**Evolutionary divergence and functional investigation of NAMLAA amidases – bacterial proteins essential to cell division.**

Sophie Rose Allen

Supervisor: Dr Mark Williams

MSc Bioinformatics (Part-Time, 2023-2024)

Birkbeck, University of London

# Abstract (200-300 words)

A

B

C

(if I can change the title perhaps):

“The case for splitting NAMLAA domain classification: how a helix insertion in the AmiC protein leads to difference in function across bacterial species”

Contents

[Abstract (200-300 words) 2](#_Toc133861786)

[Acknowledgments 4](#_Toc133861787)

[Abbreviations 5](#_Toc133861788)

[Introduction 6](#_Toc133861789)

[Materials and Methods 7](#_Toc133861790)

[Results 8](#_Toc133861791)

[Discussion 9](#_Toc133861792)

[Conclusion 10](#_Toc133861793)

[References 11](#_Toc133861794)

[Appendices 12](#_Toc133861795)

# Acknowledgments

I would like to thank and acknowledge my supervisor Dr Mark Williams for all the ongoing assistance and hugely helpful sessions throughout this project to guide both tool suggestions and the direction of analysis.

# Abbreviations

AmiC: Amidase C bacterial protein

NAMLAA: N-acetylmuramoyl-L-alanine amidase

RMSD: Root-Mean-Square Deviation

PDB: Protein Data Bank

# Introduction

Intro: What NAMLAA amidases are, what they typically do in gram negative bacteria, the biochemical pathway involved, proteins involved, focus in on AmiC, focus in again on how it has been observed in gram positive bacteria and possibly other organisms, query why this is the case, and lead to bioinformatic work to identify the structures and sequences, and if there are functional differences in the AmiC protein family (as posited by the 3NE8(?) paper and CATH database)

* From annotation bibliography: 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?)
  + 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)
* 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!
* CATH (<http://www.cathdb.info/version/v4_3_0/superfamily/2.60.40.3500/alignments>)
  + That link shows functional family separation across NAMLAA proteins, and seem to have different categories for gram negative and gram positive? Evidence that I’m going in the right direction here?
* Existing domain tools
  + 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: 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)

# Materials and Methods

The code used for all analysis can be found at the following GitHub repository: <https://github.com/sroseallen/MSc_Project> (scripts are saved in the MSc\_Project/05\_SCRIPTS folder within the repository, and all collected and derived data saved under MSc\_Project/01\_DATA).

## Data Collection

### Identification of candidate AmiC structures

The FASTA sequence for NAMLA amidase from *Escherichia coli* was downloaded from the Protein Data Bank (PDB)[ref BLASTp PDB, ref 4BIN structure]. This sequence was used in three separate structure searches on 17/05/2023: a BLASTp search restricted to the PDB[ref BLASTp], a PDB search using the search term ‘amidase\_3’, and a SCOP[ref SCOP] search using identifiers 4001130 (N-acetylmuramoyl-L-alanine amidase-like family) and 8092183 (AmiC domain). Two further searches were performed on 19/05/2023 in Pfam (identifier PF01520) and Interpro (search term amidase\_3), however these did not identify any new experimental structures[ref Pfam, ref Interpro]. A final wider search for structures was performed in the PDB using the GO annotation ‘N-acetylmuramoyl-L-alanine activity’ and search terms ‘AmiA’, ‘AmiB’ and ‘AmiC’ on 19/05/2023. The full longlist of experimental structures identified from these searches are listed in Appendix A.

Proteins were then shortlisted if they had a Pfam annotation of ‘amidase\_3’, a GO functional annotation of ‘N-acetylmuramoyl-L-alanine activity’, and an Interpro annotation of ‘Amidase\_3 domain’. BLASTp was then run using the PDB FASTA sequence for all structures in this shortlist (default parameters, against PDB database) , and an additional structural alignment for all shortlisted structures using PDBeFOLD[ref PDBeFOLD] (parameters: against the entire PDB archive, lowest acceptable match of secondary structure alignment=50%, Q-score >0.3, match individual chains and connectivity, unique and best matches only, Normal precision) to ensure all candidate AmiC proteins were included. Structures were not added to the shortlist if alignment was to another chain in a model already on the list, or if the P-score was above 3 (as these are considered statistically insignificant per the PDBeFOLD website[ref for this]).

### Identification of candidate AmiC sequences

## Sequence Analysis

## Structural Analysis

### Initial candidate structural alignment

To visualise the alignment of the shortlisted structures, the MatchMaker module of UCSC Chimera was run with the following parameters: secondary structure predictions not re-calculated, pairwise alignment, reference structure 4BIN in *E. coli*, Needleman-Waltsch alignment. Multiple structure alignment was carried out using multiple alignment in PDBeFOLD[ref PDBeFOLD]. Confirmatory alignment of all structures to all other structures in the shortlist was carried out with a local download of US-Align[ref USalign] using the below command:

./USalign -dir chains\_amidase\_domain\_only/ chain\_list.txt -mm 4 -o sup -full T > output.txt

Structure/sequence cropping was carried out in UCSC Chimera (commands used are saved in the project GitHub repository SCRIPTS folder), based on the Pfam boundaries of the ‘amidase 3’ domain as listed in the PDB.

