The rationale for constructing phylogenetic trees in protein evolution is multi-faceted. Typically, proteins are grouped for analysis based on established or hypothesized homology, and the resulting trees are used to explore the evolutionary relationships within this set. First, phylogenetic analyses can clarify the relationships among homologous proteins, helping to trace the diversification of protein families and to identify patterns of functional conservation, domain architecture changes, and the modular assembly of complex molecular systems. Second, for systems like the F-T3SS and NF-T3SS relatives, phylogenetic trees are essential for reconstructing the sequence of key evolutionary events such as gene duplications, gene fusions, domain shuffling, and the emergence of novel functionalities (Abby & Rocha, 2012; Blocker et al., 2003).

### 1.2.3 Concepts: Homology, Orthology, and Paralogy

Homology is the fundamental concept at the heart of molecular phylogenetics. In its strict sense, homology refers to similarity due to common ancestry rather than to convergent or parallel evolution (Nixon & Carpenter, 2011). Within homologous proteins, orthologs diverge after a speciation event, while paralogs diverge following a gene duplication event (Gabaldón & Koonin, 2013). Distinguishing between these is

essential, as orthologs tend to retain ancestral functions, while paralogs may acquire novel or specialized roles.

Accurate homology inference requires careful consideration of sequence conservation, domain architecture, and, three-dimensional structural information (Orengo & Thornton, 2005). In ancient protein families, such as FliPQR and their NF-T3SS homologs, the detection of homology may be challenging, as will any downstream analyses such as multiple sequence alignment and phylogenetic inference, necessitating the use of advanced computational and structural tools (Holm, 2019).

### 1.2.4 Multiple Sequence Alignment (MSA) and Its Challenges

A critical first step in phylogenetic analysis is the construction of a multiple sequence alignment (MSA), where homologous residues across different proteins are aligned in columns. This enables the identification of conserved motifs, functional residues, and the evolutionary relationships between sequences (Katoh & Standley, 2013). Modern MSA tools such as MAFFT and MUSCLE use fast Fourier transform algorithms and progressive alignment strategies to handle large, diverse datasets efficiently (Katoh et al., 2002; Edgar, 2004).

However, the reliability of MSA and, by extension, phylogenetic inference diminishes as sequence similarity decreases—a phenomenon known as the “twilight zone” (Rost, 1999; Chung & Subbiah, 1996). When sequence identity falls below 25–30%, the statistical confidence in the alignment of residues and the detection of homology drops precipitously (Holm, 2019). At these evolutionary distances, substitutions, insertions, deletions, and convergent evolution obscure the true evolutionary signal, leading to alignment errors, long-branch attraction, and false positives or negatives in

homology detection (Philippe et al., 2011).

### 1.2.5 Structure-Based Homology and the 3Di Alphabet

Protein three-dimensional (3D) structure is often more conserved than primary amino acid sequence, even over long evolutionary timescales (Orengo & Thornton, 2005; Holm, 2019). As a result, structure-based methods can be valuable for detecting deep homology in divergent protein families. High-resolution structural determination (e.g., cryo-EM, X-ray crystallography) and especially computational prediction (AlphaFold) have greatly expanded the availability of structural models for proteins (Jumper et al., 2021).

### 1.2.6 Foldseek and Advances in Structure-Based Phylogenetics

The development of tools such as Foldseek has revolutionized structure-based bioinformatics (van Kempen et al., 2023). Foldseek enables ultra-rapid comparisons of protein structures at the scale of entire proteomes by converting 3D models into 3Di "sequences" that can be aligned and searched with string-matching algorithms. This is particularly powerful for detecting remote homology—relationships that cannot be reliably inferred by sequence alone due to divergence into the twilight zone. Having a one-dimensional "sequence" encoding structural information also enables structural information to be incorporated into MSAs and phylogenetic analyses using computer algorithms originally designed for traditional sequence data (van Kempen et al., 2023; Matzke et al., 2025).

FoldMason and FAMSA3di are further advances that enable MSA based on 3Di alphabets, allowing researchers to create alignments and build phylogenies based on structural conservation (Gilchrist et al., 2024; Puente-Lelievre et al., 2024). These

approaches provide a robust path to clarify the evolutionary origins of ancient, highly diverged protein complexes, such as the flagellar FliPQR and T3SS homologs, where sequence-based methods fail.

### 1.2.7 Hybrid Approaches and Modern Phylogenetic Tree Construction

The integration of amino acid (AA) and 3Di structure-based alignments is a new frontier in molecular phylogenetics (Matzke et al., 2025; Garg & Hochberg, 2024). Hybrid MSAs—combining sequence and structure—may be more robust to alignment errors, and may be able to resolve ancient evolutionary histories even in the face of extensive sequence divergence (Garg & Hochberg, 2024; Puente-Lelievre et al., 2024). Modern tree-building tools such as IQ-TREE allow for partitioned analyses, where each data type (AA, 3Di) can use its optimal substitution model, improving the statistical support and resolution of both shallow and deep evolutionary relationships (Minh et al., 2020; Puente-Lelievre et al., 2024).

### 1.2.8. Rationale and Outlook

Given the complex evolutionary history of T3SS components—characterized by gene duplication, fusion, and divergence—integrated molecular phylogenetic analyses are essential to disentangle orthologous and paralogous relationships, identify deep homology, and reconstruct the origins of key functional modules. The advances in structure-based alphabets, rapid structural alignment tools, and combined MSA strategies now enable a new era in the study of molecular evolution, allowing for the elucidation of ancestral relationships and the modular evolution of bacterial nanomachines such as FliPQR (Abby & Rocha, 2012; Kuhlen et al., 2018; Blocker et al., 2003).

As the field progresses, the integration of bioinformatics, structural biology, and evolutionary genomics promises not only to deepen our understanding of protein evolution but also to inform new strategies for combating bacterial pathogens and engineering synthetic biological systems.

## 2. Methods

### 2.1 Overview of Analytical Strategy

The central aim of this study was to resolve the evolutionary relationships of the bacterial flagellar export apparatus proteins FliP, FliQ, and FliR and their homologs in both F-T3SS and NF-T3SS Type III Secretion Systems. Recognizing the inherent limitations of AA sequence-only approaches—especially in highly divergent protein families—this research employed a dual strategy integrating both sequence-based and structure-based alignments. By combining MSA of amino acid (AA) sequences with structural MSA based on the 3Di structural alphabet (Foldseek), it was possible to infer deep homology and evolutionary connections obscured by sequence divergence alone. The strategy included (1) careful curation and preprocessing of sequence and structure data, (2) alignment using multiple algorithms tailored to sequence or structure, (3) rigorous validation and cross-format comparison of alignment results, and (4) robust phylogenetic inference using state-of-the-art statistical methods. Each step was benchmarked for reliability and reproducibility, ensuring confidence in downstream evolutionary interpretations.

Our rationale for integrating amino acid (AA) sequence and 3Di structure-derived data was twofold: first, to exploit the higher conservation of protein tertiary structure compared to sequence, and second, to maximize the phylogenetic information

available for the detection of both recent and ancient homologies (Holm, 2019; Orengo & Thornton, 2005; van Kempen et al., 2023). By systematically comparing alignments and trees derived from each data type—and their combination—we aimed to overcome the limitations of sequence-only approaches and achieve a robust evolutionary reconstruction.

## 2.2. Sequence Alignment Methods

Amino acid multiple sequence alignments (MSAs) were generated using MAFFT version 7 (Katoh & Standley, 2013), a widely used tool for protein sequence alignment. For all datasets in this study, alignments were performed using the default MAFFT settings, which provide a balance between speed and accuracy and are suitable for typical protein sequence data. The default mode in MAFFT automatically selects the alignment algorithm based on the size and complexity of the dataset, without requiring manual adjustment of advanced parameters.

After generating the alignments, all MSAs were carefully inspected using AliView (Larsson, 2014), a lightweight alignment viewer and editor. This manual review was important to identify and correct any obvious misalignments, artificial gaps, or poorly aligned terminal regions that sometimes arise, especially in diverse or incomplete datasets. Where necessary, minor manual edits were made to improve the overall alignment quality and ensure that conserved regions were correctly aligned across all sequences.

To ensure comparability with structure-based alignments and to focus on the biologically relevant regions, only the mature protein sequences were used for alignment. For example, signal peptides at the N-terminus of FliP—which are cleaved off during post-translational processing in vivo (Macnab, 2003)—were excluded from the analysis. Removal of these regions helped avoid spurious alignment artifacts and allowed for more meaningful comparison between homologous domains.

## 2.3. Structural Alignment Methods

To overcome the limitations of sequence-based approaches in the twilight zone of low

sequence similarity, we performed structure-based alignments making use of 3Di characters, using two strategies: FAMSA3di and FoldMason.

**FAMSA3di**

FAMSA (Fast and Accurate Multiple Sequence Alignment) is a progressive MSA algorithm optimized for large protein families, using a bit-parallel implementation and in-situ alignment for high speed and memory efficiency (Deorowicz et al., 2016). In its 3DI variant, the traditional amino acid substitution matrix in the FAMSA source code is replaced with a 3Di cost matrix derived by van Kempen et al. (2023). The 3Di alphabet encodes the local three-dimensional environment of each residue as one of 20 discrete states, enabling the linear representation of tertiary structure as a sequence. This approach allows FAMSA3di to align proteins whose primary sequences are highly divergent, leveraging the higher conservation of structure to recover deep homology (Puente-Lelievre et al., 2024).

FAMSA3di alignments were performed using default parameters for guide tree construction and profile merging with the 3Di structural alphabet obtained from Foldseek.

