**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](#_Toc165099686)

[ACKNOWLEDGEMENTS 4](#_Toc165099687)

[ABBREVIATIONS 4](#_Toc165099688)

[INTRODUCTION 4](#_Toc165099689)

[MATERIALS AND METHODS 7](#_Toc165099690)

[Data Collection 7](#_Toc165099691)

[Identification of candidate AmiC structures 7](#_Toc165099692)

[Generating an AmiC sequence homologue list 8](#_Toc165099693)

[Structural Analysis 8](#_Toc165099694)

[Initial candidate structural alignment 8](#_Toc165099695)

[Sequence Analysis 9](#_Toc165099696)

[Sequence annotation and filtering 9](#_Toc165099697)

[Multiple Sequence Alignment, validation, and analysis 9](#_Toc165099698)

[Domain Fusion Analysis 10](#_Toc165099699)

[Co-factor Binding Analysis 10](#_Toc165099700)

[RESULTS 10](#_Toc165099701)

[EXPERIMENTAL STRUCTURAL ANALYSIS 10](#_Toc165099702)

[Candidate structural alignment (Experimental Data) 10](#_Toc165099703)

[SEQUENCE ANALYSIS 11](#_Toc165099704)

[Generating a candidate AmiC structure shortlist 11](#_Toc165099705)

[Generating an AmiC sequence homologue list 13](#_Toc165099706)

[Multiple Sequence Alignment and Validation 13](#_Toc165099707)

[Gram positive versus Gram negative 14](#_Toc165099708)

[Feature identification 14](#_Toc165099709)

[Sequence clustering 15](#_Toc165099710)

[SYNTHETIC STRUCTURE ANALYSIS 15](#_Toc165099711)

[DISCUSSION 17](#_Toc165099712)

[CONCLUSION 19](#_Toc165099713)

[REFERENCES 19](#_Toc165099714)

[APPENDICES 23](#_Toc165099715)

[Appendix A: Longlist of candidate NAMLAA experimental structures identified from searches of the Protein Data Bank. 23](#_Toc165099716)

[Appendix B: Chimera alignment of the full (un-cropped) 19 shortlisted structures 23](#_Toc165099717)

[Appendix C: Grouped multiple structural alignments of experimental structures with PDBeFOLD 23](#_Toc165099718)

# ACKNOWLEDGEMENTS

Molecular graphics and analyses performed with UCSF Chimera, developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco, with support from NIH P41-GM103311.

On a personal note, I would also like to thank and acknowledge my supervisor Dr Mark Williams for all the ongoing assistance and hugely helpful 1:1 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
  + [**https://link.springer.com/article/10.1007/BF00272354**](https://link.springer.com/article/10.1007/BF00272354) **- action of the N-acetylmuramoyl-L-alanine amidase as cell wall lytic enzyme?**
  + 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/AmiB amidases, shown experimentally to have the mobile helix region.
  + From 3NE8 paper: 
  + 3NE8 paper posed 3 functions for LytC-type amidases (NAMLAA): lytic from bacteriophages, lytic from sporulating bacteria, and cell separation amidases. And that those with cell separation function have the self-regulating helix.
  + 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.
* 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 (reasons why functional difference in AmiC might not separate the different functions?)
  + 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 (InterPro contains CATH-Gene3D (CATH for complete genomes), Pfam, CDD (multiple alignment databases), PROSITE, SMART, and others.)
    - 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/SPARCLE1: 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/Evolutionary-divergence-and-functional-investigation-of-NAMLAA-amidases>.

## Data Collection

### Identification of candidate AmiC structures

The FASTA sequence for AmiC protein structure 4BIN was downloaded from the Protein Data Bank (PDB)2–4. This sequence was used in three separate structure searches on 17/05/2023: a BLASTp search restricted to the PDB, a PDB search using the search term ‘amidase\_3’, and a SCOP search using identifiers 4001130 (N-acetylmuramoyl-L-alanine amidase-like family) and 8092183 (AmiC domain)5–7. 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 structures8,9. 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’ in May 202310,11. 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 the Protein structure comparison service PDBeFold at European Bioinformatics Institute (<http://www.ebi.ac.uk/msd-srv/ssm>) 12,13 (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). 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. All shortlisted protein structures were downloaded from the PDB on 19/05/2023.

### Generating an AmiC sequence homologue list

BLASTp using FASTA protein sequences for each candidate structure was performed through the BLAST web browser on May 2023, and repeated in April 2024. Sequences were cropped to the Pfam annotation for the amidase\_3 domain as listed on the PDB record for each structure. The blastp algorithm was run using the default parameters with the following specifications: Database restricted to 59,610,710 sequences in RefSeq Select14, no organism or model exclusion, 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. The top 5000 aligned sequence results were downloaded for each search. A Python script was used to combine the sequences from BLASTp into a single list and remove duplicate entries. Notably, input sequences associated with the 1XOV, 3CZX, 3QAY, and 7TJ4 were not returned in their respective BLASTp searches as the sequences were not present within RefSeq Select, and were therefore manually added to the final BLASTp single list.

The HMMER 3 phmmer (REST) web service Python-based client 15 was used to search for further sequences using the below command, although outputs were not deemed informative for this purpose as only the top 50 alignments were provided without sequence ID/UniProt accession or the FASTA sequence itself.

python hmmer3\_phmmer.py --email 'sallen10@student.bbk.ac.uk' --database 'uniprotrefprot'

