Contents

[Project Module Booklet 1](#_Toc138103513)

[Project About (Initial) 3](#_Toc138103514)

[17/05/2023 – Curate current knowledge (AmiC, day 1) 4](#_Toc138103515)

[19/05/2023 – Curate current knowledge (AmiC, day 2) 5](#_Toc138103516)

[22/05/2023 – Curate current knowledge (Amidase\_3 family) 6](#_Toc138103517)

[25/05/2023 – Curate current knowledge (Literature Search, AmiC) 7](#_Toc138103518)

[30/05/2023 – Next Steps 8](#_Toc138103519)

[31/05/2023 – Quantitative Analysis of the Mobile Helix (Day 1) 8](#_Toc138103520)

[01/06/2023 - Quantitative Analysis of the Mobile Helix (Day 2) 10](#_Toc138103521)

[04/06/2023 - Quantitative Analysis of the Mobile Helix (Day 3) 12](#_Toc138103522)

[05/06/2023 – Structural Research 12](#_Toc138103523)

[06/06/2023 – Literature Research 14](#_Toc138103524)

[08/06/2023 – Existing domain tools 15](#_Toc138103525)

[18/06/2023 – Sequence analysis, day 1 17](#_Toc138103526)

[19/06/2023 – Sequence analysis, day 2 17](#_Toc138103527)

[20/06/2023 – Annotated Bibliography + Alignments 18](#_Toc138103528)

[21/06/2023 – Conservation of sequence (helix identification?) 18](#_Toc138103529)

[22/06/2023 - Ideas for taxonomic map and presentation prep 18](#_Toc138103530)

Possible proteins involved in Septal PG separation to start looking into (from Peters et al. 2011 and Rocaboy et al., 2013) – in bold: focus on these first:

* **AmiC** (possible orthologs: Ami1 (Mycolicibacterium), PlyPSA (bacteriophage), Rv3717 (M. tuberculosis), LytH (S.aureus), AmiC2 (Nostoc puncitforme), Cw1D, Cwp6, and CD27L (C. difficile), CwIV (Bacillus), AmiP (Thermus parvatiensis)
* **AmiB**
* AmiA
* **EnvC (alt: YibP)**
* **NIpD**
* FtsZ, FtsA, ZipA, ZapA, ZapB, ZapC (Z-ring formation proteins, step before septal ring splitting)

# Project Module Booklet

* Award of MSc dependent on project completion, = 60 credits
* Will be out of 100 (85% for project and thesis, 15% for viva)
  + Supervisor gives a mark for performance and a mark for thesis
  + Second marker gives a mark for thesis
  + All marks out of 100 with a explanatory report to justify the mark.
  + Distinction/Merit/Pass/Fail
  + Viva: 15-20 mins oral exam, with Mark, Irilenia, Tristan, and external examiner (may or may not have external examiner)
* Bibliography not assessed
* Be in regular contact, agree schedule/frequency to meet.
* Mix literature with practical work
* Can ask supervisor for guidance but not for step-by-step instruction
* Notes for methods, scripts, result summaries, figures, tables for reporting back to supervisor. Basically keep a journal and record everything you try, tools you try out, results you get along the way.
* **~~Q: Expectations for part-time students in general? (are we generally expected to work on the project from now until summer next year, or just during/around summer term?)~~ no (but you won’t be stopped!)**

Bibliography:

* 5-10 references
* Each reference annotated: main results, relevance to project.
* Each entry contains a full Bioinformatics-style reference, 1-2 sentence summary of context, relevance to project, then a paragraph or two describing and evaluating the study (key results, methodologies, maybe a figure)
* When done, email a copy to Mark and arrange a meeting to discuss. Present the findings back.
* Maximum of 10, don’t include reviews (primary research only)

Thesis

* Title: Full project title, name, supervisor, degree course, institutional address
  + **~~Q: Is there a title page template we should use with university branding on it?~~ No fixed format as they won’t go into the Birkbeck repository) no specific title page format (but can use the BBK logo if you want).**
* Abstract (200-300 words)
* Contents page (Chapters and sections, tables and figures with corresponding page numbers)
* Acknowledgments
* List of non-standard abbreviations used
* Introduction (context, review of prior work and aims of project – aims should be a clear statement at the end of the intro)
* Materials and Methods (procedures followed, pitfalls in methods used, reference published technique)
  + Esp: Code development record eg A Perl script was written to translate a PDB file into a MOL2 file format
  + Closely related scripts may be in one statement
  + Short code snippets may be included in appendices single-spaced BUT only snippets or small scale programs, use GitHub as a repository for larger/bulk of the project.
* Results (data in format to make significance of results clear. Tables, graphs, titled and captioned, referred to in-text
  + **~~Q: Can we have additional figures in the appendices if broadly relevant but not directly discussed in-text, or keep these stored on a page like GitHub?~~ yes**
* Discussion (results in context of published work, identify shortcomings of project, greater significance of results demonstrated, outline of follow-up work)
* Conclusions
* References (cited in-text, use consistent format such as Bioinformatics referencing style)
* Appendices
* **~~Q: ‘Check list is provided’ for thesis/reports – where can we find this?~~ – end of booklet**

Admin

* 10,000 word maximum but not penalized if over this within reason.
* 1.5 line spaced on A4, do NOT need to print the report.
* Submit electronically, **PDF format**
* **Submit via Moodle (date TBC)** and also send identical copy to supervisor
* Paginated (page 1 is the title page)
* Tables and figures may either be woven within the main body of text or placed in order at the end of the chapter, with complex or detailed tables/output in appendices.