## Domain Fusion Analysis

## Co-factor Binding Analysis

## Statistics

# Results

## Generating an AmiC structure shortlist

From the initial structural longlist, a final shortlist of 19 candidate AmiC structures was curated, which matched on Pfam annotation, GO annotation, and Interpro domain annotation of amidase\_3 (Table 1). Two of these structures (4M6H and 4M6I) are the same protein sequence from the same bacterium, however both were retained as one structure reflects the metal-bound form of the amidase, which may affect protein structure and may therefore be relevant in subsequent analyses. In creation of the shortlist, it is notable that a search within PDBeFOLD using the structure 4BIN had a low (<=17%) percentage of alignment of residues based on alignment in space for all identified structures. Of note, the species in this shortlist varied greatly, consisting of a mixture of gram-positive and gram-negative bacteria as well as bacteriophages, but tend to be reported with a zinc ion cofactor.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **NAMLA amidase family** | **Organism** | **Gram status** | **Name of the structure in PDB** | **PDB ID** | **PDB**  **chain** | **Year in PDB** | **Ligand** | **Identified in** |
| Amidase 3 | *Escherichia coli* | - | E. coli N-acetylmuramoyl-L-alanine amidase AmiC | [4BIN](https://www.rcsb.org/structure/4BIN) | A | 2013 | Zn2+ ,Na+ | PDB search for AmiC |
| Amidase 3 | *Thermus parvatiensis* | - | AmiP amidase-3 from Thermus parvatiensis | [7B3N](https://www.rcsb.org/structure/7B3N) | A | 2020 | Zn2+, Na+, Cl-, Glycerol, SO42-, Ethanesulfonic Acid | BLASTp search 1 (restricted to PDB, FASTA seq. search of 4BIN) |
| Amidase 3 | *Neisseria meningitidis* | - | putative N-acetylmuramoyl-L-alanine amidase from *Neisseria meningitidis* | [3CZX](https://www.rcsb.org/structure/3CZX) | A | 2008 | Zn2+ | PDB search of Amidase\_3 protein family annotation |
| Amidase 3 | *Clostridium difficile* | + | Cwp6 from Clostridium difficile | [5J72](https://www.rcsb.org/structure/5J72) | A | 2016 | Citric Acid, Zn2+, Ca2+, Cl-, Na+ | BLASTp search 1 (restricted to PDB, FASTA seq. search of 4BIN) |
| Amidase 3 | *Staphylococcus aureus* | + | Structure of the S. aureus amidase LytH and activator ActH extracellular domains | [7TJ4](https://www.rcsb.org/structure/7TJ4) | B | 2022 | Zn2+ | BLASTp search 1 (restricted to PDB, FASTA seq. search of 4BIN) |
| Amidase 3 | *Clostridium difficile* | + | The crystal structure of N-acetylmuramoyl-L-alanine amidase from Clostridium difficile 630 | [4RN7](https://www.rcsb.org/structure/4RN7) | A | 2014 | Zn2+, Formic Acid, Glycerol, EPE (Ethanesulfonic Acid) | BLASTp search 1 (restricted to PDB, FASTA seq. search of 4BIN) |
| Amidase 3 | *Nostoc punctiforme* | + | N-acetylmuramoyl-L-alanine amidase AmiC2 of Nostoc punctiforme | [5EMI](https://www.rcsb.org/structure/5EMI) | A | 2015/  2016 | MES, MRD, Zn2+ | BLASTp search 1 (restricted to PDB, FASTA seq. search of 4BIN) |
| Amidase 3 | *Bartonella henselae* | - | NAMLA amidase of *Bartonella henselae* str. Houston-1 | [3NE8](https://www.rcsb.org/structure/3NE8) | A | 2010 | ACETATE ION, FORMIC ACID, GLYCEROL, Zn2+ | BLASTp search 1 (restricted to PDB, FASTA seq. search of 4BIN) |
| Amidase 3 | *Mycobacterium tuberculosis* | NA | *Mycobacterium tuberculosis* peptidoglycan amidase Rv3717 in complex with L-Alanine-iso-D-Glutamine reaction product | [4M6G](https://www.rcsb.org/structure/4M6G) | A | 2013 | Zn2+ | PDB search of Amidase\_3 protein family annotation |
| Amidase 3 | *Bacillus(* *Paenibacillus polymyxa)* | + | Catalytic domain of CwlV, N-acetylmuramoyl-L-alanine amidase from *Bacillus(Paenibacillus) polymyxa var.colistinus* | [1JWQ](https://www.rcsb.org/structure/1JWQ#entity-1) | A | 2001 | Zn2+ | BLASTp search 1 (restricted to PDB, FASTA seq. search of 4BIN) |
| Amidase 3 | *Mycobacterium tuberculosis* | NA | Crystal structure of Rv3717 reveals a novel amidase from M. tuberculosis | [4LQ6](https://www.rcsb.org/structure/4LQ6) | A | 2013 | SO4-, Zn2+, Cl-, Pt2+ | PDB search of Amidase\_3 protein family annotation |
| Amidase 3 | *Mycobacterium tuberculosis H37Rv* | NA | Structure of the reduced, metal-free form of Mycobacterium tuberculosis peptidoglycan amidase Rv3717 | [4M6H](https://www.rcsb.org/structure/4M6H) | A | 2013 | None reported | PDB search of N-acetylmuramoyl-L-alanine activity (GO annotation) |
| Amidase 3 | *Mycobacterium tuberculosis* | NA | Structure of the reduced, Zn-bound form of Mycobacterium tuberculosis peptidoglycan amidase Rv3717 (metal-free structure: [4M6H](https://www.rcsb.org/structure/4M6H)) | [4M6I](https://www.rcsb.org/structure/4M6I) | A | 2013 | Zn2+ | PDB search of Amidase\_3 protein family annotation |
| Amidase 3 | *Mycobacteroides abscessus* | + | crystal structure of the N-acetylmuramyl-L-alanine amidase, Ami1, from Mycobacterium abscessus bound to L-Alanine-D-isoglutamine | [7AGO](https://www.rcsb.org/structure/7AGO) | A | 2020 | Zn2+, D-alpha-glutamine, Alanine | PDB search of Amidase\_3 protein family annotation |
| Amidase 3 | *Mycolicibacterium smegmatis* | + | Crystal structure of the N-acetylmuramyl-L-alanine amidase, Ami1, from Mycobacterium smegmatis | [7AGM](https://www.rcsb.org/structure/7AGM) | A | 2020 | Zn2+ | PDB search of Amidase\_3 protein family annotation |
| Amidase 3 | *Mycobacteroides abscessus* | + | crystal structure of the apo form of the N-acetylmuramyl-L-alanine amidase, Ami1, from Mycobacterium abscessus. | [7AGL](https://www.rcsb.org/structure/7AGL) | A | 2020 | Zn2+ | PDB search of Amidase\_3 protein family annotation |
| Amidase 3 | *Clostridium difficile* | + | Structure of the CwlD amidase from Clostridioides difficile in complex with the GerS lipoprotein | [7RAG](https://www.rcsb.org/structure/7RAG) | B | 2021 | Zn2+, 1,2-Ethanediol | PDB search of Amidase\_3 protein family annotation |
| Amidase 3 | *Listeria monocytogenes* | NA | *Listeria monocytogenes* bacteriophage PSA endolysin PlyPSA | [1XOV](https://www.rcsb.org/structure/1XOV) | A | 2004 | Cl- ,GLUTAMIC ACID, LYSINE, 2-AMINO-2-HYDROXYMETHYL-PROPANE-1,3-DIOL, SO4-, Zn2+ | PDB search of Amidase\_3 protein family annotation |
| Amidase 3 | *Colneyvirus CD27* | + | Catalytic domain of CD27L endolysin targeting *Clostridium difficile* | [3QAY](https://www.rcsb.org/structure/3QAY) | A | 2011 | SO4-, Zn2+ | PDB search of Amidase\_3 protein family annotation |