When searching InterPro for other relevant accessions (using the search term ‘N-acetylmuramoyl-L-alanine amidase’), several protein families of relevance were identified: 1 superfamily: [IPR036505](https://www.ebi.ac.uk/interpro/entry/InterPro/IPR036505/) (PGRP domain superfamily), 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), [IPR017293](https://www.ebi.ac.uk/interpro/entry/InterPro/IPR017293/) (SH3-domain containing protein), [IPR049745](https://www.ebi.ac.uk/interpro/entry/InterPro/IPR049745/) (AmiC NAMLAA), [IPR050695](https://www.ebi.ac.uk/interpro/entry/InterPro/IPR050695/) (amidase\_3), and [IPR051206](https://www.ebi.ac.uk/interpro/entry/InterPro/IPR051206/) (amidase\_2). These families contained other domains which were not relevant for this stage of analysis, and were recorded for later domain-fusion analysis. 4 additional domains were also identified: [IPR002502](https://www.ebi.ac.uk/interpro/entry/InterPro/IPR002502/) (amidase\_2 domain), [IPR002508](https://www.ebi.ac.uk/interpro/entry/InterPro/IPR002508/) (amidase\_3 domain), [IPR021976](https://www.ebi.ac.uk/interpro/entry/InterPro/IPR021976/) (CBD\_PlyG cell wall binding domain), and [IPR048586](https://www.ebi.ac.uk/interpro/entry/InterPro/IPR048586/) (Cwp6 domain). Of these, only IPR002508 was used as a search term to download homologous sequences, since this was the only catalytic domain not associated with other protein families like amidase\_2. A Python script adapted from the InterPro website (<https://www.ebi.ac.uk/interpro/result/download/#/protein/UniProt/entry/InterPro/IPR002508/|fasta>) was used to download all IPR002508 sequences, UniProt accession IDs, taxonomic IDs and species names under the amidase\_3 entry IPR002508 and cropped the associated FASTA sequences to contain only the amidase\_3 domain as annotated by InterPro.

## Structural Analysis

### Initial candidate structural alignment

To visualise the alignment of the shortlisted structures and produce an all-vs-all pairwise structural alignment for later validation of the multiple sequence alignment, the MatchMaker tool within UCSF Chimera was run with the following parameters: secondary structure predictions not re-calculated, pairwise alignment, reference structure 4BIN in *E. coli*, Needleman-Waltsch alignment16,17. Multiple structure alignment was carried out using PDBeFOLD. Structure/sequence cropping was carried out in UCSF Chimera using a command script based on the Pfam boundaries of the ‘amidase 3’ domain as listed in the PDB. Confirmatory alignment of all structures to all other structures in the shortlist was carried out with a local download of US-align18 using the below command:

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

## Sequence Analysis

### Sequence annotation and filtering

Sequences from InterPro and BLASTp were combined using the BioPython19 module as part of a Python script, and duplicate sequences removed using a separate Python script. This script used taxonomic ID to identify unique sequences from the same species, and preferentially kept the RefSeq reference sequence from BLASTp over the InterPro sequence where applicable. This was to apply a level of caution to the InterPro sequences, which were mostly sourced from WGS preliminary or unreviewed sequences stored within UniProt, and therefore where a protein sequence has already been recorded for a species within RefSeq, the RefSeq sequence was used preferentially within the final sequence list as these have been annotated and curated by NCBI staff, and are non-redundant. Taxnomic ID was extracted within the InterPro API for InterPro sequences, and a separate Python script annotated BLASTp sequences with the taxonomic ID using a call to the NCBI REST API v220,21. If this API call failed, sequences were manually searched, as for some species (for example, species in the *Halomonas* genus) the species name did not match the name stored within the NCBI Taxonomy database22, but could be identified through manual searching of this database. Of note, some sequences from RefSeq had the same TaxID as other sequences from RefSeq; this is because the species name was used to call the REST API, and the species name was annotated at a higher order (for example, *Paenibacillus* genus instead of the species). All sequences from RefSeq were retained to avoid excluding novel species that were annotated at genus or family level.

The final list of sequences was annotated further with higher-order classification information for each species. This was done using a Python script to make an API call to Taxallnomy23 to annotate the family name for each species. A second Python script then identified and removed sequences which were annotated with non-specific family names (such as ‘Fam\_of\_environmental samples’ or ‘uncultured’), sequences which failed the API Taxallnomy call, non-bacterial *Thermococcaceae* and *Methanosarcina*, and any sequence which was partial (defined as a sequence length < 131, the threshold value determined as being more than 3 standard deviations (18.84) from the sample mean (187.63) of the amidase\_3 domain length in the candidate sequences).

### Multiple Sequence Alignment, validation, and analysis

The initial multiple sequence alignment was performed using a local download of kalign-324, chosen because kalign-3 was the most feasible tool to compute the alignment within reasonable computational time. The below command produced the initial alignment:

D:\msys2\usr\bin\kalign.exe -i keep\_seq\_taxid.txt -o D:\Documents\MSc\_Project\MSc\_Project\01\_DATA\Amidase\_3\04\_Multiple\_Alignment\alignment\_1.fa

The output from this was then filtered using OD-seq25, which is designed to detect outliers in large multiple sequence alignments by calculating an average distance matrix and predicting outliers using bootstrapping (1000 pseudo-replicates) to compute the mean and standard deviation of distance measures between an inspected sequence and the overall alignment. The following command was used to run a local instalment of OD-seq:

OD-seq -i alignment\_1.fa -r alignment\_1\_odseq.txt -o alignment\_1\_outliers.txt -c alignment\_1\_nonoutliers.txt -m affine -s 1

For OD-seq, ‘-r’ is the file name for all results, ‘-o’ is the file name for outlier sequences, ‘-c’ is the file name for non-outlier sequences, ‘-m’ is the gap penalty assignment method, and ‘-s’ is the number of standard deviations from the average distance permitted for sequence retainment. Affine was chosen as this gives a higher penalty (3 points) for opening a gap, and 1 point for each extension of an existing gap, meaning short gaps would be penalised in otherwise conserved regions without strongly penalising the long insertion region, as would be done for the Cumulative method. Alignment with kalign-3 was repeated as above after filtering. An R script was used to further filter out sequences which introduced a gap that contained information from <1% of all sequences in the alignment. Filtering was initially performed using 0.1% of sequences, however the alignment length remained around double the average amidase\_3 domain length from the candidate structures, and as amidase\_3 is a highly conserved domain, a shorter MSA obtained with a stricter threshold for information content was preferred to better ensure the sequence list contained ‘true’ amidase\_3 sequences. The MSA was visualised within Jalview26.

Validation was carried out according to the methods described by Modi and Dunbrack Jr.27 by comparing aligned residues in the structural alignment with the sequence alignment. Using the US-align all-vs-all pairwise alignment output, a Python script compared aligned residues in the structural alignment of each protein pair to the sequence alignment of the same proteins, and calculated the Positive Predictive Value (PPV): the ratio of sequence pairs aligned in both the sequences and structures to the total number of residues aligned in the sequences, and True Positive Rate (TPR): the ratio of sequence pairs aligned in both the sequences and structures to the total number of residues aligned in the structures.