# Project About (Initial)

* Bacterial cell wall: Mega-Macromolecule made from linear glycan chains interlinked with short peptides
* Dynamic, allows for growth and bacterial division.
* N-acetylmuramyl-L-alanine amidases (NAMLAA) enzymes which cleave peptide cross-links from glycan chain
* A metalloenzyme
* Modular structure containing amidase domain with cell wall/enzymatic cell wall binding domains
  + What are these domains?
* Amidase\_3 PFAM family = many NAMLAAs, maybe entirely. Essential to separation of daughter cells in cell division, cleave peptidoglycan at septum
  + Cell separation in gram negative bacteria (knockout studies show essentiality?)
  + Coordination of dissembly and reassembly of the daughter cells to maintain the structure while separating
  + Monitor the cell length, formation of the septal ring complex (protein ring which gradually gets smaller)
  + Also involved in chromosome separation
  + Amidase actually performs the cutting of the septal ring.
  + LysM – nitrogen fixing bacteria
  + Fts2: Tubulin enzyme, makes the ring
  + AmiC: Amidase, cuts the ring. AmiC is the name in E.Coli bacteria
  + Is a multi-domain protein (two domains – cell wall attachment and enzymatic action)
  + EnvC domain: ensures activation only at the septal ring
  + LytM domain: Believed that EnvC interacts with this domain? AmiC is part of EnvC on a CC stalk, want to find evidence of EnvC activation.
* **Aim: Collect and organize information on variation in the architecture of Amidase\_3 enzymes across bacterial species**
  + Gather evidence: Sequences of AmiC (?) from many species and find the correlated sequence variation in those genes
  + Architecture of the proteins also valuable data to examine how the cell wall binding domain of AmiC might be different
  + AMIN domain: Explore conserved domains in AMIN and LysM
  + Non-gram negatives: There are other architectures available with different binding domains - what are these roles, are they still cell-wall related despite not having peptidoglycan?
    - Enzyme is conserved in gram +ves – dormant spores perhaps, where the cell wall might change state and then need AmiC? Unclear what NAMLAAs are doing in gram positives, are they doing something similar with the septal ring or something different? Chromosomal separation only?
    - There is a different cell division enzyme in gram positives, yet they retain an EnvC homolog?
  + What *are* all he known architecture? Is there evidence of these interacting with the same proteins/domains?
  + AmiA and AmiB are also present, but may lack the same domains. What are these doing?
  + In gram negatives: AMIN sequence is conserves, but its unclear how AMIN recognises peptidoglycan?
* **Aim: Use sequence variation to investigate questions around function of these enzymes**
  + Suggestion: Some family members auto-inhibited by cell-wall binding domain, which are then activated by another cell division protein EnvC/NIpD
  + Are there any other suggested functional mechanisms for these enzymes in the literature?
* To learn about: Large-scale multiple sequence alignment and covariation analysis (using Gremlin?) to see if there is evidence in evolving sequences for the suggested mechanisms in the literature

# 17/05/2023 – Curate current knowledge (AmiC, day 1)

**Protein Structure Search (based on 4BIN structure in PDB)**

* **Search 1: blastp:** Used FASTA search from 4BIN PDB sequence, database Protein Data Bank proteins (pdb) (>4BIN\_1|Chain A|N-ACETYLMURAMOYL-L-ALANINE AMIDASE AMIC|ESCHERICHIA COLI (83333)MHHHHHHSSGRENLYFQGHMQVVAVRVWPASSYTRVTVESNRQLKYKQFALSNPERVVVDIEDVNLNSVLKGMAAQIRADDPFIKSARVGQFDPQTVRMVFELKQNVKPQLFALAPVAGFKERLVMDLYPANAQDMQDPLLALLEDYNKGDLEKQVPPAQSGPQPGKAGRDRPIVIMLDPGHGGEDSGAVGKYKTREKDVVLQIARRLRSLIEKEGNMKVYMTRNEDIFIPLQVRVAKAQKQRADLFVSIHADAFTSRQPSGSSVFALSTKGATSTAAKYLAQTQNASDLIGGVSKSGDRYVDHTMFDMVQSLTIADSLKFGKAVLNKLGKINKLHKNQVEQAGFAVLKAPDIPSILVETAFISNVEEERKLKTATFQQEVAESILAGIKAYFADGATLARRG)
  + 10 proteins identified by sequence (noted on Amidase\_3 analog word doc)
  + AmiC analog called LytH in S. aureus, activated by ActH (is this the same process described for AmiC and NIpD in E.coli?)
  + The AmiC analog in clostridium difficle has also been called Cwp6
  + **To do: Use multiple seq. alignment to show in Chimera where there is overlap/identity between these analogs**
    - Alignment: looks to be within catalytic domain for AmiC rather than N-terminal domain?
* Search 2: blastp: FASTA search from 4BIN (as above), use NCBI experimental non-redundant database (Clustered nr) instead of PDB
  + >100 identified clusters from NCBI accessions (no related PDB structures, sequence only)
  + Mostly Enterobacter in the top 100 (eg E. coli), several uploaded by Heidrich et al., 2001 (paper describing AmiA, AmiB, AmiC in E. coli – **to do: read this paper**)
  + **To do: Download these sequences, perform alignment?**
* Search 3: PDB, search by family (amidase\_3)
  + 17 hits, most identified in the BLASTp Search 1
  + Updated word doc with additional structures from this search
  + Searching by the N-terminal domain (AMIN) gets 2 additional hits (3JC8 and 3JC9 – type Iva pilus machinery. <https://www.science.org/doi/10.1126/science.aad2001>: this machinery binds through the cell membrane/wall of gram negative, including binding to peptidoglycan – evidence for binding to PG by AMIN domain?
  + SCOP search for N-acetylmuramoyl-L-alanine amidase-like family (4001130): 8 hits, no additional structures identified from this (all already identified via PDB search)
  + SCOP domain search: 8092183: No additional hits (only 4BIN recorded)
  + SCOP search for AMIN domain-like family (4005368): 2 additional hits (4AQZ – another type Iva pilus machine, and 5H3K – hypothetical protein structure from *Synechocystis ‘Kazusa’*. No attached publication for this structure.)
* Search 4: Pfam/Interpro sequence identity search

**To do next time:**

* Search in Pfam/Interpro for sequence identity, add structures to main Amidase\_3 list
* Multiple seq. alignment of identified structures, display in Chimera, see what is aligned between the amidase\_3 proteins
* Multiple seq. alignment of the second blastp search – for sequence-only data, what is aligned here of those with high identity with 4BIN (AmiC in E.coli)
* Literature review of the identified protein structures, esp. recent gram positive papers

# 19/05/2023 – Curate current knowledge (AmiC, day 2)

* Pfam sequence identity search: Found Amidase\_3 and AMIN domain match to the FASTA sequence from 4BIN
  + Amidase\_3: PFAM PF01520: identified in 6290 species, mostly bacterial but some viral as well
    - 7187 sequences with Amidase\_3 architecture, however 1940 with AMIN and amidase\_3, 515 with AMIN, amidase\_3, and LysM (410 total unique architectures here)
      * **Needs looking into more? Maybe record the domains in each structure? Is there a pattern (eg one gram type has LysM and the other doesn’t)**
    - 27 instances in PDB with Amidase\_3 (all already recorded in document)
    - 41 AlphaFold structure predictions in various bacteria
  + AMIN: PFAM PF11741: identified in 2765 species (all bacteria)
    - 5051 sequences with AMIN architecture, most of these are AMIN plus Amidase\_3 or AMIN plus LysM, but 633 also with secretin, secretin N and STN domains?
      * **Is this important?**
    - 3 instances in the PDB and 14 predicted alphafold structures (the 3 PDB have already been identified, see Day 1 notes)
* Interpro
  + Amidase\_3: 23 predicted pathways, including peptidoglycan recycling II but also some synthesis pathways like taurine biosynthesis? (peptidoglycan recycling: <https://metacyc.org/META/NEW-IMAGE?type=NIL&object=PWY-7883&redirect=T>)
    - 61,000+ alphafold predictions, 1013 domain architectures, 74,000 protein structures (37 reviewed by UniProt/SwissProt)
      * Reviewed structures: 1 bacillus phage Endolysin, various amidases (Bacillus subtilis, E.coli, Myco. TB mostly), 3 uncharacterized
    - Only 19 PDB structures (already recorded)
  + AMIN: just 1 pathway – peptidoglycan recycling (<https://metacyc.org/META/NEW-IMAGE?type=NIL&object=PWY-7883&redirect=T> – takes place in cytosol?)
    - 7 PDB structures and 24,000+ alphafold predictions (4 new PDB structures: 4av2, 6ve2, 6ve3, 6ve4. All PilQ complexes from *Neisseria meningitidis* or *Pseudomonas aeruginosa* (the same typeIV secretion system as identified in Day 1) – is the E.coli amidase the only structure out there with AMIN that isn’t one of these secretion systems??
    - 242 domain architectures, 29,000+ protein structures (9 reviewed by UniProt/SwissProt)
      * Reviewed structures: Mix of PilQ and AmiC