**Table 1: Shortlist of 19 candidate experimental structures of NAMLAA AmiC across different bacterial species.**

APRIL: AMIC SEQUENCE ANALYSIS

Evidence 1: sequence only clustering (multiple sequence alignment work)

* Making sure I have all sequences possible
  + 3NE8 and 4BIN were chosen because they were identified from structures in PDB and were both on the shortlist of amidase\_3 proteins from the initial structure search. When the initial 19 structures on the shortlist were aligned with PDBeFOLD and USAlign, these two proteins were the only ones with this apparent helix insertion region, and so this was used as a crude identifier for amidase\_3 function in the cell wall (based on inferences from literature and specifically the 3NE8 paper which had an early sequence alignment showing the helix was present in cell wall proteins and not in other amidase\_3 with functions like in Clostriudium difficile ie sporulating bacteria).
    - The theory at this stage was that the amidase\_3 in gram negative bacteria were intrinsically different to the amidase\_3 in gram positive bacteria, hypothesis being that function in gram negative bacteria was cell separation (in an autolytic cell cycle mechanism which requires an additional helix to ‘switch off’ the amidase activity when not in cell separation phase to avoid killing the parent cell), and in gram positive it had one of two roles (to assist in sporulation of child cells for sporulating bacteria in an autolytic cell cycle, or to kill parent cells upon infection for infectious bacteria acting in an endolytic fashion, both of which do not need a ‘switch’ as the protein needs to be on all the time to kill the parent cell). Additional group might be identified where amidase\_3 is in bacteriophages acting in the same pathway as the gram\_positive endolytic pathway to kill the target bacterial cell.
    - So, initial searches for sequences were based on BLASTp for 4BIN and 3NE8 only, which were both gram negative and had that helix in the structure alignment, to try to get a large ‘truthset’ of cell wall separation amidases with the helix region. The search was then expanded to the amidase\_3 structures found in gram\_positive bacteria (there were 7 of these). However when writing up need to talk about how this search was done for all 19 of the shortlist.
      * Needed to perform the searches on the other bacterial sequences not included in this search unless I can find a good reason to exclude them, as this is biasing my results (by just looking at what I’m expecting to find – need to include the Mycobacterial AND bacteriophage sequences as well, because it would be interesting to see if there is a clear separation in the alignment and to see which features they cluster with – do the bacteriophages indeed cluster with the gram positive infectious species? The sporulating bacteria perhaps?)
  + Searching done using BLASTp for all 19 protein sequences
    - **Search again with BLASTp**
      * Using cropped versions of the sequences, fasta files
      * Search set: RefSeq Select (chosen over the non-redundant database, as RefSeq select had one copy of a protein for each prokaryotic species, and as such thought this might be more appropriate for pulling in aligned proteins from other species without over-saturating or introducing bias from the non-redundant dataset, which includes several proteins from the same species eg many ‘Enterobacteriaceae’ and lots of E.coli sequences when searching 4BIN with the non-redundant database). 59610710 sequences in RefSeq Select database. Not restricted to an organism, did not exclude models, non-redundant RefSeq proteins, or environmental samples. Want to cast a broad net as I know identity will likely be low due to the unique N-terminal regions of these proteins.
        + Thought about a search on the experimental clustered\_nr, which is a search on a version of the nr database clustered at 90% identity and 90% sequence length, so searches pull one representative sequence from each cluster in the alignment. Uses MMseqs2 software, idea is this would pull in more taxonomic depth than the nr dataset. However, for this research, I want sequences aligned from as many species as possible, and the clustering will happen at a later stage. Don’t want to miss out any species with this clustering.
      * Algorithm: blastp (PSI-BLAST and DELTA-BLAST build a PSSM which I don’t want as each search is just for one sequence, I want to do a multiple sequence alignment for all the options later on; QuickBLASTp works best if % identity is high and I’m assuming it will be low based on my searching with a cropped protein sequence; PHI-BLAST looks at the alignment surrounding the match, so increasingly the likelihood of homology in the matches, by using a specific pattern. However I don’t know which pattern would be best to identify ‘true’ AmiC homologues, as this is partly the point of the exercise, so stuck with bastp for now.)