**FoldMason**

FoldMason is an ultrafast, scalable method for MSA designed to accommodate the rapidly expanding volume of protein structure data from AlphaFold and related resources (Gilchrist et al., 2024). FoldMason uses a **combined alphabet** encoding both 3Di and amino acid information, converting the protein structure into a one-dimensional string. It builds an all-vs-all similarity matrix, constructs a minimum

spanning tree, and performs a parallelized progressive profile alignment. The alignment is iteratively optimized to maximize the **local distance difference test (LDDT)** score, a widely used measure of structural agreement (Mariani et al., 2013).

Benchmark tests show that FoldMason achieves accuracy comparable to traditional structural alignment tools (US-align, MUSTANG, Matt) but with far greater speed, making it uniquely suited for large-scale phylogenetic applications (Gilchrist et al., 2024). FoldMason was used for both AA and 3Di data types to allow direct comparison of sequence-only, structure-only, and hybrid alignments.

## 2.4. Database and Sequence Processing

### 2.4.1 Database Construction

For this study, we wanted to sample FliPQR homologs from across the bacterial phyla, while keeping the number of sequences within computationally manageable limits, i.e., a few hundred sequences per protein. To this end, we used a list of 193 genomes selected from across 27 bacterial phyla by Puente-Lelievre et al. (2025). The sampling was informed by the large-scale genome survey of Philip et al. (2025), where 11,365 complete and non-redundant bacterial genomes were selected from the PATRIC database to maximize phylogenetic coverage and minimize redundancy as described in Philip et al. (2025).

For the homology search, we selected the FliP protein from *Escherichia coli* K-12 (UniProtKB: P0ABT6; PDB: 6S3O) as the reference query. *E. coli* K-12 is a model organism with a completely sequenced and well-annotated genome, and its FliP protein is among the best-characterized flagellar export gate components, with

abundant genetic, structural, and biochemical data available (Erhardt et al., 2010; Kuhlen et al., 2018).

For the homology search, the jackhmmer program from the HMMER suite (Eddy, 2011) was used. Jackhmmer performs iterative profile hidden Markov model (HMM) searches, which are effective for detecting distant homologs. Specifically, the FliP protein from *E. coli* K-12 was used as the input sequence (FliP_Ecoli.fasta), and the search was conducted against the full protein database (proteinall.faa) used by Puente-Lelievre et al. (2025). We used the command:

jackhmmer -o fliP_hmmer.txt --tblout fliP_hmmer.csv -A fliP_hmmer_alignment FliP_Ecoli.fasta proteinall.faa

The output included a table of significant hits (fliP_hmmer.csv), alignment information (fliP_hmmer_alignment), and a summary report (fliP_hmmer.txt). Hits above the established significance threshold were retained for downstream analyses. This approach allowed us to build a phylogenetically comprehensive set of FliPQR-related proteins for subsequent sequence alignment, phylogenetic tree construction, and structural comparison.

### 2.4.2. Signal Peptide Removal

Signal peptides and other non-homologous N-terminal extensions were annotated using UniProt and SignalP (Almagro Armenteros et al., 2019), and removed prior to alignment to avoid spurious matches and improve homology detection. For example,

residues 1–21 constitute the signal peptide of *E. coli* FliP, and were excluded from subsequent analyses.

### 2.4.3. Alignment Workflow

Three alignment strategies were systematically compared:

- **AA only**: Conventional alignment of amino acid sequences using MAFFT and FoldMason.
- **3Di only**: Alignment of 3Di structural sequences using FAMSA3di and FoldMason.
- **AA+3Di**: Concatenated alignments where 3Di sequences were appended to corresponding AA sequences.

Each dataset was processed with the above methods, generating a matrix of alignments for downstream comparison.

### 2.4.4. Cross-Format Conversion and Consistency Checks

To ensure the robustness of alignment approaches and to identify any errors introduced by algorithm-specific artifacts (such as repetitive 3Di characters in FAMSA3di alignments), cross-format conversions were performed:

- MAFFT AA alignments were converted to 3Di format using R script.
- FAMSA3di 3Di alignments were converted to AA format using R script.
- Consistency between aligned datasets was assessed by visualizing alignments in AliView, checking for the preservation of conserved regions, and identifying misaligned or ambiguous segments.

The conversion functions used are available in the Supplementary Material of Matzke et al. (2025).

### 2.4.5. Alignment Quality Control with TrimAl

Ensuring the highest quality of MSAs is paramount for robust phylogenetic inference. Even with careful alignment strategies, MSAs can include columns that are poorly conserved, highly variable, or represent alignment ambiguities—particularly in studies involving anciently diverged protein families such as FliP, FliQ, and FliR. These low-quality columns, if retained, may introduce noise, mislead model estimation, and ultimately reduce the accuracy and interpretability of phylogenetic trees.

To address this issue, we employed TrimAl (Capella-Gutiérrez et al., 2009), a widely used automated tool for alignment trimming. TrimAl systematically evaluates each column in the MSA and removes those that fall below a user-specified conservation threshold, thereby retaining only the most phylogenetically informative positions. In this study, we applied a 35% similarity cutoff: columns were retained only if at least 35% of the sequences had identical or similar residues in that position. This threshold strikes a balance between excluding excessively variable or ambiguously aligned regions and preserving sufficient informative sites for robust tree reconstruction.

### 2.4.6. Assessment of Alignment Quality

The clarity, extent, and coherence of conserved blocks showing detailed patterns of shared 3Di sequence similarity were used as the primary criteria for alignment

quality. The most biologically plausible MSA was defined as the one that best recapitulated three-dimensional superpositions of FliQ on FliP or FliR as visualised in ChimeraX and displayed consistent conservation across orthologous proteins from divergent lineages.

### 2.4.7. Manual Sanity Checks with ChimeraX

As 3Di-based alignment techniques are new, and as the FliPQR divergence is in the twilight zone where amino acid similarity is likely to be weak or undetactable, we performed some manual “sanity checks” to confirm that our MSAs were reasonable. The FliPQR protein structures of ***E. coli*** **K12** were opened in ChimeraX, and MatchMaker was used to produce structure-guided AA alignments. These were compared with computational MSAs as an independent reference, providing a "ground truth" for evaluating the congruence of different methods. The AA alignments generated from ChimeraX structural matching were also inserted into the MSAs to facilitate direct visual and quantitative comparison.

## 2.5. Internal Homology Analysis

To test hypotheses of internal duplication and domain homology (e.g., FliQ being homologous to two regions of FliR), concatenated FliQ sequences (FliQ-FliQ fusion) were constructured for both AA and 3Di sequences. The AAs were aligned to FliR using both MAFFT, and the 3Dis using FAMSA3di. Alignments were inspected for conserved blocks and structural overlap.

Manual validation was performed in ChimeraX (Pettersen et al., 2021), where the 3D structures of FliP, FliQ, and FliR were superimposed using the MatchMaker tool. The resulting structurally aligned sequences were inserted into MSAs as reference

benchmarks for evaluating the alignment accuracy of computational methods.

Once a satisfactory alignment was obtained, one of the FliQ copies was cut from the alignment. While either copy could be cut, as both FliR and FliP appear to be homologous to 2 copies of FliQ, for the purposes of the present study, the second FliQ copy was retained, due to the stronger similarity of FliQ to the C-terminal regions of FliP and FliR. This decision might conceivably be revisited in larger studies in the future that attempt to resolve the relationships of subregions of FliP and FliR.

**2.6. Phylogenetic Analysis**

A key prerequisite for reconstructing a reliable phylogenetic tree is to generate a high-quality, unified MSA. Careful visual inspections of the MSAs output by programs were used to check that a MSA successfully aligned homologous positions between the FliP, FliQ, and/or FliR groups. Correct alignment can preserve evolutionary signals and prevent the introduction of artefacts that may mislead inference of tree topology, or reduce the statistical support of key phylogenetic relationships.

To select the optimal evolutionary models for phylogenetic inference, we employed a comprehensive model-testing strategy using IQ-TREE v2 (Minh et al., 2020). Separate model selection procedures were performed for amino acid (AA), structure-based 3Di datasets, and combined AA+3Di partitioned datasets, respectively.

For individual datasets containing either AA or 3Di sequences, a broad range of substitution models was tested, including standard amino acid substitution matrices (Blosum62, Dayhoff, DCMut, JTT, JTTDCMut, LG, Poisson, Poisson+FQ, PMB,

WAG, EX2, EX3, EHO, and EX_EHO) as well as the 3DI substitution model tailored specifically for structure-derived 3Di sequences. We systematically evaluated empirical amino acid frequencies (F: empirical, FU: uniform frequencies) and rate heterogeneity among sites (E: equal rates, G: gamma distribution, R: free-rate model).

For models with AA or 3di, we used the command:

iqtree2 -s seqs.fasta -mset 3DI,Blosum62,Dayhoff,DCMut,JTT,JTTDCMut,LG,Poisson,Poisson+FQ,PMB,WAG,EX2,EX3,EHO,EX_EHO -mfreq FU,F -mrate E,G,R --ufboot 1000 -alrt 1000 -bnni --redo

For combined datasets (AA+3Di), we employed partitioned analyses to independently optimize evolutionary models for each partition (AA and 3Di). This method leverages distinct evolutionary dynamics of sequence and structural data simultaneously, thereby maximizing the phylogenetic signal from each type of information.

For combined AA+3Di datasets with partitioned analysis, we used the command:

iqtree2 -s seqs.fasta -spp BOTHp.raxml -m MFP+MERGE -madd 3DI -mdef 3DI.nexus -mset 3DI,Blosum62,Dayhoff,DCMut,JTT,JTTDCMut,LG,Poisson,Poisson+FQ,PMB,WAG,EX2,EX3,EHO,EX_EHO -mfreq FU,F -mrate E,G,R --ufboot 1000 -alrt 1000 -bnni --redo

In these workflows, the software assesses all specified models using statistical selection criteria (e.g., BIC and AIC), then automatically selects the best-fitting model for each MSA or partition. The final phylogenetic trees are estimated using these optimal models, ensuring that evolutionary rate heterogeneity and substitution

patterns are accurately accounted for.