Wider search for sequences

* Of amidase\_3 family structures in PDB: 10 are gram positive, 4 are gram negative
* PDB: search for GO annotation “N-acetylmuramoyl-L-alanine activity”
  + Found 147 deposited structures (added to summary sheet), mostly amidase\_2 and other domains but also a lot of non-bacterial structures. A lot of camels.
  + Downloaded all sequences on the summary sheet to PDB folder for further analysis
  + Identified some related domains which aren’t NAMLA but are listed under that GO annotation:
    - [PF01471](https://www.rcsb.org/search?q=rcsb_polymer_entity_annotation.annotation_id:PF01471%20AND%20rcsb_polymer_entity_annotation.type:Pfam&rt=polymer_entity) – PG\_binding\_1 domain (found in some of the Amidase\_2 family) – look into?
    - PF05036 – SPOR domain (binding to PG, involved with sporulation and division proteins like FtsN?)
    - PF01832 – Glycosaminidase (hydrolysing peptidoglycan)
    - PF00877 – NipC domain in TB (related to NIpD??)
    - PF12123 – CBC\_PlyG (cell wall binding domain in bacteria and viruses)
    - PF05257 – CHAP domain (related to amidase function, cell wall metabolism in bacteria, amidase domain of E. coli glutathionylspermidine synthetase?)

Questions

* Possible direction of research into all possible architectures of Amidase\_3 and the patterns in these?
* Should I include alphafold predictions at this stage?
* Is it worth exploring AMIN architectures further with the scarce detail around them?

**To do:**

* Align all the PDB structures in Chimera, see what is best aligned between the amidase\_3 proteins (do a couple, one for those that BLAST identified for a refined version, a version for amidase\_3 in PDB, and an all-out version with everything I found)
* For the sequence-only data: Download sequences from InterPro or BLAST top 100 and do multiple sequence alignment (what is most conserved?)
  + API call is available from InterPro to download the seqs of all protein sequences in the family, unreviewed included!
  + **How many are gram positive vs gram negative for sequences only? Is the skew towards gram positive a selection bias in PDB, or are amidase\_3’s just more common in gram positive?**
* Literature review of the identified protein structures, esp. recent gram positive papers

**More in-depth directions:**

* Further research into AMIN and Amidase\_3 pathways – get a grip into what these proteins are interacting with across their predicted pathways, and if the literature backs this up
* Look into other PG binding domains (eg PG\_binding\_domain and SPOR domain) and the overlap in architecture with Amidase\_3/Amidase\_2?

# 22/05/2023 – Curate current knowledge (Amidase\_3 family)

* Amidase\_3 Protein Structure alignments
  + MatchMaker in Chimera run for the 18 Amidase\_3 family structures (secondary structure predictions not re-calculated, pairwise alignment chosen, compared against 4BIN (AmIC in E.coli), all other setting default – Needleman-Waltsch alignment method chosen).
  + Colored by species (didn’t really see a pattern outside of the core amidase domain)
  + Colored by gram status (again no real pattern outside of the amidase domain)
  + Cut out all except the amidase domain in each of the protein families
    - Based on the PDB text file (defined boundaries of the amidase domain – auth annotation of amidase\_3 domain, NOT PFAM defined boundaries)
    - Deleted all chains not containing the amidase\_3 domain
    - Cropped to just the defined domain per PFAM annotation and PDB file
    - Re-aligned all models – before cropping, saw that the N-terminal domain was unique among the amidase\_3 family proteins, 5J72 (C. difficile) had a large unique CW-2 binding domain, and the bacteriophage in S. aureus also had a unique disordered region (contained some sheets?)
    - **On cropping and alignment only on the amidase domain: saw that gram negative structures had a helix or other structure blocking the active site, but gram positive don’t seem to. Could be worth a further look?**

**To do:**

* Difference between Amidase\_2 and Amidase\_3 protein family
* For the sequence-only data: Download sequences from InterPro or BLAST top 100 and do multiple sequence alignment (what is most conserved?)
  + API call is available from InterPro to download the seqs of all protein sequences in the family, unreviewed included!
  + **How many are gram positive vs gram negative for sequences only? Is the skew towards gram positive a selection bias in PDB, or are amidase\_3’s just more common in gram positive?**
* **Literature review of the identified protein structures, esp. recent gram positive papers**

**More in-depth directions:**

* Further research into AMIN and Amidase\_3 pathways – get a grip into what these proteins are interacting with across their predicted pathways, and if the literature backs this up
* Look into other PG binding domains (eg PG\_binding\_domain and SPOR domain) and the overlap in architecture with Amidase\_3/Amidase\_2?

# 25/05/2023 – Curate current knowledge (Literature Search, AmiC)

* **Annotated bibliography:**
  + Re-read and annotated the two papers from the project proposal
  + David baker lab: PPI coevolution modelling: annotated - one protein identified as strongly coevolving with AmiC (bepA – Beta-barrel assembly-enhancing protease)
  + Deborah marks lab EVComplex2 for E.coli PPI interactions (more useful as don’t need characterized protein structures!)
  + Search for ‘AmiC E.coli’ in PubMed, 48 results
    - Several studies using AmiC and the Tat exporter pathway mutants to explore antimicrobial effects or use in export of recombinant protein products from E. coli
    - Possible interaction with RNAse E?
    - AmiC might have a role as a exporter signal peptide? (AMIN type IV link from previous research, plus studies looking into its activity in Tat pathway)
    - Two possible candidates for interaction with AmiC? (DolP and ActS – ActS has a stronger link, 3 studies in particular, seems to prefer AmiC under standard growth conditions but AmiB in acidic pH?)

**To do:**

* For the sequence-only data: Download sequences from InterPro or BLAST top 100 and do multiple sequence alignment (what is most conserved?)
  + API call is available from InterPro to download the seqs of all protein sequences in the family, unreviewed included!
  + **How many are gram positive vs gram negative for sequences only? Is the skew towards gram positive a selection bias in PDB, or are amidase\_3’s just more common in gram positive?**

**More in-depth directions:**

* Difference between Amidase\_2 and Amidase\_3 protein family
* Further research into AMIN and Amidase\_3 family pathways – get a grip into what these proteins are interacting with across their predicted pathways, and if the literature backs this up
* Look into other PG binding domains (eg PG\_binding\_domain and SPOR domain) and the overlap in architecture with Amidase\_3/Amidase\_2?