      * Parameters: Standard (word size 0.05 (95% significance), word size 5, no restricting on maximum matches in the query range, BLOSUM62 matrix, gap costs existence 11, extension 1 (low gap costs as expecting gaps). Generated top 5000 results (maximum possible significant alignments through web interface; API requires paid subscription service)
      * Downloaded FASTA aligned sequences, and ‘text’ file which summarises the top alignments
        + Aligned sequences because I want to specifically look at how the amidase domain itself differs between species, not look at this stage at other domains in the entire protein, as will likely align according to N-terminal domains rather than the contents of the amidase domain.
      * 1JWQ: top 5000. 1XOV: 1066. 3CZX: 185. 3NE8: top 5000. 3QAY: 463. 4BIN: top 5000. 4LQ6: 2114. 4M6G: 2114. 4M6H: 2114. 4M6I: 2114. 4RN7: top 5000. 5EMI: top 5000. 5J72: top 5000. 7AGL: 1942. 7AGM: 2435. 7AGO: 1942. 7B3N: top 5000. 7RAG: top 5000. 7TJ4: top 5000.
      * Total sequences identified from BLASTp search (pre-filtering, may contain duplicates): 61,489
      * Used the script ‘unique\_blast\_seqs.py’ to combine these outputs and remove any duplicates. Total sequences after duplicate identified sequences removed: **38,687**
    - **Search again with phmmr**
      * For methods: Commands used are in ‘project notes’, day 18/06/2023
      * 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. Therefore, decided not a useful source of data for gathering sequences.
    - **Search with InterPro** and get all sequences
      * InterPro contains CATH-Gene3D (CATH for complete genomes), Pfam, CDD (multiple alignment databases), PROSITE, SMART, and others.
      * Search A: Search by text ‘Amidase\_3’.
        + Identified IPR019606 – GerMN domain: ‘The GerMN domain is a region of approximately 100 residues that is found, duplicated, in the Bacillus GerM protein and is implicated in both sporulation and spore germination. It is also found in lipoprotein LpqB. The domain is present in a number of different bacterial species both alone and in association with other domains such as Gmad1 and Gmad2. It is predicted to have a novel α-β fold.’ **Could be a signifier of function if NAMLAA domain is found with this domain in a protein?**
        + IPR002508: N-acetylmuramoyl-L-alanine amidase.
      * Search B: Browse by Entry/InterPro, search ‘N-acetylmuramoyl-L-alanine amidase’
        + 1 superfamily: [IPR036505](https://www.ebi.ac.uk/interpro/entry/InterPro/IPR036505/) (PGRP domain superfamily) (associated with type-2 bacterial amidases, with antibacterial activity).
        + 6 families: [IPR010846](https://www.ebi.ac.uk/interpro/entry/InterPro/IPR010846/) (AmiA-like), [IPR014234](https://www.ebi.ac.uk/interpro/entry/InterPro/IPR014234/) (CwlD, to do with endospores) [IPR017293](https://www.ebi.ac.uk/interpro/entry/InterPro/IPR017293/) (SH3-domain containing protein; binding to other proteins), [IPR049745](https://www.ebi.ac.uk/interpro/entry/InterPro/IPR049745/) (AmiC NAMLAA), [IPR050695](https://www.ebi.ac.uk/interpro/entry/InterPro/IPR050695/) (amidase\_3), [IPR051206](https://www.ebi.ac.uk/interpro/entry/InterPro/IPR051206/) (amidase\_2)
        + 4 domains: [IPR002502](https://www.ebi.ac.uk/interpro/entry/InterPro/IPR002502/) (proteins with Amidase\_2 domain), [IPR002508](https://www.ebi.ac.uk/interpro/entry/InterPro/IPR002508/) (proteins with amidase\_3 domain), [IPR021976](https://www.ebi.ac.uk/interpro/entry/InterPro/IPR021976/) (CBD\_PlyG, cell wall binding domain), [IPR048586](https://www.ebi.ac.uk/interpro/entry/InterPro/IPR048586/) (“This domain is found in the putative N-acetylmuramoyl-L-alanine amidase Cwp6 from Clostridioides difficile (Q183L9) and similar proteins from Clostridia.”)
      * Sequences downloaded from proteins tab of IPR002508 (proteins with amidase\_3 domain) – **70617 protein sequences.**
        + [IPR050695](https://www.ebi.ac.uk/interpro/entry/InterPro/IPR050695/) (amidase\_3), [IPR049745](https://www.ebi.ac.uk/interpro/entry/InterPro/IPR049745/) (AmiC NAMLAA), [IPR014234](https://www.ebi.ac.uk/interpro/entry/InterPro/IPR014234/) (CwlD, to do with endospores) are all relevant protein families, but include the other domains, not just the amidase\_3 domain. IPR002508 is the domain specific entry.
        + These are not either a) associated with AmiA (a different protein), b) associated with amidase\_2 domains (a different domain, where I’m focussing on amidase\_3 from AmiC), and c) apparently catalytic domains rather than binding domains
    - Then combine the different searches and remove duplicate sequences (based on duplicate NCBI IDs
      * Duplicates from BLASTp removed with the python code in ‘unique\_blast\_seqs.py’