The reliability of each tree was evaluated using 1,000 ultrafast bootstrap (UFBoot) replicates and 1,000 SH-aLRT replicates, using nearest neighbor interchange optimization to refine branch support values. The workflow for final tree estimation, using the best-fitting models and partitioning as determined above, was exemplified by the following command:

iqtree2 -s FliPRQ_AA3dis_trim35.fasta -redo -spp FliPRQ2_AA3dis_NJ.nexus -madd 3DI -mdef 3DI.nexus --ufboot 1000 -bnni | tee so2.txt &

Due to the absence of an appropriate outgroup (there are no well-established, reliably alignable proteins outside the FliPQR/NF-T3SS group), all phylogenies were midpoint-rooted. While midpoint rooting is not ideal and the position of the root must be viewed as tentative, this approach places the root in a reasonable and interpretable position, far from the tips, aiding the readability of phylogenetic trees and the discussion of evolutionary relationships.

By systematically evaluating model fit and leveraging the strengths of both AA and 3Di alignments, our phylogenetic analyses maximize the reliability and resolution of evolutionary relationships among FliP, FliQ, FliR, and their homologs.

# 3.Result

## 3.1. Quantitative Comparison of Alignment Methods

To evaluate the quality and characteristics of multiple alignment strategies for FliPQR protein homologs, we compared amino acid (AA) and 3Di structural alphabet

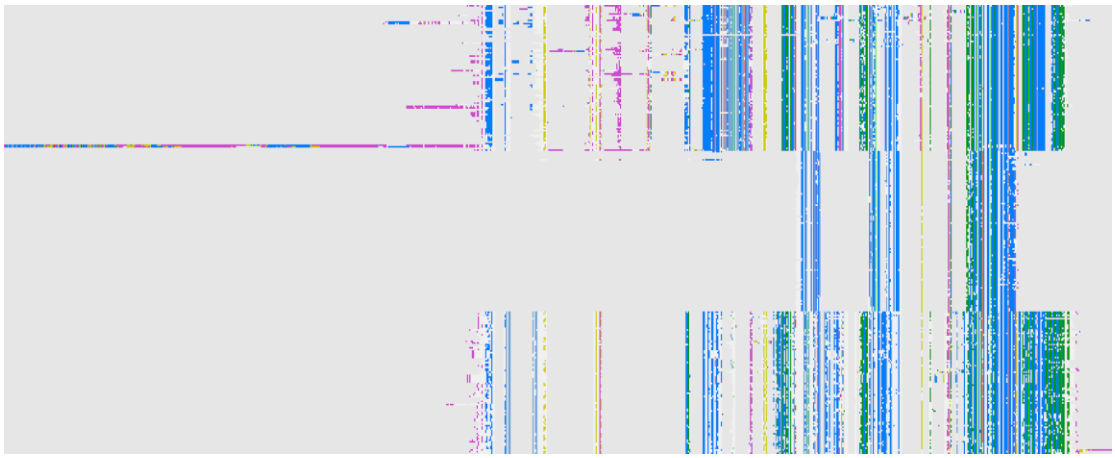
alignments using MAFFT and Foldmason, as well as reciprocal AA/3Di mapping methods. All alignments included between 275 and 277 sequences, with alignment lengths ranging from 771 to 945 residues (Figure 4). Notably, the gap percentage was consistently high across all methods (73.9%–78.6%), reflecting substantial sequence and structural diversity within the dataset. None of the alignments produced a large number of perfectly conserved AA columns; the highest observed was a single fully conserved position (0.1%) in both the 3Di-to-AA mapped alignment and one of the 3Di-based alignments, while all other approaches—including MAFFT AA and Foldmason AA alignments—showed zero perfectly conserved columns. Taken together, these results suggest that, for highly diverse datasets such as FliPQR homologs, both AA and 3Di-based alignments are challenged to identify strictly conserved positions, and high gap fractions are unavoidable. Nevertheless, 3Di-based alignments and reciprocal mapping approaches slightly outperformed pure AA alignments in detecting rare, fully conserved AA positions, and 3Di characters showed many more conserved positions than AAs (Figure 4a-f), highlighting 3Di utility for structural conservation analysis in highly divergent protein families. Full-resolution alignments and source data are provided at our GitHub repository (https://github.com/changhao200/Fli_PQR_research/tree/main/supplemental_figures). However, it can be seen in Figure 4 that the methods often disagree about how FliQ aligns to FliP and FliR, an issue we explored in follow-up analyses.

Figure 4:

a. FliPQR AA alignment done with Mafft

b. 3Di characters substituted for AAs in Mafft AA alignment in (a)

c. 3Di characters aligned with Famsa3di



d. AA characters substituted for the Famsa3di-aligned 3Di characters in ©

e. Foldmason alignment of AAs

f. Foldmason alignment of 3Di characters

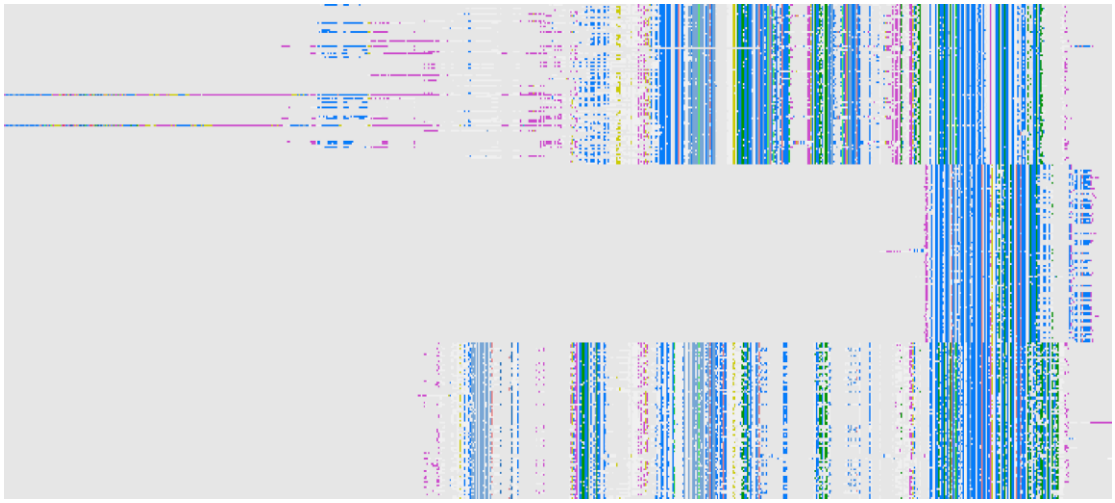


**Figure 4. Multiple sequence alignments of FliPQR core export gate proteins using amino acid and 3Di structural alphabets**. The top third of each alignment is FliP, the middle third is FliQ, and the bottom third is FliR. The colour scheme is AliView's Seaview colour scheme, but with majority-rule consensus colouring, i.e. a character is coloured only when it represents >50% of the (non-gap) data for that column. Highly conserved positions are highlighted by consistent colour blocks, while variable regions remain uncoloured. Note that the same colour scheme is used for AA and 3Di data, but the 3Di characters represent clusters of tertiary structure space, not amino acids. Panels show representative regions of alignments generated using different methods: (a) MAFFT alignment of FliPQR amino acid sequences; (b) Substitution of the 3Di structural characters into the MAFFT AA alignment; (c) FAMSA alignment of FliPQR 3Di-encoded sequences; (d) Substitution of amino acid characters into the FAMSA 3Di alignment; (e) FoldMason alignment of FliPQR amino acid sequences; (f) FoldMason alignment of FliPQR 3Di-encoded sequences. Full-resolution alignments and source data are provided at our GitHub repository (https://github.com/changhao200/Fli_PQR_research/tree/main/supplemental_figures)

### 3.2. Manual Structure-Guided Alignment Reveals Superior Performance of MAFFT for FliPQR Homologs

We attempted manual inspection of the FliPR protein structures using ChimeraX, which revealed a high degree of structural similarity between the selected homologs (Figures 5,6). Using the structural alignment in ChimeraX as a reference, the corresponding amino acid (AA) sequences were extracted for further multiple sequence alignment and method validation.

Multiple sequence alignments were performed using both MAFFT (on AA sequences) and FAMSA 3Di (on structure-encoded 3Di sequences). The results of each automated alignment were compared to the manual, structure-guided reference alignment to assess accuracy and reliability.

Among them, the AA alignment generated by MAFFT is most consistent with the manual structure alignment. Specifically, the MAFFT alignment more accurately

reflected conserved motifs and the general structural architecture observed in the ChimeraX-derived reference (see Figure 7). In contrast, alignments produced by FAMSA3Di and those based on 3Di sequence encoding exhibited greater divergence from the reference, with lower correspondence in structurally conserved regions.

Notably, conversion of the MAFFT AA alignment to 3Di sequences further increased the detection of conserved structural features, suggesting that MAFFT not only preserves sequence homology but also maintains critical aspects of structural similarity. These findings collectively indicate that, for the FliPQR dataset, MAFFT-based AA alignments provide a more reliable representation of underlying structural relationships compared to FAMSA-3Di.