# 30/05/2023 – Next Steps

* **Amidase\_3 Structural Research**
  + Focus on getting quantitative results – esp. for perceived differences between gram negative and gram positive bacteria (using PDBeFold plus another tool?)
  + If a difference seen, could incorporate AlphaFold predictions?
* **Amidase\_3 Literature Research**
  + Look up papers for the 19, ensure their cell division function has been confirmed.
* **Amidase\_3 Sequence Research**
  + For the sequence-only data: for proteins with function confirmed (ie in septal wall division), search for similarity in sequences in non-redundant NCBI dataset. (blastp)
  + How many are gram positive vs gram negative for sequences only? Is the skew towards gram positive a selection bias in PDB, or are amidase\_3’s just more common in gram positive?
  + Quantitative similarity/multiple sequence alignment
  + Expanded dataset – Isolates (API call from InterPro for seq download)
* **LytM Factors for AmiC**
  + ActS: Any structures of this? Other papers around function? How might the docking/binding of this be modelled (could compare to NlpD as this is more established as a cofactor in AmiC).
  + DolP: The same as above, evidence less confident for this factor however.
* **Programmatic tasks**
  + R package for k-means clustering
  + Identification of the amidase\_3 domain from sequences (for future cropping, multiple sequence alignment, or searching). Identify conserved regions across amidase\_3 proteins, create a pattern recognition package for these regions?

(spent this session arranging the GANTT chart and planning for the Master’s project)

# 31/05/2023 – Quantitative Analysis of the Mobile Helix (Day 1)

**Today’s Focus: Amidase\_3 Structural Research**

* *Focus on getting quantitative results – esp. for perceived differences between gram negative and gram positive bacteria* (using PDBeFold plus another tool?)
* If a difference seen, could incorporate AlphaFold predictions?

PDBeFOLD (Global comparisons)

* Note: A high level of secondary structure identity and low sequence identity could indicate the same structural family (per [PDBeFOLD instructions)](https://www.ebi.ac.uk/pdbe/docs/Tutorials/workshop_tutorials/PDBefold.pdf)
* Pairwise alignment search for 4BIN similarity (in Whole PDB archive, chain A for 4BIN, lowest acceptable match = 70%, match individual chains, connectivity, unique and best matches only. Normal precision, sorted by Q-score.)
  + No non-identical matches at 70% similarity
  + All Q scores <=0.1 except 1XOV (bacteriophage for listeria, already in structure list)
  + All sequence identities <50%
  + %seq (alignment of residues based on alignment in space) <=%17 for all identified structures
  + 80 matches identified at lower similarity threshold (80/187732 entries)
    - Submitted for multiple sequence alignment
    - MSA did not work.
* Output 1: Multiple alignment, uploaded the 19 amidase\_3 files using a PDB list text file:
  + PDB codes separates by tabulation, in the format ‘4BIN:A’ for chain A in 4BIN, for example.
  + Q score = alpha carbon backbone alignment. 1 is the highest score.
  + P score = RMSD, aligned residues, gaps, matched secondary structure elements
  + Z score = Gaussian statistics
  + RMSD = Root mean square deviation (difference between prediction values and estimator. Close to 1 for backbone = good, larger than 6 indicates poor quality OR conformational changes in the protein.
* Results from output 1:
  + Conserved across all: Glycine-Histidine pair at end of first beta sheet (Gly195, His196), glutamate at start of first helix (Glu211), histidine at end of second sheet (4BIN HIS265), glutamate in sheet 5 (Glu373)
  + For two gram negatives, there is an additional helix between 289-307 for 4BIN and 280-298 for 3NE8 that isn’t there for the gram positives or the other gram negatives and does not align to anything in the MSA.
  + Mycobacteria have a region near the start of the amidase which isn’t conserved in any other species, region contains a helix and at least one beat sheet, region is in 7AGO, 7AGL, 7AGM, 4M6G, 4LQ6. Most, not all of the mycobacteria.
  + Mycobacteria do all have a 7 residue motif, residues not all in common however some are conserved (Tyr-Ile-Gly-X-X-Gly).
  + Bacteriophage amidase had very long C-terminal domain containing beta sheets after the amidase enzyme (did not align to anything, makes sense because all other structures are bacterial)
  + RMSD pretty good, consensus scores between 0.6243 and 1.7692 (overall RMSD: 1.5525)
  + Q score less good (overall score for the alignment is 0.1255)
  + ## Structure Nres Nsse RMSD Q-score
  + 1 PDB 3ne8:A ..................... 226 13 1.0290 0.5186
  + 2 PDB 4bin:A ..................... 348 23 0.8130 0.3507
  + 3 PDB 3czx:A ..................... 182 11 1.6146 0.5581
  + 4 PDB 7b3n:A ..................... 169 11 1.7739 0.5743
  + 5 PDB 1jwq:A ..................... 179 12 0.9407 0.6663
  + 6 PDB 5j72:A ..................... 638 51 1.1552 0.1788
  + 7 PDB 4rn7:A ..................... 186 12 0.7130 0.6666
  + 8 PDB 7rag:B ..................... 197 12 0.9918 0.5995
  + 9 PDB 3qay:A ..................... 180 11 1.7692 0.5400
  + 10 PDB 7ago:A ..................... 216 15 0.6897 0.5760
  + 11 PDB 7agl:A ..................... 212 15 0.6870 0.5871
  + 12 PDB 7agm:A ..................... 218 15 0.6632 0.5729
  + 13 PDB 5emi:A ..................... 180 13 0.8254 0.6766
  + 14 PDB 7tj4:B ..................... 176 12 1.0820 0.6586
  + 15 PDB 1xov:A ..................... 317 25 1.5939 0.3223
  + 16 PDB 4m6g:A ..................... 216 15 0.6425 0.5799
  + 17 PDB 4lq6:A ..................... 213 15 0.6333 0.5888
  + 18 PDB 4m6i:A ..................... 185 12 0.6243 0.6787
  + 19 PDB 4m6h:A ..................... 190 12 0.6263 0.6607
* Output 2: Gram negative alignment (4 structures, same settings as output 1)
  + RMSD 2.243 and Q-score 0.2389 for these (alignment is somehow worse just in gram negative amidases)
  + 4BIN is the only one with an N-terminal part
  + Helix does not align with AmiP amidase or the Neisseria meningitidis NAMLA (**are the first two the same enzyme but the other two not? Functional difference?)**
  + Conservation of secondary structures across the 4 gram negatives
* Output 3: Gram positive alignment (10 structures, same settings as output 1)
  + RMSD 1.519 and Q-score 0.1616 (RMSD much better for just gram positive bacteria)
  + Secondary structures well conserved across the bacteria even though sequence is not
* Output 4: Mycobacteria genus only (7 structures, same settings as output 1)
  + RMSD 0.4795, Q-score 0.7921 (very well aligned on the surface, however sequence identity for the final four is the same, therefore questioning how good this is)
  + Removed the three duplicates (retained 4LQ6:A) and re-ran with 4 non-redundant structures, and new RMSD was 0.4678, Q-score 0.9316 (actually improved the alignment!)
  + All secondary structures conserved even if sequence wasn’t
  + But on the whole sequence was conserved as well as structure.
* Output 5: Re-run of output 1 with the duplicate mycobacteria removed (16 structures total)
  + RMSD slightly higher than before (1.712) and Q-score: 0.1256.