      * Interpro sequences: Most of these are unreviewed and from WGS shotgun sequencing, and so the data is preliminary, may not be 100% accurate, and annotated with caution in UniProt. Level of strictness applied to the sequences from here as a result; therefore, if taxid matches a species already identified from BLASTp, then the sequence is not added to the master list of AmiC/amidase\_3 sequences.
      * Interpro sequences also needed cropping as the download is for the entire protein, not just the domains, so need to crop to match BLASTp output or this will affect the alignment and align N-termini together instead of amidase\_3 domains.
      * Used API call adapted from InterPro download API page for IPR002508 (saved as “interpro\_download\_seqs.py”) to gather domains from Uniprot gathered under the correct domain in InterPro and add any non-duplicates to the BLASTp list of sequences.
        + Added source organism name to the API call header
        + Extracted the boundaries for the domain for each protein, and crop the FASTA sequence according to that domain.
        + Identify all taxids for sequences in the BLASTp file using the ‘ncbi\_taxid\_call.py’ script
        + Removed any sequences with taxids from the InterPro file which have a match in the BLASTp file
        + **Final sequence number:**
* Check alignment - re-align, run through annotation and filtering scripts
  + <https://www.nature.com/articles/s41598-019-56499-4#Sec8> – reference to read, describes a workflow for getting a good alignment for a similarly low-identity protein family, includes steps like making sure conserved motifs were aligned across most sequences.
    - For NAMLAA, what would these motifs be? Can I pick them out using structural alignment?
    - <https://github.com/evolbioinfo/goalign> - tool for editing MSAs
  + Has a validation methods as well using all-vs-all pairwise structural alignments of proteins with known structures, using FATCAT rigid mode. For each sequence pair between two aligned sequences, they calculate the number of aligned residue pairs that are in both the sequence alignment and the structural alignment. True positive rate is then the ratio between the number of these correctly aligned pairs and the general number of aligned residues in the structural alignment. Predictive positive value (PPV) is the ratio between the number of correctly aligned pairs and the general number of aligned residues in the sequence alignment.
* Annotate sequences with secondary structure data?
  + SPOT-1D-LM is back up again, is this a viable option for annotation?
* Conversion of sequence features of possible interest to classification model (for each in/del region – feature - add % of non-gapped base coverage in the alignment. Also do for secondary structures in the model.)
  + Need to note what all the features are
* Use PCA to cluster sequences based on the features identified from the MSA
  + And from this, need to note what all the SIGNIFICANT features are, ie which principal components were used to cluster the sequences, and hopefully the top (or high ranking) feature will be the mobile helix of interest, but if not that could also be interesting – is there another feature of the structure which is at play in invoking these different functions?
* What do the clusters look like? Do ‘gram negative’ and ‘gram positive’ cluster differently? Do species cluster together? Where do the bacteriophage and mycobacterium species cluster? (to bring up in discussion in more detail as well)
* Species phylogeny (some sort of tree perhaps?), do clusters align with the phylogenetic tree in terms of where the sequences were clustered? (second check, does my PCA model make sense?)
  + <https://www.nature.com/articles/s41598-023-47496-9> - phylogeny for thousands of sequences
  + Also <https://www.nature.com/articles/s41598-019-56499-4#Sec8> – phylogeny workflow including how to bootstrap the alignment and create a distance matrix
    - <https://itol.embl.de/tree/13124980200338471572642166> - nice presentation of phylogeny here, can I do something similar?
* If I can find an existing clustering algorithm for protein sequences, use this, do the results match up with the PCA model? Are the same patterns seen? (confirmatory results to prove the features I identified were relevant and useful in separating these different proteins out)
  + <https://huggingface.co/blog/AmelieSchreiber/plm-persistent-homology-msa-replacement> - clustering algorithm which doesn’t involve MSAs
    - <https://genome.cshlp.org/content/33/7/1145> - paper which I can cite for this methodology but blog puts things in a slightly more readable term for me
  + <https://www.nature.com/articles/s41586-023-06832-9> - AFCluster algorithm which clusters protein sequences on large-scale (paper uses alphafold database to show confirmation differences but also describes this clustering algorithm)
    - <https://github.com/HWaymentSteele/AF_Cluster> - github page for using the clustering algorithm