### Conservation of regions in the final MSA

A conserved region was defined a continuous section of at least 3 columns where average occupancy of the columns was >90%, there was not >1 continuous column of <90% occupancy (gap-rich columns), and the region was not composed of >50% gap-rich columns. Regions were identified based on the JalView conservation quality scores, and boundaries determined using the above rules by viewing the alignment within JalView26,28,29. Logos were generated for conserved regions using WebLogo30, coloured by charge (positive=green, negative=red) and polarity (polar=blue, non-polar=black). Conservation scores were generated for each column of the MSA using Jalview web services28,29, and visualised by being mapped onto the residues for 4BIN and 3NE8 using custom UCSF Chimera attribute assignment files and rendering by attribute.

Sequences were annotated as being gram-positive or gram-negative based on the species phyla using the R package AMR31. Species not recognised by AMR were manually searched in the NCBI Taxonomy Browser22 to identify their phyla and annotate them as gram-positive or gram-negative using the same rulings as AMR. To quantify the level of conservation of each species, species were annotated using a Python script depending on % of occupied columns (i.e. non-gap positions) across I-6, and species gram-status and phylum reviewed for patterns.

### Feature identification and Sequence Clustering

For each insertion region identified, a Python script assigned each sequence a value of “0” if the sequence occupied <50% of the region with information, and a value of “1” if the sequence occupied ≥50% of the region, giving each sequence a binary fingerprint of 8 digits long based on which insertion regions the sequence appeared to have. These fingerprints were then used to cluster sequences based on which insertion regions they contained – this was done using an R script to perform PCA to cluster sequences.

## Domain Fusion Analysis

In earlier Interpro searching, several accessions of relevance were identified.

## Co-factor Binding Analysis

# RESULTS

## EXPERIMENTAL STRUCTURAL ANALYSIS

### Candidate structural alignment (Experimental Data)

[talk about the 4 AmiA structures here and how they didn’t align very well with the other 19 structures!]

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 domain32,33. 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. As the visualisation 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 in this alignment.

**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 US-align 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).

|  |  |  |  |
| --- | --- | --- | --- |
| **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 |

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

## SEQUENCE ANALYSIS

### Generating a candidate AmiC structure shortlist

From an initial structural longlist constructed of possible NAMLAA structures across species, a final shortlist of 19 candidate AmiC structures was curated, which matched on Pfam annotation, GO annotation, and an Interpro domain annotation of amidase\_3 (Table 1).

The lengths of the sequences for these structures varied widely in length from 175-638, however length of the annotated amidase\_3 domain varied less significantly, with a range from 159-215. 5J72 and 4RN7 are both from *Clostridium difficile*, however were deposited by different research groups and present different structures and sequences, where 5J72 is much longer. It is possible 4RN7 is a ‘cropped’ structure representing just the catalytic domain of this protein. 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. The species in this shortlist vary 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.

### Generating an AmiC sequence homologue list

A BLASTp search for the top 5000 significantly aligned RefSeq sequences was performed for each of the candidate AmiC amidase\_3 domains, which identified in total 61,489 sequences. The number of identified sequences is given for each protein in Table 1. After removing duplicate sequences, BLASTp identified 39,303 candidate AmiC homologous sequences.

Interpro was then searched to identify further sequences. Of the search terms used to identify sequences potentially homologous to the AmiC amidase domain, the only domain that was not associated with amidase\_2 and was a catalytic domain instead of a binding domain was the ‘amidase\_3’ domain (IPR002508). Download of sequences annotated as IPR002508 identified 70,617 non-unique sequences, and after combining these with the sequences identified from BLASTp and removing duplicates, the final unique number of sequences was 57,277.