Unfortunately the visualiser for the alignment was not working in PDBeFOLD

Suspicion that not all of these 19 structures are doing the same function (are not all AmiC for example) – to explore their functions in the literature at later date.

# 01/06/2023 - Quantitative Analysis of the Mobile Helix (Day 2)

Blastp for the other 18 structures

* -4 of the structures because they are duplicate sequences (structure in complex with a molecule for example)
* Attempt to identify even more structures based on similarity, added search results to excel sheet
  + 5J72: Clostridium difficile potential amidase, had 4 structures with higher similarity than to other amidases (probably due to N-terminal domain)
  + 1XOV bacteriophage amidase had 7 additional structures found with similarity in BLASTp. A variety, but includes alanine racemase in S. aureus (which is involved in cell wall biosynthesis)
  + No other new structures identified in PDB sequence similarity search
  + Simple search for AmiA, AmiB, and AmiC in the PDB found 3 additional structures all from late 1990s in Pseudomonas aeruginosa, identified AmiC as having periplasmic binding domain and regulating the Amidase operon, as well as a few more potential NAMLAA in gam negative species by searching AmiA. Additional 4 structures found.
* Re-did the PDEeFOLD analysis for the new 19 structures
  + RMSD much higher for the new 4 compared to all the other structures previously searched
  + Overall RMSD: 2.858, Overall Q-score: 0.01216 (worse than without the new structures)
  + Difference in secondary structures for the new four as well
  + 1 PDB 3ne8:A ..................... 226 13 0.8553 0.2046
  + 2 PDB 4bin:A ..................... 348 23 0.8374 0.1333
  + 3 PDB 3czx:A ..................... 182 11 1.6479 0.2110
  + 4 PDB 7b3n:A ..................... 169 11 1.1354 0.2588
  + 5 PDB 1jwq:A ..................... 179 12 0.9554 0.2536
  + 6 PDB 5j72:A ..................... 638 51 0.8474 0.0726
  + 7 PDB 4rn7:A ..................... 186 12 0.8658 0.2481
  + 8 PDB 7rag:B ..................... 197 12 1.0242 0.2273
  + 9 PDB 3qay:A ..................... 180 11 1.7365 0.2081
  + 10 PDB 7ago:A ..................... 216 15 0.9087 0.2120
  + 11 PDB 7agl:A ..................... 212 15 0.9020 0.2163
  + 12 PDB 7agm:A ..................... 218 15 0.9186 0.2097
  + 13 PDB 5emi:A ..................... 180 13 1.0318 0.2484
  + 14 PDB 7tj4:B ..................... 176 12 0.8867 0.2613
  + 15 PDB 1xov:A ..................... 317 25 1.5956 0.1229
  + 16 PDB 4lq6:A ..................... 213 15 0.9838 0.2119
  + 17 PDB 4h4j:A ..................... 236 18 3.9441 0.0776
  + 18 PDB 2p1g:A ..................... 229 15 3.7287 0.0858
  + 19 PDB 4q68:A ..................... 235 18 3.6350 0.0862
  + 20 PDB 4q5k:A ..................... 237 17 3.8969 0.0785
* Therefore is this evidence for the other four self-described as AmiA to be a different enzyme to the others with a different function? (and therefore the others are actually AmiC possibly?)
  + These four seem to be separate from the others in BLASTp searches, all the other amidase\_3 appeared in similarity searches of all the others (except for the bacteriophage, 1XOV, and the Neisseria meningitidis one, 3CZX (which had no similarity matches).
  + All the others (except the bacteriophage) have been reported with Zinc ion ligands, but the four AmiA have not (reported with sodium ion instead).
* To structure list: noted which search parameters they appeared in (appeared in more = better link to amidase\_3 function?)

**Next session:** Use PDBeFOLD for the 16 structures, have any structures been missed?

Local alignment in that mobile helix. Explore other alignment tools, not just PDBeFOLD. (Global distance test GDT-TS, RMSDlocal, lDDT (local distance difference test, superposition-free, <https://swissmodel.expasy.org/lddt>), TM-Align (<https://zhanglab.ccmb.med.umich.edu/TM-align/> ), DSSP (<https://swift.cmbi.umcn.nl/gv/dssp/> ), PROMOTIF, CE (<http://cl.sdsc.edu/>).

Then: look into published functional data for each structure (are they proven to be acting in the cell wall division pathway?)

# 04/06/2023 - Quantitative Analysis of the Mobile Helix (Day 3)

**PDBeFOLD for the other structures**

* Search parameters: Against Whole PDB archive, lowest acceptable match of secondary structure alignment of 50% (with other close matches shown if none within limits), matching individuals chains and connectivity (unique and best matches only), precision Normal. Chose lower sequence similarity than default due to observed low sequence identity in the original 19 identified sequences with NAMLAA predicted function.
* Where similarity is identified in multiple chains for the same structure, have ensured all scores are recorded.
* Added results with a Q-score above 0.3 to document (based on MSA of the original 19 structures where all Q-scores were above 0.3 threshold)
* If alignment was to another chain in the same model, Q-score and RMSD was not recorded.
* Structure not included if P-score was below 3, as match is considered statistically insignificant (per PDBeFOLD stats advice)
* Of note: 4BIN only had one other match with these parameters (3NE8)
* Of note: 5J72 had no matches at 50% SSE, therefore repeated at 40%. No matches that weren’t the same model with a P value above 3.
* Of note: only 27 matches with 1XOV, and only one of these (a cell wall binding endolysin domain in bacteriophage) was above threshold for P and Q scores.

Total structure list after searching: 24 (the original 19 plus 5 additional un-annotated family structures, all gram negative, possibly AmiA and which don’t align with any significance with any of the original 19).

