

Nicholas J. Matzke^{1,2}, Caroline Puenta-Lelievre^{1,2}, Pietro Ridone³, Jordan Douglas^{2,4,5}, Kaustubh Amritkar⁶, Betül Kaçar⁶, Matthew Baker³, Ashar J. Malik^{7,8}, David Ascher^{7,8}, Jane Allison¹, Anthony Poole^{1,2}, Daniel Lundin⁹, Matthew Fullmer^{1,2}, Remco Bouckert^{2,10}, Hyunbin Kim¹¹, Martin Steinegger^{11,12,13}, Piper Craven¹⁴, Jiahe Zhang¹⁴, Changhao Li¹⁴, Arya Kaul¹⁴, Masafumi Obara¹⁴, Yogapriya Subramanian¹⁴

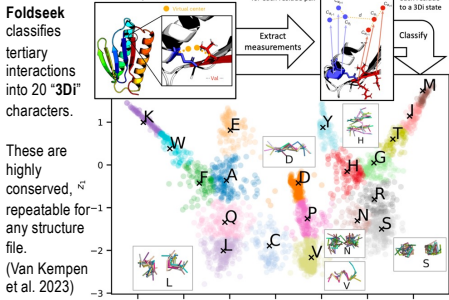
¹School of Biological Sciences, University of Auckland
²Centre for Computational Evolution, University of Auckland
³School of Biotech. & Biomolecular Sciences, Univ. New South Wales
⁴Department of Physics, University of Auckland
⁵Division of Ecology & Evolution, Research School of Biology, ANU

⁶Department of Bacteriology, University of Wisconsin-Madison, Madison WI 53715, USA
⁷School of Chemistry and Molecular Biosciences, The University of Queensland
⁸Computational Biology and Clinical Informatics, Baker Heart & Diabetes Institute, Melbourne
⁹Department of Biochemistry and Biophysics, Stockholm University, Sweden
¹⁰School of Computer Science, University of Auckland

¹¹School of Biological Sciences, Seoul National University, South Korea
¹²Artificial Intelligence Institute, Seoul National University
¹³Institute of Molecular Biology and Genetics, Seoul National University
¹⁴Post-graduate student researchers

Introduction: Structural Phylogenetics

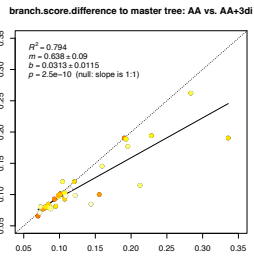
Protein structure is more conserved than amino acid (AA) sequence, but has been hard to include in model-based statistical phylogenetics. This changed with Puenta-Lelievre et al. (2023), which proposed and tested the use of Foldseek's "3Di" (3-dimensional) structural alphabet.



Data	Model	lnL	ΔlnL	Δ	BIC
EX+FU+G	-16315.2	35.9	106		33187.1
Blossum62+G	-16279.2	0.0	123		33204.5
LG+F+G	-16289.1	9.8	123		33224.2
EX+FU+G	-16299.1	19.8	125		33254.6
Poisson+F+G	-16888.9	609.7	123		34423.9
3Di+F+G	-17950.8	1671.5	123		36547.6
3Di+F+G	-8054.9	0.0	123		16755.9
Poisson+F+G	-8329.4	274.5	123		17304.9
Blossum62+F+G	-8485.0	430.0	123		17516.0
EX+FU+G	-8620.9	566.0	125		17898.3
LG+F+G	-8734.1	679.1	123		18114.1
EX+FU+G	-10415.9	2361.0	106		21388.6

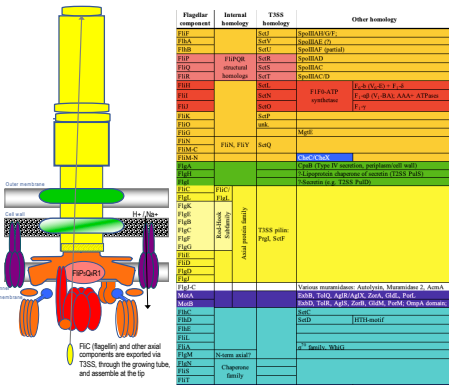
Ground truth: Kannan et al. (2014) phylogeny of 56 mitochondrial proteins present in the Last Eukaryote Common Ancestor.

We compared distances of single-protein AA trees (x-axis) and AA+3Di trees (y-axis) to the ground truth. **Result:** improvement for shorter (yellow) & less conserved (right) proteins.