MAY-JUNE: AMIC PROTEIN STRUCTURAL ANALYSES

Evidence 2a: structural clustering (from experimental data)

## Candidate structural alignment

An initial alignment to visualise similarity between these 19 structures showed structural variation in the N-terminal domain of all proteins, with these domains being unique to each protein and little similarity between species or gram-staining status (Appendix B). Specifically, 5J72 (*C. difficile*) had a large unique CW-2 binding domain, and the bacteriophage targeting *S. aureus* also had a unique disordered region. When cropping each protein to the C-terminal amidase\_3 domain only, the domain appeared to be highly conserved, with the exception of an additional α-helix present for some structures (Figure 1).

**Figure 1: Chimera visualisation of alignment of the 19 shortlisted structures.** The additional α-helix has been coloured in red, and is present for the following PDB structures in the alignment: 4BIN, 3NE8

[to paste here]

The 19 cropped candidate structures were then aligned for more detailed examination in PDBeFOLD[ref PDBeFOLD]. As the visualisation in Chimera showed high variability outside of the amidase\_3 domain, these structures were cropped to this domain only before alignment. Overall, the RMSD demonstrated a close alignment of the carbon alpha backbone across the 19 structures (RMSD = 1.655), however a low Q-score (0.3561), which may reflect a longer alignment length from additional gaps. This is an improvement from the multiple alignment performed with the uncropped structures (RMSD = 1.5525, Q-score 0.1255)(Appendix C). Notably, when examining the alignment of these structures, 4BIN and 3NE8 both had an additional helix structure which was absent in the alignment from the other structures (Figure 2), reflecting the visualisation of the alignment. These structures were both from gram-negative bacteria, however structures from other gram-negative bacteria (7B3N and 3CZX) lack the helix.