**US-Align**

* Using updated US-align (comparison of 3D structures, allows MSA of multiple protein files, updated from previous TM-align)
* Ran US-align locally with the 19 structures in amidase\_3 family (hypothetical from research, disregarded AmiA suspicious ones).
  + Local download, output as PyMOL viewable files for the alignment
  + Sequence-independent TM score for each
  + Command run: ./USalign -dir chains/ chain\_list.txt -mm 4 -o sup
* Of note: 7RAG (amidase in complex with lipoprotein from Clostridium difficile) had a very low TM score (,0.01), but did align with others on BLASTp similarity search on PDB structures, and PDBeFOLD aligned it with several other NAMLAAs.
* Bacteriophage and Staph. A LytH also had poor TM scores (around the 2.4 mark)

# 05/06/2023 – Structural Research

**Amidase\_3 Structural Research**

* Used US-align to look at secondary structures between the 24 total.
  + ./USalign -dir chains/ chain\_list.txt -mm 4 -o sup -full T > output.txt
  + This produces a full alignment output with all structures aligned to all other structures.
* Used US-align between gram positive and gram negative in the 24.
  + Repeated for 9 gram negative
  + Repeated for 10 gram positive
  + Repeated for all Mycobacteria
* Repeated alignment with N-terminal domains removed
  + Cropped the 19 down per below chimera code (using author assigned residue boundaries for the catalytic domain. Did not do for the potential AmiA orthologs as these were undefined in the PDB with no protein family annotations, therefore uncertainty about bounds of the domain).
  + # cropping all except the amidase domain
  + # 1jwq is just the catalytic domain, no cropping needed
  + select #1:6-173 #2:5-174 #3:182-397 #4:3-175 #5:190-405 #6:5-208 #7:29-232 #8:29-232 #9:29-232 #10:119-293 #11:441-610 #12:463-644 #13:19-222 #14:45-248 #15:19-222 #16:7-175 #17:43-231 #18:1-291
  + select invert
  + Actions -> Atoms/Bonds -> delete
  + Also removed extraneous chains (used the best match chains from original BLASTp search)
  + Saved each as individual models, realigned the 24 using USalign.
  + This did improve the TM score for 7RAG considerably; glitch in system, have to use Chain B for this because chain A in 7RAG is a lipoprotein.
  + TM score around 0.3 for the potential AmiA orthologs, but around 0.7 for all the others.
* Repeated PDBeFOLD for the 24 (cropped to domains as above)
  + Uploaded edited structures as a tarball
  + Repeated with just the 19 as the 24 produced an RMSD above 3 and Q-score of around 0.2 (these 5 possible AmiA seemed to throw off the alignment)
  + Overal RMSD: 1.655, Overall Q-score 0.3561
  + Notably very well conserved secondary structure alignment:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1jwq.pdb:A |  | - | - | - | s | H | s | h | S | S | h | s | h | s | h | h | - |  |  |  |  |  |
| 1xov.pdb:A | - | - | - | s | H | s | h | S | S | h | s | h | s | h | h | - |
| 3czx.pdb:A | - | - | - | s | H | h | - | S | S | h | s | h | s | h | - | - |
| 3ne8.pdb:A | - | - | - | s | H | h | - | S | S | h | h | s | h | s | h | h |
| 3qay.pdb:A | - | - | - | s | H | s | h | S | S | h | s | h | s | h | - | - |
| 4bin.pdb:A | - | - | - | s | H | h | - | S | S | h | h | s | h | s | h | h |
| 4lq6.pdb:A | s | h | s | s | H | h | - | S | S | h | s | h | s | h | h | - |
| 4m6g.pdb:A | s | h | s | s | H | h | - | S | S | h | s | h | s | h | h | - |
| 4m6h.pdb:A | - | - | - | s | H | h | - | S | S | h | s | h | s | h | h | - |
| 4m6i.pdb:A | - | - | - | s | H | h | - | S | S | h | s | h | s | h | h | - |
| 4rn7.pdb:A | - | - | - | s | H | h | - | S | S | h | s | h | s | h | h | - |
| 5emi.pdb:A | - | - | s | h | H | h | - | S | S | h | s | h | s | h | h | - |
| 5j72.pdb:A | - | - | s | h | H | h | - | S | S | h | h | s | h | s | h | h |
| 7agl.pdb:A | s | h | s | s | H | h | - | S | S | h | s | h | s | h | h | - |
| 7agm.pdb:A | s | h | s | s | H | h | - | S | S | h | s | h | s | h | h | - |
| 7ago.pdb:A | s | h | s | s | H | h | - | S | S | h | s | h | s | h | h | - |
| 7b3n.pdb:A | - | - | - | s | H | s | h | S | S | h | s | h | s | h | - | - |
| 7rag.pdb:B | - | - | - | s | H | h | - | S | S | h | s | h | s | h | h | - |
| 7tj4.pdb:B | - | - | - | s | H | s | h | S | S | h | s | h | s | h | h | - |

* + 4BIN and 3NE8 still have what looks like an additional helix (the proposed mobile helix region) – these were the structures from E. coli and Bartonella henselae, both gram negative. However other gram negative (the Thermus parvatiensis AmiP 7B3N and the 3CZX from Neisseria meningitidis) don’t have this additional helix in the alignment. To look into proposed functions of all four, maybe only 3NE8 and 4BIN are AmiC, this might be the identifier for function?
* ClustalOmega alignment of the 19 FASTA sequences and HMMsearch
  + Again, cut the 5 AmiA possibilities as sequences and structures too different
  + Used all default settings
  + Wanted to get an alignment to run through hmmsearch and see if any more orthologs/protein family members could be identified
  + Did without cutting down to the amidase domain first to see what the output would be from hmmsearch
  + HMMsearch found 314 significant hits in the reference proteome database

# 06/06/2023 – Literature Research

* **Amidase\_3 Literature Research**
  + Look up papers for the 24, ensure their cell division function has been confirmed.
  + 4BIN and 3NE8 are AmiC and AmiB amidases, shown experimentally to have the mobile helix region.
  + 3NE8 paper posed 3 functions for LytC-type amidases (NAMLAA): lytic from bacteriophages, lytic from sporulating bacteria, and cell separation amidases (which are the ones I’m after!) And that those with cell separation function have the self-regulating helix, as why would an amidase whose function is to lyse the cell wall to death want to be regulated?
  + From 3NE8 paper: 
  + Clostrium difficle is sporulating, therefore the amidases here which had poor alignment structurally were probably in this functional class alone in the dataset (explains poor TM scores for 7RAG, 5J72)
  + Change of tact: Can we use the sequence data to identify the helix and functionally classify the NAMLA into sub-families? Because ‘NAMLAA’ or Amidase\_3 isn’t enough to distinguish these variety of functions on their own, clearly!