Overall, this comparison highlights the importance of cross-validating automated sequence alignments with experimental structure-based data. Relying solely on structure-encoding methods such as 3Di or alignment algorithms without validation against known structures may lead to misinterpretation of conserved regions, particularly in protein families with complex evolutionary histories or low sequence identity

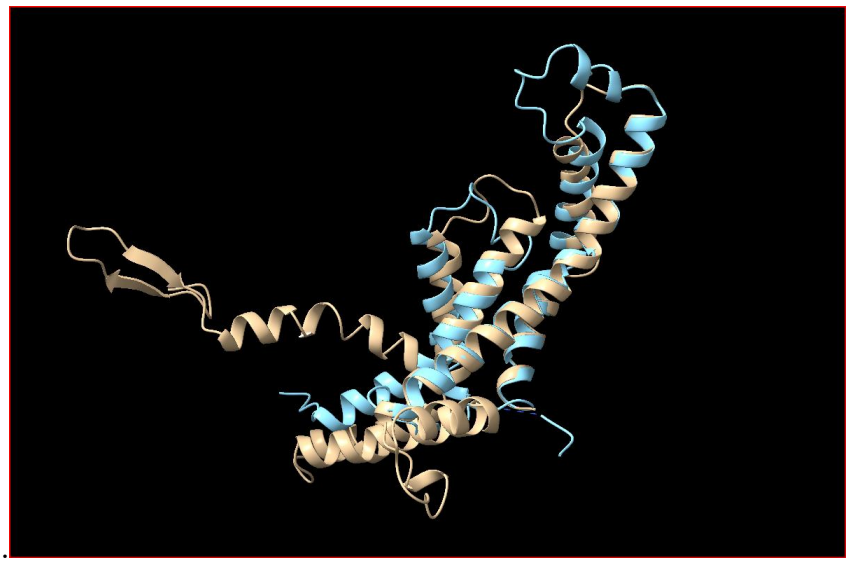


**Figure 5.** FliP and FliR superimposed in with ChimeraX's Matchmaker. Light blue is FliP, and yellow is FliR.

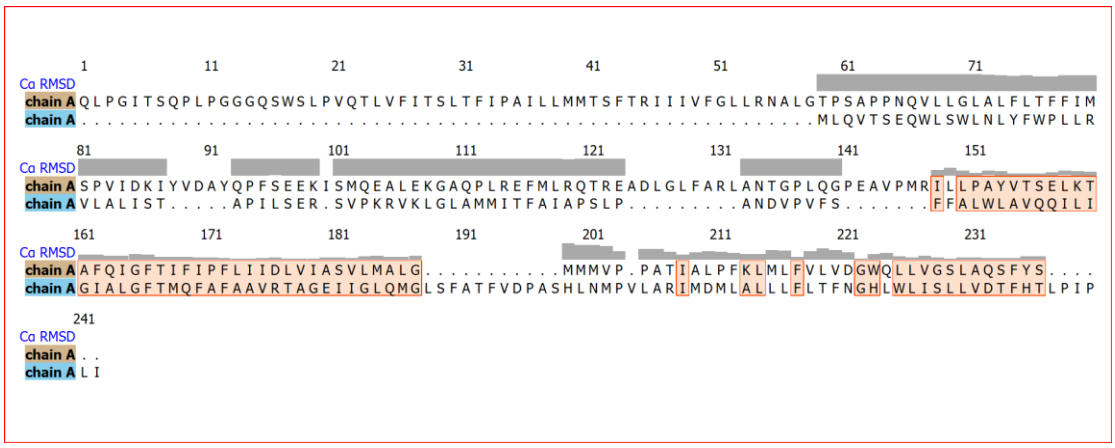
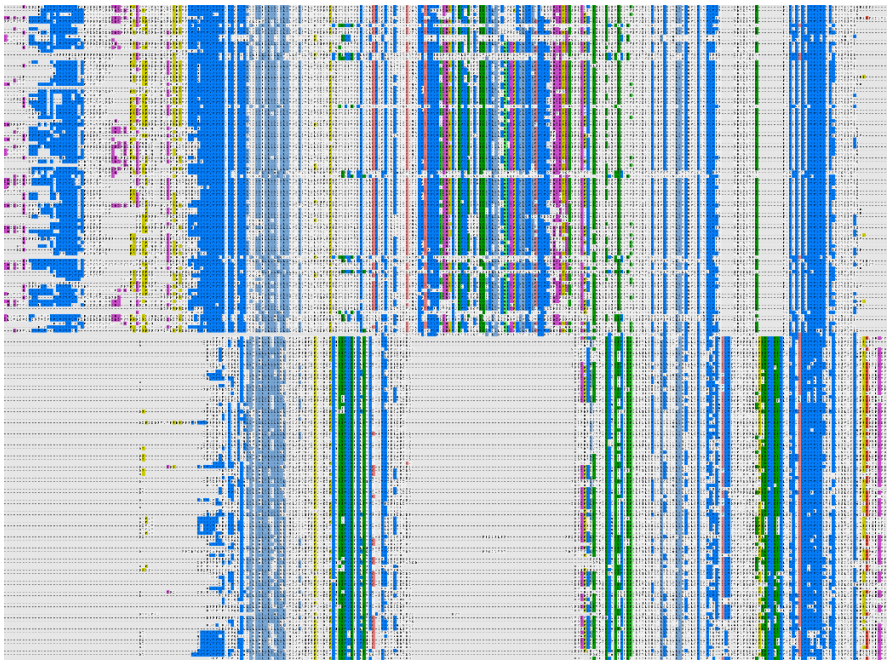



**Figure 6.** AA alignment of FliPR in ChimeraX. Cα RMSD represents the degree of structural divergence along the backbones of the superposed structures.

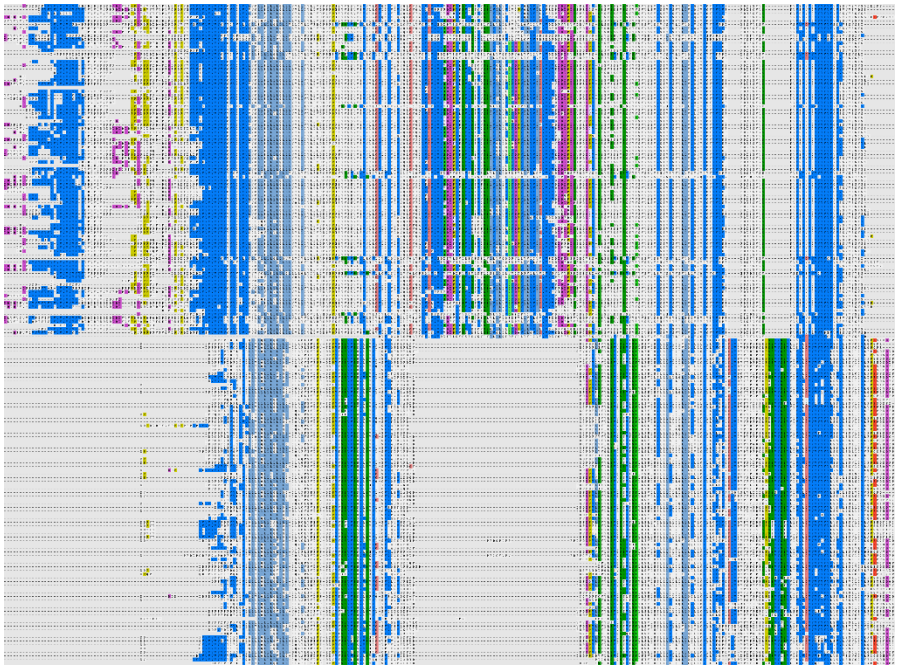
Figure 7: Manually assisted alignments of FliP and FliR

a. Mafft AA (cut off FliP signal peptide and FliR's Q region)

b. 3di match Mafft AA (cut off FliP signal peptide and FliR‘s Q region)



c. 3di match Mafft AA (cut off FliP signal peptide and FliR's Q region)



d. Famsa 3di (cut off FliP signal peptide and FliR's Q region)