**Figure 2: Alignment of the 19 structures in PDBeFOLD.** The insertion in 3NE8 and 4BIN has been highlighted in yellow.

A white background with black lines

Description automatically generated

Confirmatory alignment with USalign produced an average TM-score of 0.83417 for the 19 cropped structures, with all structures having an individual TM-score above 0.8 with the exception of three structures from gram-negative bacteria; 4BIN, 3NE8, and 3CZX (Table 2). This is an improvement over the uncropped structures, with an average TM-score of 0.58520 (Appendix C).

**Table 2: PDBeFOLD and USalign scores from multiple alignment of the 19 cropped structures.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Structure** | **PDBeFOLD RMSD** | **PDBeFOLD Q-score** | **USalign TM-score** |
| 3NE8 | 1.1502 | 0.5314 | 0.74304 |
| 4BIN | 0.9537 | 0.5699 | 0.77128 |
| 3CZX | 1.7016 | 0.5697 | 0.79843 |
| 7B3N | 1.7498 | 0.5651 | 0.85630 |
| 1JWQ | 0.9662 | 0.6479 | 0.88316 |
| 5J72 | 1.2414 | 0.6005 | 0.83297 |
| 4RN7 | 0.7880 | 0.6842 | 0.88708 |
| 7RAG | 1.0864 | 0.5987 | 0.81271 |
| 3QAY | 1.8619 | 0.5341 | 0.81289 |
| 7AGO | 0.7946 | 0.5863 | 0.81590 |
| 7AGL | 0.7904 | 0.5867 | 0.81593 |
| 7AGM | 0.7998 | 0.5858 | 0.81706 |
| 5EMI | 0.9011 | 0.6906 | 0.88620 |
| 7TJ4 | 1.1036 | 0.6442 | 0.88514 |
| 1XOV | 1.5857 | 0.5955 | 0.82427 |
| 4M6G | 0.7719 | 0.5885 | 0.81727 |
| 4LQ6 | 0.7409 | 0.5914 | 0.81818 |
| 4M6I | 0.7515 | 0.6882 | 0.88625 |
| 4M6H | 0.7498 | 0.6693 | 0.87099 |

* Get all structures using lecture 9 search features
* FoldSeek to cluster structures (predicted and experimental)
  + Quick note: searched 3NE8 in PDB on Foldseek and top results all matched the 18 structures identified last year, all amidase\_3, then a sudden shift to other enzymes while maintaining high alignments
* Functional link? (see if hypothesis holds true?)

Evidence 2b: structural clustering (from in-silico data, Alphafold)

* FoldSeek to identify alphafold structures
* Get these from interpro also
* Data sense check: Do these structures have the same patterns as expected from experimental data? What does the alphafold quality score say? (have a threshold to remove structures if quality not good enough)
* Integrate experimental data with in silico data and re-cluster

Evidence 2c: sequence to structure and cluster assignment

* Generate structures from the identified sequences with alphafold
* Assign these to clusters – any which don’t go into suitable cluster, remove, as unlikely to be part of similar family. Endpoint: have a full truthset of sequences which are ‘true’ AmiC amidases

Evidence 2d: Structural clusters link to different AmiC function

* Any with do go to clusters: does this match the sequence clustering? What sort of patterns are seen? Is it the same as sequences? (hypothesis: separation of gram positive and negative to backup the observation by the 3NE8(?) paper and CATH family separation of NAMLAA)
* Look at the structural and sequence differences between representatives from each cluster, what are these? What might they do? (eg that helix insertion in gram negative bacteria), link to discussion and try to find papers which describe possible function of these regions.

JUNE: Evidence 3: Cofactor binding

* Need to do

JULY?: Evidence 4: domain fusion

* Need to do
* Start with AMIN domains
  + 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)
  + 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.)
  + 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
  + 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)
  + Identified some related domains which aren’t NAMLA but are listed under that GO annotation in PDB (N-acetylmuramoyl-L-alanine activity):
    - [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?)
* Return to the shortlist of 19 and look at the N-terminal domains – what are these? What do they do?
* Use CATH searches and funfam matches to identify these
* Different domain fusions = different functions and different classes? Different clusters?
* From InterPro search for ‘Amidase\_3’ 09/04/2024:
  + IPR019606 – GerMN domain: ‘The GerMN domain is a region of approximately 100 residues that is found, duplicated, in the Bacillus GerM protein and is implicated in both sporulation and spore germination. It is also found in lipoprotein LpqB. The domain is present in a number of different bacterial species both alone and in association with other domains such as Gmad1 and Gmad2. It is predicted to have a novel α-β fold.’ **Could be a signifier of function if NAMLAA domain is found with this domain in a protein?**
  + [IPR021976](https://www.ebi.ac.uk/interpro/entry/InterPro/IPR021976/) (CBD\_PlyG, cell wall binding domain), found with InterPro search for N-acetylmuramoyl-L-alanine amidase