**Table 1: Shortlist of 19 candidate experimental structures of NAMLAA AmiC across different bacterial species.** Ligand, organism and sequence information taken from the PDB entry for each structure. Chain listed is the best-matched chain identified by the respective identification method for this shortlist. PDB entry reference is listed for each structure, as well as the publication where available.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **PDB ID** | **Year in PDB** | **PDB**  **chain** | **Organism** | **Ligand** | **Full Sequence Length** | **Amidase domain length** | **Identification Method** | **Sequences identified from BLASTp** | **PDB Reference** | **Publication** |
| [4BIN](https://www.rcsb.org/structure/4BIN) | 2013 | A | *Escherichia coli* | Zn2+ ,Na+ | 403 | 215 | PDB search ‘AmiC’ | 5000 | 2 | 3 |
| [7B3N](https://www.rcsb.org/structure/7B3N) | 2020 | A | *Thermus parvatiensis* | Zn2+, Na+, Cl-, Glycerol, SO42-, Ethanesulfonic Acid | 177 | 159 | BLASTp search, 4BIN reference sequence | 5000 | 34 | 35 |
| [3CZX](https://www.rcsb.org/structure/3CZX) | 2008 | A | *Neisseria meningitidis* | Zn2+ | 182 | 169 | PDB search ‘Amidase\_3’ | 185 | 36 | *Not applicable* |
| [5J72](https://www.rcsb.org/structure/5J72) | 2016 | A | *Clostridium difficile* | Citric Acid, Zn2+, Ca2+, Cl-, Na+ | 638 | 181 | BLASTp search, 4BIN reference sequence | 5000 | 33 | 32 |
| [7TJ4](https://www.rcsb.org/structure/7TJ4) | 2022 | B | *Staphylococcus aureus* | Zn2+ | 176 | 165 | BLASTp search, 4BIN reference sequence | 5000 | 37 | 38 |
| [4RN7](https://www.rcsb.org/structure/4RN7) | 2014 | A | *Clostridium difficile* | Zn2+, Formic Acid, Glycerol, Ethanesulfonic Acid | 188 | 174 | BLASTp search, 4BIN reference sequence | 5000 | 39 | *Not applicable* |
| [5EMI](https://www.rcsb.org/structure/5EMI) | 2015/2016 | A | *Nostoc punctiforme* | MES, MRD, Zn2+ | 180 | 169 | BLASTp search, 4BIN reference sequence | 5000 | 40 | 41 |
| [3NE8](https://www.rcsb.org/structure/3NE8) | 2010 | A | *Bartonella henselae* | ACETATE ION, FORMIC ACID, GLYCEROL, Zn2+ | 234 | 215 | BLASTp search, 4BIN reference sequence | 5000 | 42 | 43 |
| [4M6G](https://www.rcsb.org/structure/4M6G) | 2013 | A | *Mycobacterium tuberculosis* | Zn2+ | 225 | 203 | PDB search ‘Amidase\_3’ | 2114 | 44 | 45 |
| [1JWQ](https://www.rcsb.org/structure/1JWQ#entity-1) | 2001 | A | *Paenibacillus polymyxa* | Zn2+ | 179 | 170 | BLASTp search, 4BIN reference sequence | 5000 | 46 | *Not applicable* |
| [4LQ6](https://www.rcsb.org/structure/4LQ6) | 2013 | A | *Mycobacterium tuberculosis* | SO4-, Zn2+, Cl-, Pt2+ | 218 | 203 | PDB search ‘Amidase\_3’ | 2114 | 47 | 48 |
| [4M6H](https://www.rcsb.org/structure/4M6H) | 2013 | A | *Mycobacterium tuberculosis* | None reported | 225 | 203 | PDB search ‘N-acetylmuramoyl-L-alanine activity’ | 2114 | 49 | 45 |
| [4M6I](https://www.rcsb.org/structure/4M6I) | 2013 | A | *Mycobacterium tuberculosis* | Zn2+ | 225 | 203 | PDB search ‘Amidase\_3’ | 2114 | 50 | 45 |
| [7AGO](https://www.rcsb.org/structure/7AGO) | 2020 | A | *Mycobacterium* | Zn2+, D-alpha-glutamine, Alanine | 233 | 203 | PDB search ‘Amidase\_3’ | 1942 | 51 | 52 |
| [7AGM](https://www.rcsb.org/structure/7AGM) | 2020 | A | *Mycolicibacterium smegmatis* | Zn2+ | 230 | 203 | PDB search ‘Amidase\_3’ | 2435 | 53 | 52 |
| [7AGL](https://www.rcsb.org/structure/7AGL) | 2020 | A | *Mycobacterium* | Zn2+ | 233 | 203 | PDB search ‘Amidase\_3’ | 1942 | 54 | 52 |
| [7RAG](https://www.rcsb.org/structure/7RAG) | 2021 | B | *Clostridium difficile* | Zn2+, 1,2-Ethanediol | 210 | 188 | PDB search ‘Amidase\_3’ | 5000 | 55 | 56 |
| [1XOV](https://www.rcsb.org/structure/1XOV) | 2004 | A | *Listeria monocytogenes* | Cl- ,GLUTAMIC ACID, LYSINE, 2-AMINO-2-HYDROXYMETHYL-PROPANE-1,3-DIOL, SO4-, Zn2+ | 326 | 167 | PDB search ‘Amidase\_3’ | 1066 | 57 | 58 |
| [3QAY](https://www.rcsb.org/structure/3QAY) | 2011 | A | *Colneyvirus* | SO4-, Zn2+ | 180 | 172 | PDB search ‘Amidase\_3’ | 463 | 59 | 60 |

### Multiple Sequence Alignment and Validation

To identify conserved and insertion regions across species and cluster these accordingly, a multiple sequence alignment (MSA) for the sequence list was generated. To ensure a broad coverage of all possible AmiC homologues, sequence identification may have included non-AmiC proteins with a similar catalytic domain. To adjust for these less similar protein sequences pre-alignment, filtering based on species was applied to restrict sequences to bacterial sequences, with the exception of viral sequences, as these may originate from bacteriophages which are highly relevant to this analysis. Partial sequences (4,233) and sequences with a missing or unclassified taxonomic lineage (3,441) were also removed before alignment. After filtering, 49,554 sequences remained for the initial MSA.

To remove sequences which are unlikely to be AmiC homologues due to having an unusually divergent sequence from other homologues while still retaining sequences with insertions that are likely to be valuable in clustering of features, the initial multiple sequence alignment performed with k-align was further filtered in two stages:

* Outlier detection program OD-seq25 removed distant sequences (8,596), which removed 5 of the candidate structure sequences from the sequence list: 3QAY (a bacteriophage virus), 4LQ6, 7AGL, 7AGM (the three candidate sequences from the family *Mycobacteriaceae*), and 7B3N (from *Thermaceae*, which is known to be distinct from other bacterial families through conserved signature insertions/deletions61).
* A separate R script removed sequences that introduced gaps which represented information from <1% of the data (see Methods). This further removed 5 candidate sequences: 1JWQ, 1XOV, 3CZX, 5EMI, and 5J72.

The final post-filtered alignment contained 26,661 sequences and was 351 residues long including gaps.

As described by Modi and Dunbrack Jr in 201927, the MSA was then validated by using a structural alignment of the candidate structures retained in the final alignment by examining the conserved (aligned) residues in the structural alignment versus the sequence alignment, and calculating the True Positive Rate (TPR) and Positive Predictive Value (PPV) for each. On average, TPR was 0.946 and PPV was 0.976 (TPR and PPV for each of the 5 structures are in Appendix D), although gaps were still present in the final alignment as anticipated, which might be indicative of evolutionary changes for each species (e.g. insertions for certain species as the bacteria evolve). There were 8 total potential insertions which are described in Table 2, the shortest being I-5 (1 residue) and the longest being I-6 (72 residues). To explore which regions these low-occupancy insertions refer to, conservation was mapped onto the structures of the 4BIN and 3NE8 structures (Figure 3), which showed high conservation in the core of the amidase\_3 structure and poor conservation of the outer helical structures. Region I-6, the longest insertion, maps onto both structures as being the proposed mobile helix.