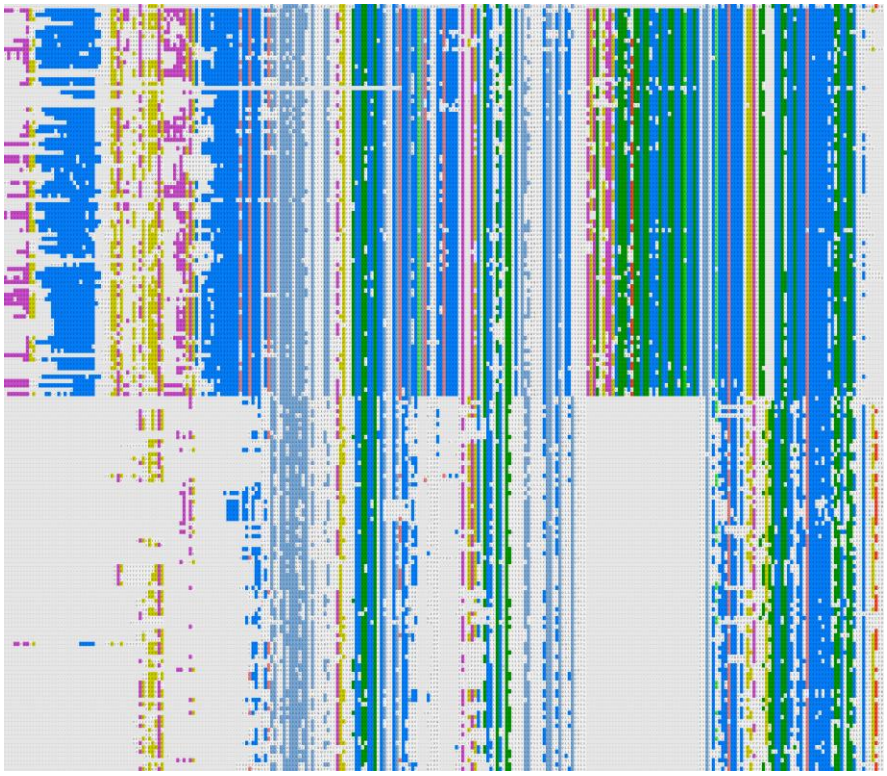


**Figure 7. Trimmed multiple sequence alignments of FliP, FliQ, and FliR proteins using amino acid and 3Di alphabets.** All alignments have the FliP signal peptide region and the FliR "FliQ region" (the C-terminus most similar to FliQ) removed to focus on the region most likely to be homologous between FliP and other proteins. Alignments were visualized in AliView with the default color scheme and "majority rule" setting, such that only columns with greater than 50% sequence conservation are colored. The darkened region in the center of the two AA MSAs highlights the alignment derived from ChimeraX-based structure-guided alignment. (a) MAFFT alignment of FliPR amino acid sequences (trimmed) (FliPR_Mafft_AA.png). (b) 3Di alphabet sequences mapped onto the MAFFT AA alignment (trimmed) (FliPR_3di_match_mafft_AA.png). (c) Amino acid sequences mapped onto the FAMSA 3Di alignment (trimmed) (FliPR_AA_match_famsa3di.png). (d) FAMSA alignment of FliPR 3Di-encoded sequences (trimmed) (FliPR_Famsa3di.png). Full-resolution alignments and additional data are provided at our GitHub repository (https://github.com/changhao200/Fli_PQR_research/tree/main/supplemental_figures).

### 3.3. Multiple Sequence Alignment Reveals Deep Conservation and Homology Across FliPQR and NF-T3SS Protein Families

The final MSA constructed for both amino acid (AA) sequences and structure-derived 3Di codes made use of the FliQ-FliQ concatenation strategy, and aligned FliQQ against FliR and then FliP. This provided the most visually convincing alignment hypothesis, as it seemed to resolve an apparent problem wherein a single FliQ had columns assigned to some of two FliQ-homologous regions found in each of FliP and FliR. This finding was surprising, as Kuhlen et al. 2018 and Beeby et al. 2020 characterized FliR as essentially FliP + FliQ. Instead it appears that the proper homology statement is approximately FliP = FliQQ = FliR. After the alignment, the first FliQ was cut as described above, to provide an alignment for use in estimating the evolutionary relationships among FliP, FliQ, FliR, and NF-T3SS homologs. Both

alignments exhibited extensive regions of columnar conservation in the 3Di characters, with clearly delineated blocks of well-aligned residues spanning the majority of sequences included in the dataset. This high degree of uniformity and regularity in the alignments attests to the success of the alignment and curation pipeline, which was specifically optimized to preserve homologous positions while minimizing the inclusion of ambiguously aligned or non-homologous columns.

MSA (415 sequences, 460 columns) of the FliPQQR dataset revealed significant differences in conservation between the two regions of the alignment. In the amino acid (AA) sequence portion (columns 1-230), only 6 consensus columns were identified, indicating very low AA sequence conservation across the dataset. In contrast, the structure-based 3di code segment (columns 231-460) had 138 consensus columns, indicating a significantly higher degree of structural conservation between these proteins (Figure 8).


**Figure 8. The final alignment used for subsequence phylogenetic analysis.**

Left: amino acids; Right: 3Di characters. From top to bottom, the 3 blocks are FliP, FliQ, and FliR.Colour Scheme: AliView's default color scheme is used, but with the "majority rule" setting, where only columns that are more than 50% conserved are coloured. Full-resolution alignments and additional data are provided at our GitHub repository (https://github.com/changhao200/Fli_PQR_research/tree/main/supplemental_figures).

The preservation of conserved regions across both data types—AA and 3Di—not only confirms the deep homology among the core components of the flagellar export apparatus and their NF-T3SS counterparts, but also demonstrates the utility of integrating sequence and structure-based information. The presence of strong

conservation in the 3Di structural characters, even in regions where AA sequence identity has fallen into the "twilight zone," provides independent structural corroboration of evolutionary relationships that might otherwise be obscured by rapid sequence divergence. This dual approach enhanced the validity of the inferred homologous groupings and provided the basis for the phylogenetic analyses performed in this study.

The overall clarity and absence of extensive gaps or misaligned regions in the final MSAs further confirm that alignment was effective. This is particularly important in light of the ancient divergence and structural variability characteristic of the FliPQR protein family and their T3SS homologs. The resulting alignments are thus well-suited for partitioned phylogenetic analyses, where both AA and 3Di data can be modeled separately to exploit their respective evolutionary signals.

### 3.4. Phylogenetic tree estimation

The results of the midpoint-rooted phylogenetic analysis (Figure 9).present a detailed hypothesis of the evolutionary relationships among the core export gate proteins of the bacterial flagellum and their NF-T3SS homologs. Three principal clades corresponding to FliP, FliQ, and FliR are well-resolved in the phylogeny, each supported by strong internal branch confidence and representing the diversity of these protein families across the sampled bacterial taxa. Within the FliP group, canonical

flagellar FliP proteins from a broad spectrum of species cluster together, while a subset of non-flagellar T3SS homologs—including SctR, EscR, and YscR—are observed as a tightly grouped branch nested within the FliP clade. This clustering indicates that these non-flagellar export apparatus proteins are most closely related to FliP and likely originated from a flagellar ancestor before acquiring their specialized roles in injectisome systems (Figure 9).

A particularly notable feature of the phylogenetic tree is the relationship between FliQ and FliR. Unlike FliP, which is separated from the other two families, the FliQ proteins do not form a completely independent clade. Instead, FliQ sequences appear as a highly supported branch within the broader FliR clade. This arrangement is consistently recovered in the analysis and reflects a particularly close evolutionary connection between these two protein families.

Within the FliQ branch, there is a distinct subgroup composed of NF-T3SS proteins, including SctS, YscS, EscS, and HrcS. These proteins are unified in the tree as a derived cluster within FliQ and show high sequence and structural similarity to their flagellar homologs, further emphasizing their close evolutionary link. Similarly, the FliR clade contains not only canonical flagellar FliR proteins but also a derived cluster of non-flagellar homologs—SctT, Esct, Ysct, and Hrct—grouped as a well-supported branch within the FliR family. The clear pattern across the entire tree is that NF-T3SS homologs do not form separate, basal clades; rather, they are consistently positioned as internal, lineage-specific subgroups within the larger flagellar protein families.

The overall topology indicates that the diversification of T3SS export gate proteins in non-flagellar systems was shaped by the recruitment and subsequent adaptation of

flagellar protein families, rather than by parallel, independent origins. The nesting of FliQ within FliR, along with the internal positioning of each NF-T3SS cluster, underscores the modular and stepwise nature of T3SS evolution. These results provide strong support for a model in which the specialized components of the injectisome export gate originated from flagellar ancestors (Abby and Rocha, 2012) and diversified in the context of the broader evolution of bacterial motility and virulence systems. The findings also align with and extend previous structural and phylogenetic research, offering a clearer resolution of the relationships among these essential molecular machines.