# Discussion

Discussion

* Sequence/structure searching: Low % in alignment (low sequence identity), and low similarity matches in PDBeFOLD, possibly due to the N-terminal regions and uniqueness of these? (expand on this and what the different N-terminal domains were, and possible functional differences).
* AmiA differences: Domain annotation, but also when including these in the PDBeFOLD initial alignment, the RMSD and Q-score become worse (RMSD: 2.858, Overall Q-score: 0.01216). There is also difference in the secondary structure, which perhaps causes the worse quality scores, and also notable that these did not appear in initial similarity searches and do not have a zinc ion ligand listed in their PDB entries list the others, instead having a sodium ion. What might this mean?
* 5J72: Longer N-terminal domain than the others, no matches of significance when searched in PDBeFOLD, what is this? Why is it so different? Bacteriophage amidases coming with a long N-terminal domain (5J72), to bind to cell wall?
* Gram positive/negative splitting
* Insertion region identification and what that means in relation to the clusters the insertions were found in
* Bring in studies on the helix being used to block the enzyme activity when not needed, and why this might be absent for other bacteria
* If this aligns with the tree (is this a deletion? An insertion? Do proteins with/without the helix have a common ancestor, is the protein even related anymore if the helix is missing? Should this be indicative of a different protein family entirely?)
* Other structural findings go here (eg evidence 2d section, papers backing up what features which separate the structures do)
* Cofactor evidence goes here

# Conclusion

Conclusion

* Some statement on the markers for differentiating the protein by function, and argue that the absence of the helix could indicate different protein function (even though the enzyme acts by the same activity and technically has the same function, the bacterial action of the protein is so different between species clusters that perhaps there should be a separation in the way the domain is classified based on absence/presence of that helix near the structural core?)

# References

# Appendices

## Appendix A: Longlist of candidate NAMLAA experimental structures identified from searches of the Protein Data Bank.

[to paste here]

## Appendix B: Chimera alignment of the full (un-cropped) 19 shortlisted structures

[to paste here]

## Appendix C: Grouped multiple structural alignments of experimental structures with PDBeFOLD

[to paste here, probably a table]

1. Uncropped 19 (PDBeFOLD)
2. Uncropped 19 (USAlign)
3. Gram-negative
4. Gram-positive

From Martin Steinegger talk 29/02/2024:

* Classifier trained on mills of protein sequences: random forest, used a large transformer to train on small number of data points (169)
* Combined with homology search to rank candidate proteins according to protein performance/function being similar to an input protein, even with low sequence identity
* (**ColabFold and ESM1b databases used, notebook available for homologous protein search – seekrank.steineggerlab.com) – pre print out soon 2024 ‘discovery of highly active…’)**
* Predicted structure databases released in 2022: Uniprot TrEMBL, ESMatlas-MGnify
* Isoform.io: Structure-guided isoform identification for the human transcriptome – predictable structures more likely to be real proteins than unpredictable structures (pLDDT: higher pLDDT = more likely the protein is real. Over 70 is pretty good.). Some isoforms have better structural prediction than MANE select, this website lists the 940 proteins this is the case for.
* Foldseek: 1D. Reduce 3D structure into 1D string. Each residue represented by a structural state letter. Describes tertiary interactions – the nearest 3D neighbour in space (simplified to the virtual centre of a residue, distance between these virtual centres measured), features extracted from the nearest neighbour (torsion angles and distances), discretise these different feature combinations by feeding into an encoder/decoder. Creates 20 different states. Called 3Di sequences.
  + Foldseek-TM: Improvement to Foldseek with higher precision/recall to TM-align?
  + ‘giant genes are rare but implicated in cell wall degradation by predatory bacteria’, west-roberts et al 2023, biorxiv pre-print (might be worth a read?)
* ‘Fast and accurate protein structure search with Foldseek’, van Kempen and Kim et al 2023 Nature biotech <https://www.nature.com/articles/s41587-023-01773-0>
* Clustering of AlphaFold TrEMBL with structures:
* A diagram of a protein universe using mms

  Description automatically generated
  + **Clustering predicted structures at the scale of the known protein universe, Barrio and Yeo et al** [**https://www.nature.com/articles/s41586-023-06510-w**](https://www.nature.com/articles/s41586-023-06510-w)
  + Evolutionary analysis, ‘dark clusters’ (no annotation in PDB, Pfam, TIGRfams) and new domain annotation from structural similarity!
    - **Lowest common ancestor analysis**: take tree of life and your cluster, and compute where in the tree these proteins in the cluster would meet. Quantifies how specific a cluster is (is it unique to one organism? One family? Etc.)
    - **Cluster.foldseek.com: Any uniport identifier, identify clusters (AFDB clusters)**
* Phylogeny: 3di can help with multiple sequence alignments! Pre-print here: (<https://www.biorxiv.org/content/10.1101/2023.12.12.571181v2.full.pdf>)
* Structure alignment: <https://github.com/steineggerlab/foldmason>