# 08/06/2023 – Existing domain tools

* **Existing tool research**
  + UniProt/SwissProt: Provides assertion method (manual vs automatic), but draws in information from other tools or databases (does not appear to run the tools themselves in domain identification).
    - Includes both experimentally derived evidence and computational evidence. For amidase\_3 (amiC), had just automatic annotation from SMART ‘inferred from signature match’.
  + [InterPro](https://www.ebi.ac.uk/interpro/about/interpro/): Lists as ‘IPR002508’.
    - Incorporates the models from 13 member databases including SMART, adds the GO annotations on top of that
    - I think matches the different signatures and profiles produced from the 13 databases and combines them into a consensus where they are different to assign them a family.
  + SMART: Lists as ‘Ami\_3’ (SM00646), reported in UniProt, Pfam, etc. Family includes phages as well as bacteria (therefore does not distinguish between specific function, although acknowledges phage endolytic and autolytic functional difference in-text – endolytic not having signal peptide and autopytlic having a C- or N-terminal cell wall binding domain).
    - Combines Uniprot, stable ensembl proteomes. Contains 137mill+ proteins from 537,000 species+.
    - 1998 paper: ‘Manually curated hidden Markov models’
    - Uses manual curation to annotate domains (using multiple sequence alignment and selecting a cut-off threshold for similarity to cluster each domain manually).
    - Gapped multiple sequence alignments (encompassing secondary structures of known tertiary structures), found candidate homologues using HMMer and two other iterative tools (MoST and WiseTools) which were statistically significant in their alignments but also includes some homologues experimentally shown to operate in the same biological context.
    - True positive/true negatives manually decided upon for each domain by assigning am SWise similarity score (scoring algorithm) threshold based on ‘published homology arguments’
    - Protein profiles generated from the alignments, then constructed hidden Markov models (assuming to identify patterns in the domains which can be applied to match with new sequences being entered?)
  + ECOD (domain ID e3ne8A1 for Amidase 3, using the protein family Amidase 3). F-class classification still contains both endolysins and autolysins.
    - Can be automatically assigned (was automatic for 3NE8).
    - F-group = domains with significant sequence similarity, mostly mapped Pfam families and some HHsearch clusters.
    - Also provides automatic groupings based on architecture of the structure, possible homologs based on high sequence/structure scores, functional similarity, features, topology.
    - Only uses structures derived in the PDB, emphasises distant evolutionary relationships.
    - Has a classification pipeline that gets more and more specific (starting with the SCOP superfamilies and pairing protein structures based on low lidentity, around 40%), but for F-class, uses hidden markov modelling like SMART does (HMMER and HHsearch-based clustering)
  + CDD/SPARCLE: By the NCBI, architecture ID for NAMLA is 11436722
    - Searches based on protein sequence, uses RPS-BLAST to produce position-specific scoring matrices against a protein query (these are pre-calculated against the conserved domain database)
    - Manually curated at least for the NAMLA architecture
    - Does not distinguish between sporulating Clostridium difficile and the normal function cell cycle autolytic bacteria.
  + GO Annotations
    - A mix of manual annotations and computational predictions for gene products, indicated by evidence code.
    - Manual curation based on papers (functional evidence).
    - 99% are computational annotations: project annotations from one species to another based on orthology (like the EggNOG database), prediction based on curated rules, prediction based on sequence features, mapping of concepts to GO terms(?)
    - Annotations from: InterPro2GO, UniProt Keywords2GO, UniProt Subcellular Location2GO, EC2GO, UniRule2GO, Ensembl and Ensembl Genomes, UniPathway2GO, Gene Ontology Consortium, RNACentral
      * InterPro2GO: Manually curated file by looking between proteins and InterPro and selecting an appropriate GO term for all proteins in the family (using stats in keywords, comments, annotation conservation). 90-100% accurate.
      * Keyword2GO: Manually curated and based on literature/sequence analysis checks (from TrEMBL entries on keywords in nucleotide sequence, and using Spearmint decision tree program and RuleBase rules algorithm). 91-98% accurate.
  + PANTHER
    - N/A for NAMLA, no organism matches in their database, but uses protein placement in a phylogenetic tree to inform classification of sequences (TreeGrafter)
  + PROSITE (N/A didn’t have Amidase 3 as a domain, although describes it under the LytM entry)

*Can I do this better with machine learning? Have identified problem with identification of amidase\_3 region, since other than the key ones identified through sequence and structure alignment, there is a real variety in protein structures and possible functions, and therefore annotations may not be taking advantage of pattern recognition in relation to function. Can we use ML to a) identify patterns in sequence domains, b) use patterns in secondary structures in the domains, and c) use both sequence and structure patterns to more accurately annotate domains? Use NAMLAA as my test case, utilising hypothetical protein structures from AlphaFold and deposited NCBI sequences for bacteria.*

* Put together my research from past two weeks to present to Mark on Friday.

# 18/06/2023 – Sequence analysis, day 1

* **Amidase\_3 Sequence Research – Find sequences (WGS)**
  + *For the sequence-only data: for proteins with function confirmed (ie in septal wall division), search for similarity in sequences in ~~non-redundant NCBI dataset. (blastp)~~ RefSeq WGS*
  + *Phmmer (using both 4BIN and 3NE8, Reference Proteomes)*
  + *Expanded dataset – Isolates (API call from InterPro for seq download) – not non-redundant, includes isolates, leave out as only want reference genome data for now.*

BLASTp searches:

* Search 1: 3NE8 cropped VVLDPGHGGIDGGARGVTGILEKDVTLAFARALRDELQKGSHTIVALTRDSDIFLRLSERVKKAQEFDADLFISIHADTIDVHSLRGATVYTISDEASDAIAKSLAESENKVDLLDGLPKEESLELTDILLDLTRRETHAFSINFANNVVSNLSKSHINLINNPHRYADFQVLKAPDVPSVLIEIGYLSNKEDEKLLNNPQWRKQMAASIAYSIRQ
* Search 2: 4BIN cropped

IMLDPGHGGEDSGAVGKYKTREKDVVLQIARRLRSLIEKEGNMKVYMTRNEDIFIPLQVRVAKAQKQRADLFVSIHADAFTSRQPSGSSVFALSTKGATSTAAKYLAQTQNASDLIGGVSKSGDRYVDHTMFDMVQSLTIADSLKFGKAVLNKLGKINKLHKNQVEQAGFAVLKAPDIPSILVETAFISNVEEERKLKTATFQQEVAESILAGIKA

* Downloaded descriptions csv, aligned sequences, and ‘text’ file
* Searched in RefSeq Select proteins to get reference sequences for each protein
* BLASTp API requires Google Cloud Shell or AWS subscription (free trial incl. entering payment info)

Phmmer: Ran using local REST client rather than web service (as produced too many outputs for download).

* python hmmer3\_phmmer.py --email 'sallen10@student.bbk.ac.uk' --database 'uniprotrefprot' VVLDPGHGGIDGGARGVTGILEKDVTLAFARALRDELQKGSHTIVALTRDSDIFLRLSERVKKAQEFDADLFISIHADTIDVHSLRGATVYTISDEASDAIAKSLAESENKVDLLDGLPKEESLELTDILLDLTRRETHAFSINFANNVVSNLSKSHINLINNPHRYADFQVLKAPDVPSVLIEIGYLSNKEDEKLLNNPQWRKQMAASIAYSIRQ
* python hmmer3\_phmmer.py --email 'sallen10@student.bbk.ac.uk' --database 'uniprotrefprot' IMLDPGHGGEDSGAVGKYKTREKDVVLQIARRLRSLIEKEGNMKVYMTRNEDIFIPLQVRVAKAQKQRADLFVSIHADAFTSRQPSGSSVFALSTKGATSTAAKYLAQTQNASDLIGGVSKSGDRYVDHTMFDMVQSLTIADSLKFGKAVLNKLGKINKLHKNQVEQAGFAVLKAPDIPSILVETAFISNVEEERKLKTATFQQEVAESILAGIKA
* Phmmer output gives alignment (only top n alignments, around 50?) but not sequence ID/accessions, species, or individual sequences. Also doesn’t give all 11,000 sequence alignments with query sequence (scrap, use NCBI BlastP only?)