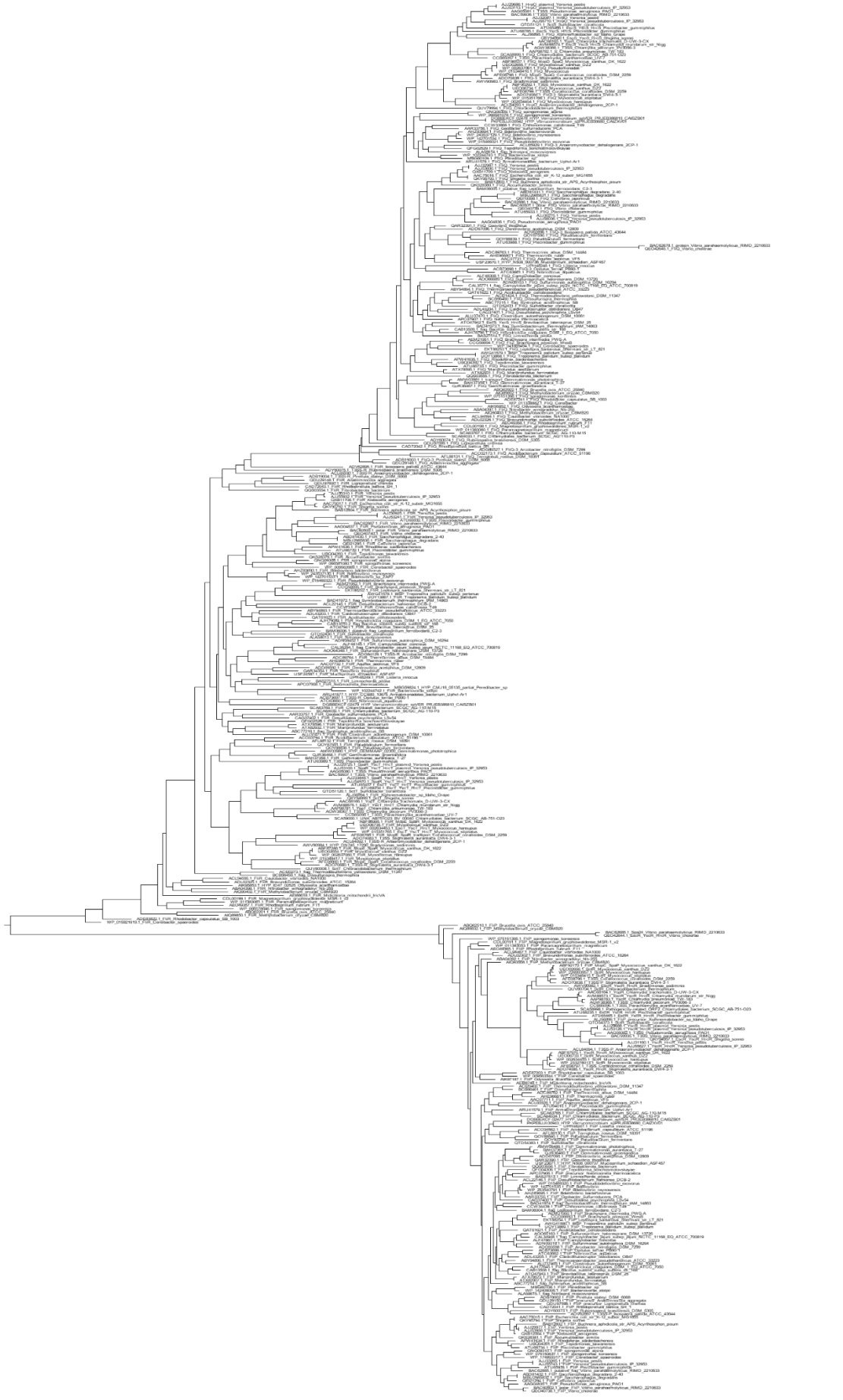


**Figure 9. Maximum likelihood phylogenetic tree of FliP, FliQ, and FliR core export gate proteins.**

The tree was reconstructed using IQ-TREE based on a partitioned alignment of amino acid and 3Di sequences. Node support values are shown as ultrafast bootstrap (UFBoot) percentages. The tree is midpoint-rooted for display.

All major clades corresponding to FliP, FliQ, and FliR are labeled. The alignment and full-resolution tree files are provided as Supplemental Data and are available at our GitHub repository (https://github.com/changhao200/Fli_PQR_research/tree/main/supplemental_figures).

### 3.5. FliQQ Homology and Alignment Patterns with FliP and FliR

In the process of MSA and phylogenetic analysis of the flagellar export gate proteins, a striking observation emerged when comparing FliQ and FliR sequences. Initial alignments using standard methods such as MAFFT, as well as structure-guided approaches, frequently resulted in apparent misalignments between FliQ and FliR. Upon closer inspection, including manual curation and visualization in AliView and ChimeraX, it became evident that these misalignments were not solely due to methodological artifacts. Rather, the pattern suggested the presence of multiple homologous regions shared between FliQ and FliR, indicative of a more complex evolutionary relationship than can be captured by conventional pairwise alignment.

To rigorously explore this relationship, we constructed a specialized MSA that concatenated two FliQ protein sequences and aligned this composite with FliP and FliR proteins. The resulting alignment, visualized in AliView (Figure 10). The resulting shows clear blocks of similarity between specific regions of FliQQ and

discrete segments within FliR. This finding indicates that FliR harbors at least two domains that are homologous to the repeated regions in FliQ, suggesting a history of internal duplication or domain shuffling during the evolution of FliR from a FliQ-like ancestor.

**Figure 10. Multiple sequence alignment of FliP (top), FliQ-FliQ (middle), and FliR (bottom) proteins. Left: AAs. Right: 3Dis for the same residue positions.**

The alignment shows both amino acid and 3Di-encoded partitions for all sequences, with regions trimmed using a 35% gap threshold. The FliP signal peptide and FliR Q region have been removed to

focus on conserved structural cores. Alignments were visualized using AliView with the default color scheme and "majority rule" setting, such that only columns with greater than 50% conservation are colored. Full alignment files and partition metadata are available as Supplemental Data and via our GitHub repository (https://github.com/changhao200/Fli_PQR_research/tree/main/supplemental_figures).

In addition, the alignment highlighted that the concatenated FliQQ sequence also shares regions of significant homology with certain conserved motifs in FliP. This broader pattern of modular similarity points to a deeper evolutionary relationship among all three proteins, consistent with the notion that they arose through ancient duplication and recombination events.

Importantly, the use of concatenated FliQ sequences proved critical for revealing these relationships. Standard single-sequence alignments often resulted in a highly-gapped FliQ block, as different parts of FliQ were spread across the repeated homologous regions in FliR. The concatenation approach enabled the identification of distinct, well-aligned blocks, providing visual and quantitative support for the hypothesis that FliR is structurally modular.

### 3.6 Structural Alignment of FliQ and FlhB Reveals Alternative and Modular Homologous Regions

During the research, we discovered that there may be potential homology between another bacterial flagellum protein, FlhB, and FliQ, a possibility noted by Kuhlen et al. (2018). As FlhB integrates into the base of the FliPQR structure (Kuhlen et al., 2018) as part of its function of regulating the opening of the export gate, it may be

plausible that it shares some structural elements, To further investigate the relationship between flagellar export apparatus proteins, we performed structure-based alignments between FliQ and FlhB using multiple approaches, including US-align, ChimeraX MatchMaker, and FoldMason. Each of these tools is validated for global and local protein structure comparison, but as shown in Figure 11, the resulting superpositions exhibited distinct differences in the regions identified as most closely aligned.

For US-align, the algorithm generated a structural alignment in which specific α-helical bundles of FliQ and FlhB were closely superimposed, suggesting one set of putative homologous domains. In contrast, ChimeraX's MatchMaker alignment resulted in a different overlap, with other helices and segments identified as the primary homologous regions. FoldMason, employing both visual superposition and a 3Di structural alphabet encoding, revealed yet another configuration: its structure-based alignment (blue/yellow highlighted a partially distinct set of helices, and its 3Di alignment further supported the presence of recurring structural motifs across multiple segments of the proteins.

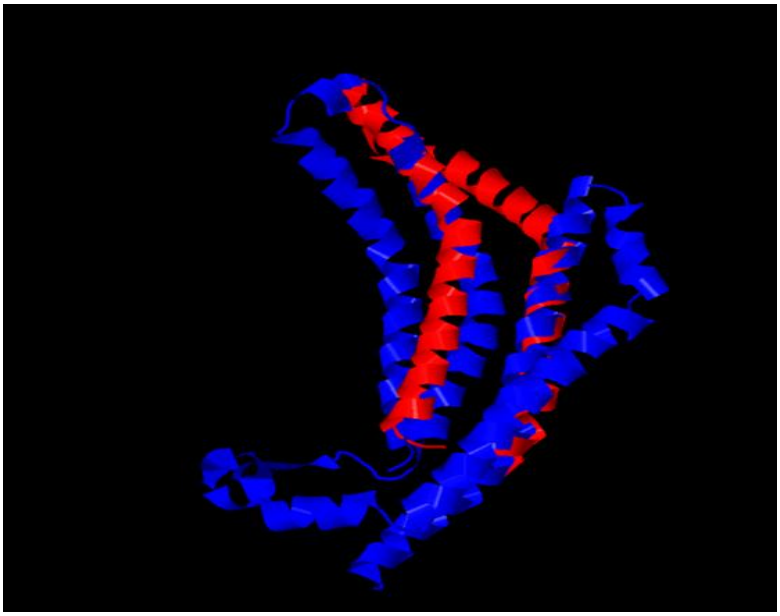
These discrepancies can be attributed to both methodological differences and the underlying modularity of the proteins. US-align and ChimeraX utilize different optimization criteria-US-align maximizes global structural similarity using TM-score and local distance differences (Zhang & Skolnick, 2005; Zhang et al., 2022), while ChimeraX considers secondary structure and sequence similarity (Goddard et al., 2018). FoldMason, on the other hand, uses a structure-based alphabet to capture local backbone conformations and is particularly sensitive to repeated or modular structural features (Gilchrist et al., 2024).

Importantly, the FoldMason structural alignment provides a particularly clear and intuitive visualization of the spatial relationship between FliQ and FlhB, making it easier to observe potential regions of homology by eye. Compared to US-align and ChimeraX, the FoldMason image highlights the helical overlap and modular correspondence in a way that is immediately accessible for visual assessment. Furthermore, if the multiple sequence alignment (MSA) analysis also demonstrates conserved or homologous segments between FliQ and FlhB, this would further support the interpretation drawn from the structure-based alignments.

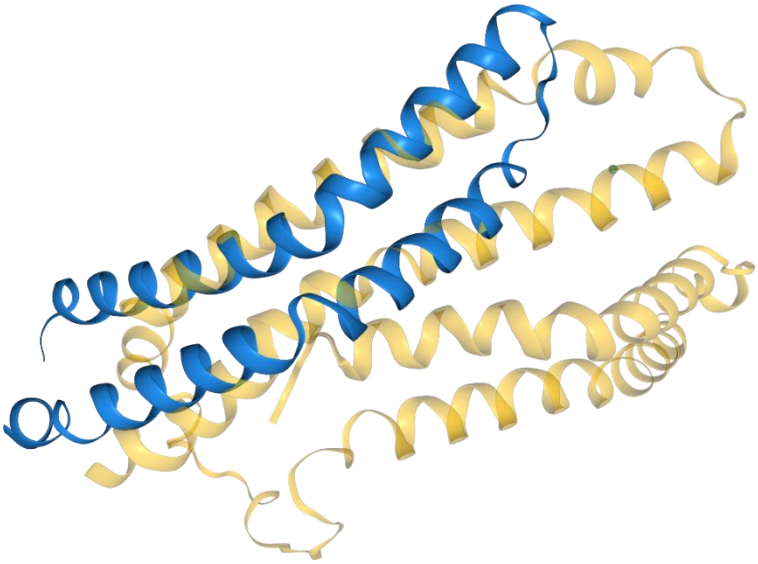
In summary, the combination of different structural alignment approaches, especially FoldMason images, enables a practical and straightforward assessment of potential homology between FliQ and FlhB.

a.Chimerax

b. Usalign



c. Foldmason



d. Foldmason 3Di alignment between FliQ and FlhB

**Figure 11: Schematic representation and 3di align of the FliQ and FlhB**

**structures shown in Chimerax, USalign, and Foldmason.**

# 4.Discussion

## 4.1. Evolutionary Insights from Modular Homology

This study investigated the evolutionary relationships and modular homology among core flagellar export gate proteins—FliP, FliQ, and FliR—and their NF-T3SS homologs. Through an integrated combination of sequence-based and structure-based MSA, alongside advanced phylogenetic analyses, we identified complex patterns of homology, duplication, and divergence that have shaped the architecture of the bacterial Type III export apparatus.