**Table 2: Descriptive summary of final multiple sequence alignment of 26,661 sequences, length 351 residues.** Conserved region IDs = ‘C-X’, Insertion region IDs = ‘I-X’. Logos for conserved regions created using WebLogo30. Full web logo for all regions can be found in Appendix E.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Region** | **Region Columns** | **Region Length** | **Average Occupancy (%)** | **Consensus sequence** | **Logo** |
| C-1 | 1-9 | 9 | 98.68 | IVIDPGHGG | A black and red text with green and red letters  Description automatically generated with medium confidence |
| I-1 | 10-45 | 36 | 31.14 | -- | -- |
| C-2 | 46-67 | 22 | 94.02 | TRNEKDVTLAIARKLRALLEAE |  |
| I-2 | 68-75 | 8 | 23.51 | -- | -- |
| C-3 | 76-89 | 14 | 93.21 | ARVVLTRDGDRYFV |  |
| I-3 | 90-106 | 17 | 10.05 | -- | -- |
| C-4 | 107-129 | 23 | 99.83 | LRERVEIARKAGADLFVSIHA  DA |  |
| I-4 | 130-140 | 11 | 37.01 | -- | -- |
| C-5 | 141-144 | 4 | 93.81 | SARG |  |
| I-5 | 145 | 1 | 5.66 | -- | -- |
| C-6 | 146-148 | 3 | 96.27 | ASV |  |
| I-6 | 149-220 | 72 | 46.89 | -- | -- |
| C-7 | 221-233 | 13 | 90.92 | SGLKLAKAVLKEL |  |
| I-7 | 234-267 | 34 | 43.76 | -- | -- |
| C-8 | 268-313 | 46 | 96.16 | FAVLKAAPDIPSVLVETGFISNPEE  EALLKSAAYQQKLAEAIADGI |  |
| I-8 | 314-351 | 38 | 12.71 | -- | -- |

A structure of a cell

Description automatically generatedA structure of a cell

Description automatically generated

**Figure 3: BLOSUM62 conservation scores mapped onto two candidate AmiC structures.** Visualisation of structures was carried out using UCSF Chimera.Left: 4BIN. Right: 3NE8. Pink colouration = highly conserved residue, blue colouration = poorly conserved residue. The proposed mobile helix has been highlighted in each structure.

### Conservation of I-6 across species

Having defined the boundaries for conserved and less-conserved regions, and specifically the boundaries of the proposed mobile helix in region I-6, each species in the alignment was then examined to explore any species clustering patterns based on relative conservation of I-6. Based on the hypothesis that the mobile helix is linked to differing function in gram-negative (cell wall division) and gram-positive (endolytic behaviour) bacteria, species were also annotated with their predicted gram staining status (Figure 4).

A graph with numbers and a number

Description automatically generated with medium confidence**Figure 4: Raw count of per-species occupancy of the I-6 region, annotated by gram staining status.**

Looking at this conservation, it is evident that there is an association between gram-staining status and occupancy in this insertion region, where gram-negative bacteria are more likely to have a considerable (>40-residue insertion) insertion spanning this region, while gram-positive bacteria and bacteriophages tend to lack this insertion (insertion of <14 amino acids). Few bacteria occupy an insertion between these two distinct groups, showing a clear binary difference between ‘large insertion’ and ‘low/no insertion’.

Some gram-negative bacteria appear to lack this considerable insertion, aligning better with gram-positive bacteria with regard to this region. On further investigation, these bacteria share a common phylum (*Cyanobacteriota*). A considerable number of gram-positive bacteria occupied 37 columns, marking them as a unique group separate from gram-negative and even other gram-positive bacteria. With the exception of 3 species, these belonged to the *Streptomycetaceae* family.

### Feature identification

* 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
    - Table 2 for regions
    - Other features: Secondary structures?
  + Either: A binary fingerprint for each sequence eg species 1 = [0,0,1,1,1,1,1,0,1,1,0] where 0 = missing region and 1 = has region which can then be clustered
    - Would need to define when a 0 becomes a 1 = 75% occupancy threshold?
      * Region-specific: Can define per the I-6 region? Would need to look into
    - Simpler, might not need PCA
    - Primary candidate I think.
  + Or: A raw count model eg species 1 = [30, 5, 10, 15, 40, 32, 1, 1] which can then be clustered
    - Could use PCA on this?
    - Can include features like number of helices, number of beta sheets
    - Physiochemistry
      * No. each residue type? (neg, pos, polar?)
      * A binary Y/N to having polar at pos. X if pos. X is relevant for enzymatic function?
    - Annotate sequences with secondary structure data?
      * SPOT-1D-LM is back up again, is this a viable option for annotation?
* 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?)
  + Simple phylogenetic tree using one representative sequence from each family? Do myself?
  + <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?
  + Phylogeny: Does the phylogeny match what I expect (eg a divergence shown where the helix is inserted/deleted depending on what was needed for function)? Is there a split between has helix vs does not have helix?

### Sequence clustering

* 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

## SYNTHETIC STRUCTURE ANALYSIS

(May-June)

* 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: 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