# 19/06/2023 – Sequence analysis, day 2

Monday

* **Amidase\_3 Sequence Research – Analyse sequences**
  + Identification of other conserved features (Quantitative similarity/multiple sequence alignment)
    - Clustalo: .\clustalo.exe -i 7161\_unique\_sequences.txt -t Protein –outfmt=clustal –output-order=tree-order > output.txt
    - Also output in fasta format
  + Repeated with kalign
    - .\kalign.exe -i .\7161\_unique\_sequences.txt -o output\_fasta.fa

# 20/06/2023 – Annotated Bibliography + Alignments

Tuesday

For annotated biblio:

* Finish and send to Mark for discussion on Friday

# 21/06/2023 – Conservation of sequence (helix identification?)

Analysis of alignments

* Secondary structure predictions?
* Conserved regions?
* Split into gram negative and gram positive in clustal omega output?

Looked into JalView

Need gram positive – blastp for these, same search parameters, for alignment with gram negatives

* VFIDPGHGGNDKGTESKTSNRYEKDLNLQIAKKLANKLSKQKDIQVVVSRTDDTYISLKDRAILANNSSADVLVSIHLNAEKNGNTATGIETWYRNKATDGSKELAQTVQSTIVSYVKVRDRGIVENNFEVLRESNMPAILIECGFLTTPSEEQKIINEKYQDQLAEGIVQGVLS (4rn7)
* VVIDPGHGGKDSGAPGLGGLLEKDVILPIGKRVAAILEQHGVQAVLTRDADFFVELQGRVEIAERVNATAFVSIHANSVDNRPDVNGLEVYYYDSGYALAEVVRNTILQNIDTIKNRGTRKARFYVLRKSSMPSILVETGYMTGREDNPRLASREYQNQMAEAIARGILK (5emi)
* VVIDPGHGGSDSGATSGLNGGAQEKKYTLNTALATTEYLRSKGINVVMTRDTDKTMALGERTALSNTIKPDLFTSIHYNASNGSGNGVEIYYKVKDKNGGTTKTAASNILKRILEKFNMKNRGIKTRTLDNGKDYLYVLRNNNYPAILVECAFIDNKSDMDKLNTAEKVKTMGTQIGIGIED (5j72)
* VVLDPGHNGANDSSINNQVPDGRGGTKSCQTSGTATDGGYPEHTFTWNTVLLIRQQLTQLGVRTAMTRGDDNKLGPCIDKRAEIENSYNPDAVVSIHADGGPAGGHGFHVNYSNPPVNAVQGEPTLRFAKTMRDSLQAAGLTPATYIGTGGLYGRSDLAGLNLAQHPKVLVELGNMKNAQDSAMMTSPEGRSKYAQAVVQGIVA (7agl)
* VFLDPGHNGANDASIGRQVPTGRGGTKNCQESGTATDDGYPEHSFTWDTTLRVRAALTALGVRTAMSRGNDNALGPCVDERAAMANSLRPHAIVSIHADGGPPTGRGFHVLYSSPPLNAAQSGPSVQFAKVMRDQLAASGIPPATYIGQGGLNPRSDIAGLNLAQFPSVLVECGNMKNPVDSALMKSPEGRQKYADAIVRGIAG (7agm)
* IILDAGHGGIDPGALNKDKSTSEKDINLAITLKLRELIESSGGLVILTREDDSSLYKEENNKTTRQKYNENLKNRKEIISNSNANMFVSIHLNAFEQSKYYGAQTFYPKDKQDSKELSKCIQEELKRVVDKTNNREVKPRDDIYLLKDNNIPSVLIECGFLSNEKECKLLTDETYQEKIAWAIYIGIQK (7rag)
* LQGKTIVLDPGHGGSDQGASSNTKYKSLEKDYTLKTAKELQRTLEKEGATVKMTRTDDTYVSLENRDIKGDAYLSIHNDALESSNANGMTVYWYHDNQRALADTLDATIQKKGLLSNRGSRQENYQVLAQTKVPAVLLELGYISNPTDETMIKDQLHRQILEQAIVDGLKIYFSA (7tj4)
* Repeated with kalign

# 27/06/2023 – Link genus/species to tax ID

Using Batch Entrez API call against genus/species

Future research

* K-means clustering for RMSD/Q-score (structural analyses visuals)

Alignments

* Download whole genomes for these bacteria
* Alignment of whole genomes, not just the amidase\_3 for the identified bacteria. Any other possible conserved genes?
* **LytM Factors for AmiC**
  + ActS: Any structures of this? Other papers around function? How might the docking/binding of this be modelled (could compare to NlpD as this is more established as a cofactor in AmiC).
  + DolP: The same as above, evidence less confident for this factor however.
* **Programmatic tasks**
  + Identification of the amidase\_3 domain from sequences (for future cropping, multiple sequence alignment, or searching). Identify conserved regions across amidase\_3 proteins, create a pattern recognition package for these regions? If have more evidence that the helix is a recognition/conserved sequences in gram negative autolytic ‘Group A’
  + R package for k-means clustering in structures? (would need to incorporate AlphaFold predictions as dataset is incomplete currently?)

# 25/02/2023 – Starting again with data searching (experimental structures only)

* FoldSeek: Looking in GMGCL 2204 (global microbial gene catalog from EMBL), MGnify-ESM30 v1 (from EMBL-EBI, microbiome data, structures from AlphaFold?), PDB100 220122
  + Using 3di/AA smith waterman local pairwise alignment
* Look in GMGC v1.0 separately, see if it picks up anything else (<https://gmgc.embl.de/>)
  + From the paper: “non-redundant gene catalogue of 303 million species-level genes (clustered at 95% nucleotide identity) from 13,174 publicly available metagenomes across 14 major habitats and use it to show that most genes are specific to a single habitat. The small fraction of genes found in multiple habitats is enriched in antibiotic-resistance genes and markers for mobile genetic elements. By further clustering these species-level genes into 32 million protein families, we observed that a small fraction of these families contain the majority of the genes (0.6% of families account for 50% of the genes). The majority of species-level genes and protein families are rare. Furthermore, species-level genes, and in particular the rare ones, show low rates of positive (adaptive) selection, supporting a model in which most genetic variability observed within each protein family is neutral or nearly neutral.”
* Look in MGnify separately as well