A key finding is the modular homology observed between FliQ and FliR, particularly evident when two FliQ sequences are concatenated and compared to FliR. This modularity supports the hypothesis that FliR evolved through internal duplication or domain shuffling involving FliQ-like ancestors, an insight reinforced by visual and structural alignment using AliView, US-align, and ChimeraX. Furthermore, the detection of significant similarity between FliQQ and FliP points to even deeper evolutionary connections among the three proteins, consistent with a model of repeated duplication and recombination events during the evolution of the T3SS core.

## 4.2. Phylogenetic Tree Topology and Rooting

Our midpoint-rooted phylogenetic trees provided robust support for three principal clades corresponding to FliP, FliQ, and FliR. The placement of FliQ as a branch

within the FliR clade is particularly notable, echoing the modular relationships observed in sequence alignments. NF-T3SS proteins—including SctR, SctS, SctT, and their homologs—formed derived subclades within each family, underscoring the evolutionary trajectory from canonical flagellar export proteins to their virulence-associated derivatives. This is consistent with previous comparative genomic and structural analyses that have posited the evolution of injectisome T3SS systems from ancestral flagellar machineries (Abby & Rocha, 2012; Blocker et al., 2003).

A persistent challenge in such analyses is the rooting of deep phylogenies. While we employed midpoint rooting for interpretability, the true evolutionary root remains tentative due to the absence of suitable outgroups and the inherent limitations of rooting ancient, highly diverged protein families (Philippe et al., 2011; Puente-Lelievre et al., 2025). Nonetheless, the observed topology, with the flagellar proteins FliP, FliQ, and FliR occupying basal positions and NF-T3SS homologs as derived branches, is congruent with both functional specialization and previous structural studies.

### 4.3. Benefits and Limitations of Sequence and Structure-Based Analyses

Comparing sequence-based (MAFFT) and structure-based (FAMSA3di, FoldMason) alignments proved useful in helping to judge the utility of various alignments produced by these programs, especially because the FliPQR divergence lays in the "twilight zone" where sequence similarity alone is insufficient for reliable inference (Rost, 1999; Holm, 2022). Structure-based approaches, particularly those leveraging the 3Di alphabet and advanced structural comparison tools, enabled the detection of homologous domains even in the face of substantial sequence divergence. However, our findings also highlight that different alignment tools—such as US-align and

ChimeraX—can yield alternative plausible matches between modular protein regions, emphasizing the importance of comparing multiple, independent methods to help assess quality (van Kempen et al., 2023; Gilchrist et al., 2024).

The choice and quality of MSA remain central to accurate phylogenetic reconstruction. In this work, careful quality control—including the use of the TrimAl tool to remove poorly conserved regions—was important to emphasize the phylogenetic signal and mitigate the confounding effects of misalignments or non-homologous regions (Capella-Gutiérrez et al., 2009; Tan et al., 2015). Visual validation in AliView and benchmarking against structural alignments from ChimeraX further increased confidence in the reliability of our final alignment.

# 5. Limitations

## 5.1. Limitations in MSA and Homology Detection

A significant portion of this study's approach relies on the accuracy of multiple sequence alignments and structural homology assignments. But there are still some limitations. First, the dataset was primarily built from publicly available protein databases. Despite stringent curation and the removal of clear redundancies, a small proportion of sequences may still contain errors such as incomplete or repeated segments, and misannotations stemming from inconsistent or outdated database entries. Such artifacts can introduce substantial noise into MSAs, affecting alignment quality and confounding downstream analyses.

Second, the application of structure-based alignments, particularly those derived from

Foldseek-generated 3Di sequences, while powerful for revealing remote homology, is inherently limited by the accuracy of the underlying protein structure models. These structures are often computational predictions (e.g., from AlphaFold2), and may not always reflect the true conformational state of the protein. Misassigned structural features or errors in the 3Di alphabet could result in the misidentification of homologous regions or the loss of genuine evolutionary signals.

Third, the modular and repetitive architectures of FliP, FliQ, and FliR complicate the detection of true homologous relationships. Both sequence and structure aligners may match only portions of these proteins or confuse paralogous and convergent features with genuine homology. This is especially problematic in families with internal duplications or rearrangements, potentially leading to ambiguous or conflicting alignment results.

Fourth, the computational algorithms and models employed for both sequence and structure alignment carry inherent biases. The choice of alignment tool, scoring matrix, and algorithmic parameters can influence which regions are deemed homologous and, ultimately, the inferred relationships among protein families.

Finally, to reduce these errors, we chose to manually check and verify the alignments using visualization tools such as ChimeraX and quality control software such as TrimAl, but some issues (such as fragmented alignments, ambiguous boundaries, or rare misannotations) may still exist.

### 5.2. Limitations in Phylogenetic Tree Reconstruction and Interpretation

Parallel to the challenges in alignment, several important limitations were encountered in phylogenetic tree construction and interpretation:

First, the phylogenetic trees generated in this study using IQ-TREE are inherently unrooted. While midpoint rooting was applied to facilitate biological interpretation and to infer likely evolutionary origins, this method is not definitive. The position of the root remains tentative due to possible rate differences between lineages, which may affect the inference of ancestral nodes and evolutionary directions.

Second, the complexity of FliP, FliQ, and FliR—marked by internal repeats and duplications—creates additional difficulty in distinguishing between orthology and paralogy. Consequently, some tree branches may represent mixtures of orthologous and paralogous relationships, complicating evolutionary interpretations and ancestral state reconstructions.

Third, although we carefully selected multiple available models to achieve the best fit, no single model can perfectly explain the heterogeneity and complexity present in the FliPQR protein family. Therefore, model specification errors may affect the tree topology and branching.

Fourth, horizontal gene transfer (HGT) is a well-documented phenomenon in T3SS genes. Undetected HGT events may interfere with the reconciliation of gene trees with species trees, resulting in gene trees that do not accurately reflect the evolutionary history of organisms. However, FliPQR are typically arranged tandemly in genomes (they originally given these names when labeled alphabetically in the order in *E. coli*), and are co-expressed, so are probably less likely to have independent histories than most proteins.

Fifth, protein naming errors and annotation errors in source databases sometimes lead to unexpected or mixed evolutionary branch structures among proteins claimed in the Genbank database to be FliP, FliQ, FliR, or their respective nonflagellar homolog proteins. In addition, it is important to remember that while homology and structural similarity strongly suggest evolutionary relationships, they do not necessarily confirm functional equivalence.

Finally, these limitations highlight the inherent complexity and ongoing development of molecular evolution studies. Continued progress in structure prediction, broader taxonomic sampling, more detailed functional analysis, and improved phylogenetic methods are essential to resolve these uncertainties. Such advances will further illuminate our understanding of the evolutionary dynamics of bacterial flagella and T3SS.

Taken together, these limitations highlight the complexity and ongoing nature of molecular evolutionary analyses. Continued progress in structure prediction, taxonomic sampling, functional characterization, and phylogenetic methods will be

critical to resolving remaining uncertainties and further elucidating the evolutionary dynamics of bacterial flagella and T3SS systems.

## 6. Summary

This study advances our understanding of the evolutionary dynamics underlying the core export gate of the bacterial flagellum and its NF-T3SS relatives. Through rigorous MSA, structural comparison, and phylogenetic inference, we demonstrate that FliP, FliQ, and FliR form a deeply interconnected set of proteins, shaped by modular evolution, gene duplication, and domain recombination. The congruence of sequence and structure-based evidence with phylogenetic tree topology supports the view that the diversity of the T3SS export apparatus is rooted in the evolutionary plasticity of these ancestral flagellar proteins.

Our integrative methodology not only clarifies the relationships among FliP, FliQ, and FliR but also provides a robust framework for studying other ancient, modular protein complexes. Continued advances in structure prediction, sequence alignment algorithms, and the expansion of high-quality reference databases will further enhance our ability to resolve the deepest branches of the tree of life and understand the molecular mechanisms that drive the evolution of complex cellular machines.

## 7. Reference List

Abby, S. S., & Rocha, E. P. C. (2012). The non-flagellar type III secretion system evolved from the bacterial flagellum and diversified into host-cell adapted systems.