Introduction: Bacterial Flagellum

The flagellum is an ion-powered rotating nanomachine. It is made of ~30 core proteins that cooperate to self-assemble and function. To study its evolution, we first assemble non-flagellar homologs:



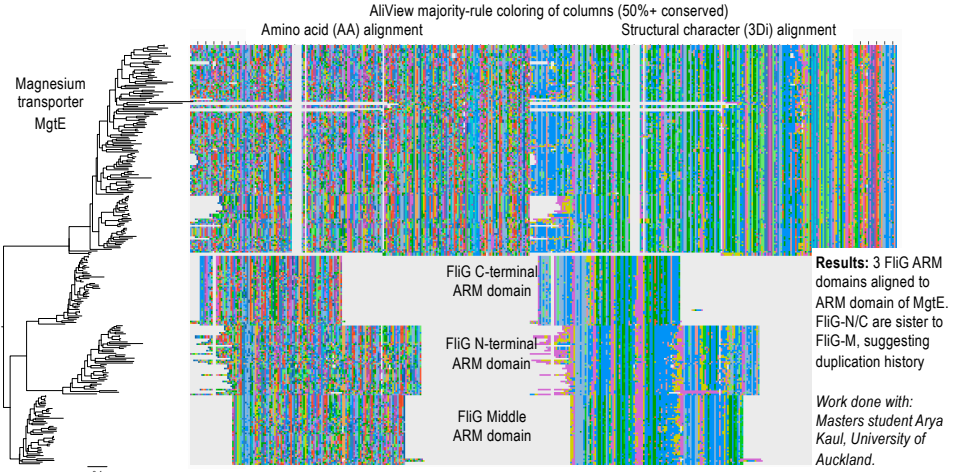
Methods

Nonflagellar homologs of flagellum proteins are so diverged they are in the Twilight Zone or Midnight Zone (weak/no AA similarity). To attempt phylogenies, we followed Matzke et al. (2025):

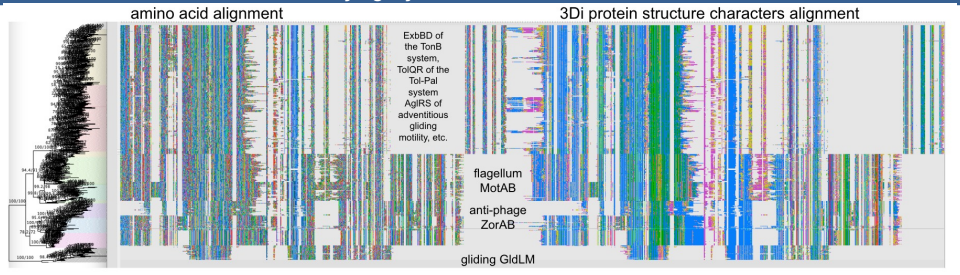
- JackHMMER-sampled known homolog families from ~200 genomes sample to represent major bacterial phyla. Manually filtered out unconvincing hits.
- Downloaded AlphaFold monomer structures (Varadi et al. 2022) for all proteins.
- Converted to 3Di with Foldseek (Van Kempen et al. 2024), aligned with FAMS3Di, checked, and assembled partitioned alignments using R pipeline (Matzke et al. 2025).
- IQ-TREE with model selection on AA/3Di partitions, as in Puenta-Lelievre et al. (2024).

These are the **first-ever phylogenies** of Flg-MgtE, of MotAB-multiple outgroups, and of FlpQR. They were used to test position of nonflagellar-T3SS, and history of internal duplications.