JULY(if time): Evidence 4: Cofactor binding

* Need to do

# DISCUSSION

Discussion

* Candidate experimental structural alignment
  + High TM scores and average Q-scores indicate conservation for these structures – closely related
  + Including AmiA structure (x5) reduced average Q-score and TM-score significantly(need to put in Appendix C), therefore some confirmatory proof that these structures were all AmiC (as per 4BIN structure).
    - 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?
  + However, 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?
    - 3NE8 and 4BIN: 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).
    - Clostrium difficle is sporulating, therefore the amidases here which had poor alignment structurally were probably in this functional class alone in the dataset (explains poorer TM scores for 7RAG, 5J72 when not cropped) (appendix C)
    - 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?
* Sequence analysis
  + Removal of structural candidate sequences at later stage filtering (which isn’t due to quality of the sequence, and more to do with that these sequences did not fit with the majority of other identified sequences well enough).
    - None of the candidate sequences with an associated structure were removed prior to multiple sequence alignment – so the sequences weren’t removed due to poor quality or looking incomplete, therefore sequences removed because they were too distant from the others in the set.
    - Thermus family being separate anyway
    - Colnyvirus being fairly distinct too
    - Mycobacteria family being removed with OD-seq (although Mycolicibacterium being retained). Is there a difference functionally? Do some mycobacteria have amidase\_3 function and others do not? Cell wall structure being different to gram negative despite cell wall presence, etc etc.
    - The 5 sequences being removed with the R script – why? Is there a pattern?
      * Why was one clostriudium difficle remove and one was not? Is one AmiC (the retained one) and one not AmiC (the 5J72 one, which is supposedly the exact same species and enzyme as the retained one, however it had a very long N-terminus. Not AmiC after all?)
        + Notably: of the 5 removed at this stage, 4 are retained by adjusting the threshold from 1% information to 0.1% information – the only sequence which is still too distinct to retain is 5J72.
  + Conservation from the alignment mapped onto structures: high conservation in the enzymatic core and poor conservation on the outside of this core – as expected, commonly seen for enzyme families. Specifically links in as the proposed mobile helix is also poorly conserved despite having a potential function, so likely that these enzymes have similar function (since both structures where available and sequences are highly conserved having similar sequence properties and overlapping structures in space), although not all have a helix which is present in the enzymatic core – therefore same enzyme, different function?
  + Look at the conserved region logos – any conserved residues in particular? Same amino acids, same polarity/charge? What does this mean for the enzymatic site function as these conserved regions map well onto the enzymatic core? Can we say again that despite differing species and differing gram status, the enzymatic function is retained across species, and the difference is in how the enzyme is used/activated (possibly by way of the mobile helix)?
  + Limitation: As sequence identity is so low for these proteins, possible that sequence included in the multiple sequence alignment are not true AmiC homologues, which may affect the final phylogeny if we assume all sequences are AmiC homologues (and possibly conflate non-helix containing AmiC sequence/structures with other proteins entirely, meaning the set of AmiC gram positives might not actually be AmiC – weakning the strength of the conclusion that AmiC with helix are different to AmiC without helix, might be that the only AmiC in there are ones with a helix). As not really much separation in the function of amidase\_3 domain in current domain annotations, this is to be expected, have tried to adjust for this known-unknown with quality filtering steps and validated with a structural alignment of candidate AmiC structures – this alignment also showed that there are large structural differences between known AmiC and AmiA proteins, and therefore sequences introducing large gaps in the alignment may not be true AmiC homologues, however certainty of this is not complete and could only be resolved by elucidating more structures of AmiC proteins (that have been confirmed to be AmiC experimentally through their biochemical role in the bacterial cell) to confirm there are AmiC with and without this helix that match with their functional role.
* Gram positive/negative splitting
  + General limitation: Acknowledge that this is a non-biological term, not very specific and boundaries can be blurred (see: Mycobacterium tuberculosis), less of a definitive term and more of a general term to relate the sequence/structure differences to a biological function. Also acknowledge that annotation of species as gram negative/positive is based on the classification of each species, and there is a chance that some of the enzymes have been mis-classified as the wrong species, especially for WGS shotgun sequences from InterPro. Therefore, have tried to define those ‘gram negative’ with a helix as ‘cell wall separation’ function, and those ‘gram positive’ as alternative sporulating or lytic functions, and state that this classification is a general term to group the families with/without the helix. Have also tried to apply quality filter to the InterPro sequences, and preferentially includes sequences from species that had been curated for the RefSeq database.
  + Gram negative having the I-6 insertion, gram positive tend not to. Supports the hypothesis that this insertion is relevant for a function in gram-negative that is not used in gram-positive bacteria – cell-wall separation would require such an insertion if the function of this insertion is to ‘switch off’ the enzymatic activity, whereas gram positive bacteria – which do not have a cell wall – would require the enzyme for a different reason that may not require such ‘switching off’ – for example, it would not be useful for an enzyme used in endolytic bacteria designed to kill neighbouring or parent bacterial cells to have an off switch, since uncontrolled amidase activity would result in cellular breakdown, which would be the goal of such endolytic activity. I-6 as a marker of function?
  + However the separation is not perfect – there are gram negative bacteria lacking the insertion. These share *Cyanobacteriota* as their phylum. Why? Are all Cyanobacteriota lacking the insertion or just some? Do these bacteria have alternative activity? Do the bacteria in fact have multiple copies of the AmiC NAMLAA, where one copy is for cell wall activity and one is for endolytic activity? Do these bacteria have paralogs for multiple purposes?
  + Some gram positive have 37 occupied columns and appear unique. These are almost exclusively *Streptomycetaceae*. Why? What are these? Why do they have a considerable insertion region when the other gram positives don’t seem to?
    - May not be particularly significant after all; 441/1299 species with 37 occupied columns are different Streptomyces strains that could not be filtered earlier (as their records did not have a species name, just a genus and an alphanumeric ‘species’ which appears to be a strain ID).
* 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

1. Marchler-Bauer, A., Derbyshire, M.K., Gonzales, N.R., Lu, S., Chitsaz, F., Geer, L.Y., Geer, R.C., He, J., Gwadz, M., Hurwitz, D.I., et al. (2015). CDD: NCBI’s conserved domain database. Nucleic Acids Res. *43*, D222-226. 10.1093/nar/gku1221.

2. Kerff, F., Rocaboy, M., Herman, R., Sauvage, E., and Charlier, P. (2013). Crystal structure of the E. coli N-acetylmuramoyl-L-alanine amidase AmiC. https://doi.org/10.2210/pdb4bin/pdb.

3. Rocaboy, M., Herman, R., Sauvage, E., Remaut, H., Moonens, K., Terrak, M., Charlier, P., and Kerff, F. (2013). The crystal structure of the cell division amidase AmiC reveals the fold of the AMIN domain, a new peptidoglycan binding domain. Mol. Microbiol. *90*, 267–277. 10.1111/mmi.12361.

4. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res. *28*, 235–242. 10.1093/nar/28.1.235.

5. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410. 10.1016/S0022-2836(05)80360-2.

6. Andreeva, A., Howorth, D., Chothia, C., Kulesha, E., and Murzin, A.G. (2014). SCOP2 prototype: a new approach to protein structure mining. Nucleic Acids Res. *42*, D310–D314. 10.1093/nar/gkt1242.

7. Andreeva, A., Kulesha, E., Gough, J., and Murzin, A.G. (2020). The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. Nucleic Acids Res. *48*, D376–D382. 10.1093/nar/gkz1064.

8. Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., Sonnhammer, E.L.L., Tosatto, S.C.E., Paladin, L., Raj, S., Richardson, L.J., et al. (2021). Pfam: The protein families database in 2021. Nucleic Acids Res. *49*, D412–D419. 10.1093/nar/gkaa913.

9. Paysan-Lafosse, T., Blum, M., Chuguransky, S., Grego, T., Pinto, B.L., Salazar, G.A., Bileschi, M.L., Bork, P., Bridge, A., Colwell, L., et al. (2023). InterPro in 2022. Nucleic Acids Res. *51*, D418–D427. 10.1093/nar/gkac993.

10. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene Ontology: tool for the unification of biology. Nat. Genet. *25*, 25–29. 10.1038/75556.

11. The Gene Ontology Consortium, Aleksander, S.A., Balhoff, J., Carbon, S., Cherry, J.M., Drabkin, H.J., Ebert, D., Feuermann, M., Gaudet, P., Harris, N.L., et al. (2023). The Gene Ontology knowledgebase in 2023. Genetics *224*, iyad031. 10.1093/genetics/iyad031.

12. Krissinel, E., and Henrick, K. (2004). Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Crystallogr. D Biol. Crystallogr. *60*, 2256–2268. 10.1107/S0907444904026460.

13. Krissinel, E., and Henrick, K. (2005). Multiple Alignment of Protein Structures in Three Dimensions. In Computational Life Sciences, M. R. Berthold, R. C. Glen, K. Diederichs, O. Kohlbacher, and I. Fischer, eds. (Springer), pp. 67–78. 10.1007/11560500\_7.

14. O’Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. *44*, D733-745. 10.1093/nar/gkv1189.

15. Finn, R.D., Clements, J., and Eddy, S.R. (2011). HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. *39*, W29–W37. 10.1093/nar/gkr367.

16. Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., and Ferrin, T.E. (2004). UCSF Chimera--a visualization system for exploratory research and analysis. J. Comput. Chem. *25*, 1605–1612. 10.1002/jcc.20084.

17. Meng, E.C., Pettersen, E.F., Couch, G.S., Huang, C.C., and Ferrin, T.E. (2006). Tools for integrated sequence-structure analysis with UCSF Chimera. BMC Bioinformatics *7*, 339. 10.1186/1471-2105-7-339.

18. Zhang, C., Shine, M., Pyle, A.M., and Zhang, Y. (2022). US-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes. Nat. Methods *19*, 1109–1115. 10.1038/s41592-022-01585-1.

19. Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., et al. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics *25*, 1422–1423. 10.1093/bioinformatics/btp163.

20. Sayers, E.W., Bolton, E.E., Brister, J.R., Canese, K., Chan, J., Comeau, D.C., Connor, R., Funk, K., Kelly, C., Kim, S., et al. (2022). Database resources of the national center for biotechnology information. Nucleic Acids Res. *50*, D20–D26. 10.1093/nar/gkab1112.

21. Federhen, S. (2012). The NCBI Taxonomy database. Nucleic Acids Res. *40*, D136-143. 10.1093/nar/gkr1178.

22. Schoch, C.L., Ciufo, S., Domrachev, M., Hotton, C.L., Kannan, S., Khovanskaya, R., Leipe, D., Mcveigh, R., O’Neill, K., Robbertse, B., et al. (2020). NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database J. Biol. Databases Curation *2020*, baaa062. 10.1093/database/baaa062.

23. Sakamoto, T., and Ortega, J.M. (2021). Taxallnomy: an extension of NCBI Taxonomy that produces a hierarchically complete taxonomic tree. BMC Bioinformatics *22*, 388. 10.1186/s12859-021-04304-3.

24. Lassmann, T. (2020). Kalign 3: multiple sequence alignment of large datasets. Bioinformatics *36*, 1928–1929. 10.1093/bioinformatics/btz795.

25. Jehl, P., Sievers, F., and Higgins, D.G. (2015). OD-seq: outlier detection in multiple sequence alignments. BMC Bioinformatics *16*, 269. 10.1186/s12859-015-0702-1.

26. Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M., and Barton, G.J. (2009). Jalview Version 2—a multiple sequence alignment editor and analysis workbench. Bioinformatics *25*, 1189–1191. 10.1093/bioinformatics/btp033.

27. Modi, V., and Dunbrack, R.L. (2019). A Structurally-Validated Multiple Sequence Alignment of 497 Human Protein Kinase Domains. Sci. Rep. *9*, 19790. 10.1038/s41598-019-56499-4.

28. Troshin, P.V., Procter, J.B., Sherstnev, A., Barton, D.L., Madeira, F., and Barton, G.J. (2018). JABAWS 2.2 distributed web services for Bioinformatics: protein disorder, conservation and RNA secondary structure. Bioinformatics *34*, 1939–1940. 10.1093/bioinformatics/bty045.

29. Troshin, P.V., Procter, J.B., and Barton, G.J. (2011). Java bioinformatics analysis web services for multiple sequence alignment—JABAWS:MSA. Bioinformatics *27*, 2001–2002. 10.1093/bioinformatics/btr304.

30. Crooks, G.E., Hon, G., Chandonia, J.-M., and Brenner, S.E. (2004). WebLogo: A Sequence Logo Generator. Genome Res. *14*, 1188–1190. 10.1101/gr.849004.

31. Berends, M.S., Luz, C.F., Friedrich, A.W., Sinha, B.N.M., Albers, C.J., and Glasner, C. (2022). AMR: An R Package for Working with Antimicrobial Resistance Data. J. Stat. Softw. *104*, 1–31. 10.18637/jss.v104.i03.

32. Usenik, A., Renko, M., Mihelič, M., Lindič, N., Borišek, J., Perdih, A., Pretnar, G., Müller, U., and Turk, D. (2017). The CWB2 Cell Wall-Anchoring Module Is Revealed by the Crystal Structures of the Clostridium difficile Cell Wall Proteins Cwp8 and Cwp6. Structure *25*, 514–521. 10.1016/j.str.2016.12.018.

33. Renko, M., Usenik, A., and Turk, D. (2017). Cwp6 from Clostridium difficile.

34. Freitag-Pohl, S., and Pohl, E. (2022). AmiP amidase-3 from Thermus parvatiensis.

35. Jasilionis, A., Plotka, M., Wang, L., Dorawa, S., Lange, J., Watzlawick, H., van den Bergh, T., Vroling, B., Altenbuchner, J., Kaczorowska, A.-K., et al. (2023). AmiP from hyperthermophilic Thermus parvatiensis prophage is a thermoactive and ultrathermostable peptidoglycan lytic amidase. Protein Sci. Publ. Protein Soc. *32*, e4585. 10.1002/pro.4585.