*PLoS Genetics, 8*(9), e1002983. https://doi.org/10.1371/journal.pgen.1002983

Abrusci, P., Vergara-Irigaray, M., Johnson, S., Beeby, M. D., Hendrixson, D. R., Roversi, P., Friede, M. E., Deane, J. E., Jensen, G. J., Tang, C. M., & Lea, S. M. (2012). Architecture of the major component of the type III secretion system export apparatus. *Nature Structural & Molecular Biology*, *20*(1), 99–104. https://doi.org/10.1038/nsmb.2452

Almagro Armenteros, J. J., Tsirigos, K. D., Sønderby, C. K., Petersen, T. N., Winther, O., Brunak, S., ... & Nielsen, H. (2019). SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nature Biotechnology, 37*(4), 420–423. https://doi.org/10.1038/s41587-019-0036-z

Blocker, A., Komoriya, K., & Aizawa, S.-I. . (2003). Type III secretion systems and bacterial flagella: Insights into their function from structural similarities. *Proceedings of the National Academy of Sciences*, *100*(6), 3027–3030. https://doi.org/10.1073/pnas.0535335100

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

Chen, S., Beeby, M., Murphy, G. E., Leadbetter, J. R., Hendrixson, D. R., Briegel, A., ... & Jensen, G. J. (2011). Structural diversity of bacterial flagellar motors. *EMBO Journal, 30*(14), 2972–2981. https://doi.org/10.1038/emboj.2011.186

Chevance, F. F. V., & Hughes, K. T. (2008). Coordinating assembly of a bacterial macromolecular machine. *Nature Reviews Microbiology, 6*(6), 455–465. https://doi.org/10.1038/nrmicro1887

Chung, S. Y., & Subbiah, S. (1996). A structural explanation for the twilight zone of protein sequence homology. *Structure*, *4*(10), 1123–1127. https://doi.org/10.1016/S0969-2126(96)00119-0Cornelis, G. R. (2006). The type III secretion injectisome. *Nature Reviews Microbiology, 4*(11), 811–825. https://doi.org/10.1038/nrmicro1526

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

Diepold, A., & Armitage, J. P. (2015). Type III secretion systems: The bacterial flagellum and the injectisome. *Philosophical Transactions of the Royal Society B: Biological Sciences, 370*(1679), 20150020. https://doi.org/10.1098/rstb.2015.0020

Eddy, S. R. (2011). Accelerated profile HMM searches. *PLoS Computational Biology, 7*(10), e1002195. https://doi.org/10.1371/journal.pcbi.1002195

Eisen, J. A. (1998). Phylogenomics: Improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Research, 8*(3), 163–167. https://doi.org/10.1101/gr.8.3.163

Erhardt, M., Namba, K., & Hughes, K. T. (2010). Bacterial nanomachines: The flagellum and type III injectisome. *Cold Spring Harbor Perspectives in Biology, 2*(11), a000299. https://doi.org/10.1101/cshperspect.a000299

Penny, D. (2004). Inferring Phylogenies.—Joseph Felsenstein. 2003. Sinauer Associates, Sunderland, Massachusetts. *Systematic Biology*, *53*(4), 669–670. https://doi.org/10.1080/10635150490468530

Gabaldón, T., & Koonin, E. V. (2013). Functional and evolutionary implications of

gene orthology. *Nature Reviews Genetics, 14*(5), 360–366. https://doi.org/10.1038/nrg3456

Galán, J. E., Lara-Tejero, M., Marlovits, T. C., & Wagner, S. (2014). Bacterial type III secretion systems: Specialized nanomachines for protein delivery into target cells. *Annual Review of Microbiology, 68*, 415–438. https://doi.org/10.1146/annurev-micro-092412-155725

Gilchrist, C. L. M., Puente-Lelievre, C., Allsopp, L. P., ... & Matzke, N. J. (2024). FoldMason: Ultrafast, scalable multiple structure alignment using structure-based alphabets. *bioRxiv*. https://doi.org/10.1101/2023.12.12.571181

Holm, L. (2019). DALI and the persistence of protein shape. *Protein Science*, *29*(1), 128–140. https://doi.org/10.1002/pro.3749

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

Jumper, J., Evans, R., Pritzel, A., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *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. *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. https://doi.org/10.1093/nar/gkf436

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution, 30*(4), 772–780. https://doi.org/10.1093/molbev/mst010

Kojima, S., & Blair, D. F. (2004). The bacterial flagellar motor: structure and function of a complex molecular machine. *International Review of Cytology*, *233*, 93–134. https://doi.org/10.1016/S0074-7696(04)33003-2

Kuhlen, L., Abrusci, P., Johnson, S., Gault, J., Deme, J., Caesar, J. J. E., ... & Lea, S. M. (2018). Structure of the core of the type III secretion system export apparatus. *Nature Structural & Molecular Biology, 25*(7), 583–590. https://doi.org/10.1038/s41594-018-0086-9

Larsson, A. (2014). AliView: A fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics, 30*(22), 3276–3278. https://doi.org/10.1093/bioinformatics/btu531

Li, W., & Godzik, A. (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics, 22*(13), 1658–1659. https://doi.org/10.1093/bioinformatics/btl158

Macnab, R. M. (2003). How bacteria assemble flagella. *Annual Review of Microbiology, 57*, 77–100. https://doi.org/10.1146/annurev.micro.57.030502.090832

Mariani, V., Biasini, M., Barbato, A., & Schwede, T. (2013). lDDT: A local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics, 29*(21), 2722–2728. https://doi.org/10.1093/bioinformatics/btt473

Matzke, N. J., Puente-Lelievre, C., & Baker, M. A. B. (2025, May 9). A Pipeline for

Generating Datasets of 3-Dimensional Tertiary Interaction Characters for Model-Based Structural Phylogenetics. https://doi.org/10.31219/osf.io/5uhkx_v1

Minamino, T., & Imada, K. (2015). The bacterial flagellar motor and its structural diversity. *Trends in Microbiology*, *23*(5), 267–274. https://doi.org/10.1016/j.tim.2014.12.011

Minamino, T., Kinoshita, M., & Namba, K. (2019). Directional Switching Mechanism of the Bacterial Flagellar Motor. *Computational and Structural Biotechnology Journal*, *17*, 1075–1081. https://doi.org/10.1016/j.csbj.2019.07.020

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

Morimoto, Y. V., & Minamino, T. (2014). Structure and function of the bi-directional bacterial flagellar motor. *Biomolecules, 4*(1), 217–234. https://doi.org/10.3390/biom4010217

Nixon, K. C., & Carpenter, J. M. (2011). On homology. *Cladistics*, *28*(2), 160–169. https://doi.org/10.1111/j.1096-0031.2011.00371.x

Orengo, C. A., & Thornton, J. M. (2005). Protein families and their evolution—a structural perspective. *Annual Review of Biochemistry, 74*, 867–900. https://doi.org/10.1146/annurev.biochem.74.082803.133029

Pettersen, E. F., Goddard, T. D., Huang, C. C., Meng, E. C., Couch, G. S., Croll, T. I., ... & Ferrin, T. E. (2021). UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Science, 30*(1), 70–82.

https://doi.org/10.1002/pro.3943

Philip, J. S., Grewal, S., Scadden, J., Puente-Lelievre, C., Matzke, N. J., McNally, L., & Baker, M. A. B. (2025). Mapping the loss of flagellar motility across the tree of life. *The ISME Journal*. https://doi.org/10.1093/ismejo/wraf111

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. *PLoS Biology, 9*(3), e1000602. https://doi.org/10.1371/journal.pbio.1000602

Puente-Lelievre, Caroline; Ridone, Pietro; Douglas, Jordan; Amritkar, Kaustubh; Kaçar, Betül; Baker, Matthew; Matzke, Nicholas J. (2025). Molecular and structural innovations of the stator motor complex at the dawn of flagellar motility. *bioRxiv*, 604496. https://www.biorxiv.org/content/10.1101/2024.07.22.604496v1.abstract

Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Engineering, 12*(2), 85–94. https://doi.org/10.1093/protein/12.2.85

Samatey, F. A., Imada, K., Nagashima, S., Vonderviszt, F., Kumasaka, T., Yamamoto, M., & Namba, K. (2001). Structure of the bacterial flagellar protofilament and implications for a switch for supercoiling. *Nature, 410*(6826), 331–337. https://doi.org/10.1038/35066504

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

Tan, G., Matthieu Muffato, Ledergerber, C., Herrero, J., Goldman, N., Alonso, M., & Christophe Dessimoz. (2015). Current Methods for Automated Filtering of Multiple

Sequence Alignments Frequently Worsen Single-Gene Phylogenetic Inference. *Systematic Biology*, *64*(5), 778–791. https://doi.org/10.1093/sysbio/syv033

Terashima, H., Kojima, S., & Homma, M. (2008). Flagellar motility in bacteria: Structure and function of flagellar motor. *International Review of Cell and Molecular Biology, 270*, 39–85. https://doi.org/10.1016/S1937-6448(08)01402-0

Tohru Minamino, & Kinoshita, M. (2023). Structure, Assembly, and Function of Flagella Responsible for Bacterial Locomotion. *Ecosal Plus*, *11*(1). https://doi.org/10.1128/ecosalplus.esp-0011-2023

Turner, L., Ryu, W. S., & Berg, H. C. (2000). Real-time imaging of fluorescent flagellar filaments. *Journal of Bacteriology, 182*(10), 2793–2801. https://doi.org/10.1128/JB.182.10.2793-2801.2000

van Kempen, M., Kim, S., Tumescheit, C., Mirdita, M., Gilchrist, C. L. M., Söding, J., & Steinegger, M. (2023). Foldseek: Fast and accurate protein structure search. *Nature Biotechnology, 41*(5), 818–823. https://doi.org/10.1038/s41587-023-01773-0

Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., ... & Velankar, S. (2022). AlphaFold Protein Structure Database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research, 50*(D1), D439–D444. https://doi.org/10.1093/nar/gkab1061

Wadhams, G. H., & Armitage, J. P. (2004). Making sense of it all: Bacterial chemotaxis. *Nature Reviews Molecular Cell Biology, 5*(12), 1024–1037. https://doi.org/10.1038/nrm1524

Zhang, Y., & Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research, 33*(7), 2302–2309.

https://doi.org/10.1093/nar/gki524

Zhang, Y., Holm, L., Steinegger, M., & Söding, J. (2022). Structural alignment in the era of machine learning: challenges and solutions. *Nature Reviews Molecular Cell Biology, 23*(5), 319–332. https://doi.org/10.1038/s41580-021-00444-6