Results: Flg-MgtE

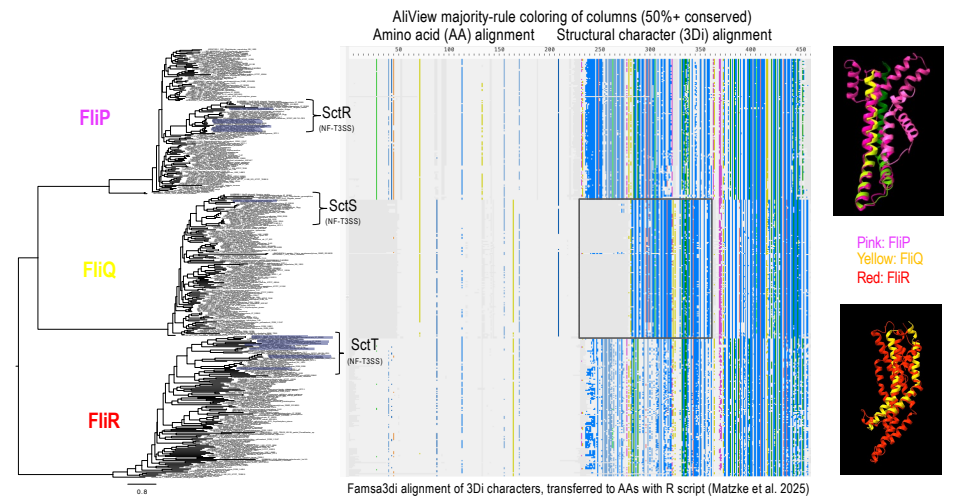


Results: AA+3Di Phylogeny: MotAB and ExbBD/ToIQR, ZorAB, GldLM



Results: FlpQR

The Flp₂Q₁R₁ complex forms a export gate at the base of the axial tube, and cryo-EM structures revealed they are structural homologs (Kuhlen et al. 2020; right). However, AA sequence similarity is in the Midnight Zone. Result: FlpR clearly has 2 FlpQ regions. Rooting the duplicates against each other, nonflagellar T3SS homologs SctRST are clearly derived, confirming Abby & Rocha 2012. Work done with: Masters student Changhao Li, U. Auckland.



Conclusions

- Structural phylogenetics enables us to erect and test phylogenetic hypotheses for protein divergences into and beyond the Twilight Zone.
- For Flg, the "cog teeth" that allow MotAB to rotate the flagellum both clockwise and counterclockwise, we inferred a duplication history of the 3 ARM domains homologous to MgtE's ARM domain.
- MotAB results confirm & extend Puenta-Lelievre et al. (poster)
- FlpQR results show FlpR homologous to FlpQ+FlpQ, suggest Flp-FlpQ relationship, and three confirm nonflagellar T3SS are derived.

Future work: combine all core flagellum proteins for best-ever estimate of flagellum root & early divergences; use to map assembly of features.

References / Acknowledgements

Karren S. Rogovin, Kozen EV (2016). MitoCOCs: clusters of orthologous genes from mitochondria & implications for the endosymbiont. BMC Evol Biol 14:237
Kuhlen L, Johnson S, Zeller A, Baute S, Dorne JC, Cassar AE, Dabor R, Fisher J, Wagner S, Lea SM (2020). The substrate specificity switch FlpB assembled into the export gate to regulate type three secretion. Nature Communications 11: 2296
Matzke NJ, Puenta-Lelievre C, Baker Matthew AB (2025). A Pipeline for Generating Datasets of 3-Dimensional Tertiary Interaction Characters for Model-Based Structural Phylogenetics. A Pipeline for Generating Datasets of 3-Dimensional Tertiary Interaction Characters for Model-Based Structural Phylogenetics. Chapter in: Evolutionary Genomics: Methods and Protocols, Series Title: Methods in Molecular Biology, Series Editor: Gustavo Castano-Andrade
Matzke NJ, Li Changhao (2025). Protein structure characters in the light of phylogenetic systematics. GSF Preprints submitted to Genome Biology and Evolution
Special Issue on Structural Phylogenetics, July 1 2025, pp. 1-18. <https://doi.org/10.1101/2025.07.01.658888>
Puenta-Lelievre C, Malik AJ, Douglas J, Ascher D, Baker MAB, Allison J, Poole A, Lundin D, Fullmer M, Bouckert R, Kim H, Obara M, Steinegger M, Matzke NJ (2025). Tertiary interaction characters enable fast model-based structural phylogenetics beyond the twilight zone. bioRxiv 217817
Puenta-Lelievre C, Ridone P, Douglas J, Amritkar K, Kaçar B, Baker MAB, Matzke NJ (2024). Molecular and structural innovations of the stator motor complex at the base of flagellar motility. bioRxiv 504460
Palen M, Matzke N (2026). "From the Origin of Species to the Origin of Bacterial Flagella." Nature Reviews Microbiology 4:10, 784-790.
Van Kempen et al. (2024). Fast and accurate protein structure search with Foldseek. Nature Biotechnology 42: 243-246.
Varadi et al. (2024). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nature-Alpha Research 5(01), 0439-0444.
Funding support: NJM, CPL, and MAAB were supported by HESP grant RGP0060/2021. MAAB and NJM were supported by ARC DP240100462, and NJM was additionally supported by New Zealanders Higher Society RDF 21-LJOA-040. Mastercard Grant 18-LJOA-034, and University of Auckland DRDF Research Fund #3727963. CPL and NJM were supported by University of Auckland, Faculty of Science Research Development Fund, FoS RDF-#3732317.