36. Zhang, R., Zhou, M., Bargassa, M., Joachimiak, A., and Midwest Center for Structural Genomics (2011). The crystal structure of the putative N-acetylmuramoyl-L-alanine amidase from Neisseria meningitidis.

37. Page, J.E., Skiba, M.A., Kruse, A.C., and Walker, S. (2022). Structure of the S. aureus amidase LytH and activator ActH extracellular domains.

38. Page, J.E., Skiba, M.A., Do, T., Kruse, A.C., and Walker, S. (2022). Metal cofactor stabilization by a partner protein is a widespread strategy employed for amidase activation. Proc. Natl. Acad. Sci. *119*, e2201141119. 10.1073/pnas.2201141119.

39. Tan, K., Mulligan, R., Kwon, K., Anderson, W., Joachimiak, A., and Center for Structural Genomics of Infectious Diseases (2014). The crystal structure of N-acetylmuramoyl-L-alanine amidase from Clostridium difficile 630.

40. Büttner, F.M., and Stehle, T. (2016). N-acetylmuramoyl-L-alanine amidase AmiC2 of Nostoc punctiforme.

41. Büttner, F.M., Faulhaber, K., Forchhammer, K., Maldener, I., and Stehle, T. (2016). Enabling cell–cell communication via nanopore formation: structure, function and localization of the unique cell wall amidase AmiC2 of Nostoc punctiforme. FEBS J. *283*, 1336–1350. 10.1111/febs.13673.

42. Tan, K., Rakowski, E., Buck, K., Joachimiak, A., and Midwest Center for Structural Genomics (2010). The crystal structure of a domain from N-acetylmuramoyl-l-alanine amidase of Bartonella henselae str. Houston-1.

43. Yang, D.C., Tan, K., Joachimiak, A., and Bernhardt, T.G. (2012). A conformational switch controls cell wall-remodelling enzymes required for bacterial cell division. Mol. Microbiol. *85*, 768–781. 10.1111/j.1365-2958.2012.08138.x.

44. Prigozhin, D.M., Mavrici, D., Huizar, J.P., Vansell, H.J., Alber, T., and TB Structural Genomics Consortium (2013). Structure of the Mycobacterium tuberculosis peptidoglycan amidase Rv3717 in complex with L-Alanine-iso-D-Glutamine reaction product.

45. Prigozhin, D.M., Mavrici, D., Huizar, J.P., Vansell, H.J., and Alber, T. (2013). Structural and Biochemical Analyses of Mycobacterium tuberculosis N-Acetylmuramyl-l-alanine Amidase Rv3717 Point to a Role in Peptidoglycan Fragment Recycling \*. J. Biol. Chem. *288*, 31549–31555. 10.1074/jbc.M113.510792.

46. Yamane, T., Koyama, Y., Nojiri, Y., Hikage, T., Akita, M., Suzuki, A., Shirai, T., Ise, F., Shida, T., and Sekiguchi, J. (2003). Structure of the catalytic domain of CwlV, N-acetylmuramoyl-L-alanine amidase from Bacillus(Paenibacillus) polymyxa var.colistinus.

47. Kumar, A., Kumar, S., Kumar, D., Mishra, A., Dewangan, R.P., Shrivastava, P., Ramachandran, S., and Taneja, B. (2013). Crystal structure of Rv3717 reveals a novel amidase from M. tuberculosis.

48. Kumar, A., Kumar, S., Kumar, D., Mishra, A., Dewangan, R.P., Shrivastava, P., Ramachandran, S., and Taneja, B. (2013). The structure of Rv3717 reveals a novel amidase from Mycobacterium tuberculosis. Acta Crystallogr. D Biol. Crystallogr. *69*, 2543–2554. 10.1107/S0907444913026371.

49. Prigozhin, D.M., Mavrici, D., Huizar, J.P., Vansell, H.J., Alber, T., and TB Structural Genomics Consortium (2013). Structure of the reduced, metal-free form of Mycobacterium tuberculosis peptidoglycan amidase Rv3717.

50. Prigozhin, D.M., Mavrici, D., Huizar, J.P., Vansell, H.J., Alber, T., and TB Structural Genomics Consortium (2013). Structure of the reduced, Zn-bound form of Mycobacterium tuberculosis peptidoglycan amidase Rv3717.

51. Blaise, M. (2020). crystal structure of the N-acetylmuramyl-L-alanine amidase, Ami1, from Mycobacterium abscessus bound to L-Alanine-D-isoglutamine.

52. Küssau, T., Van Wyk, N., Johansen, M.D., Alsarraf, H.M.A.B., Neyret, A., Hamela, C., Sørensen, K.K., Thygesen, M.B., Beauvineau, C., Kremer, L., et al. (2020). Functional Characterization of the N-Acetylmuramyl-l-Alanine Amidase, Ami1, from Mycobacterium abscessus. Cells *9*, 2410. 10.3390/cells9112410.

53. Blaise, M., and Alsarraf, H.M.A.B. (2020). Crystal structure of the N-acetylmuramyl-L-alanine amidase, Ami1, from Mycobacterium smegmatis.

54. Blaise, M. (2020). crystal structure of the apo form of the N-acetylmuramyl-L-alanine amidase, Ami1, from Mycobacterium abscessus.

55. Eckenroth, B.E., and Doublié, S. (2021). Structure of the CwlD amidase from Clostridioides difficile in complex with the GerS lipoprotein.

56. Feliciano, C.A., Eckenroth, B.E., Diaz, O.R., Doublié, S., and Shen, A. (2021). A lipoprotein allosterically activates the CwlD amidase during Clostridioides difficile spore formation. PLOS Genet. *17*, e1009791. 10.1371/journal.pgen.1009791.

57. Korndörfer, I.P., and Skerra, A. (2005). The crystal structure of the listeria monocytogenes bacteriophage PSA endolysin PlyPSA.

58. Korndörfer, I.P., Danzer, J., Schmelcher, M., Zimmer, M., Skerra, A., and Loessner, M.J. (2006). The Crystal Structure of the Bacteriophage PSA Endolysin Reveals a Unique Fold Responsible for Specific Recognition of Listeria Cell Walls. J. Mol. Biol. *364*, 678–689. 10.1016/j.jmb.2006.08.069.

59. Mayer, M.J., Garefalaki, V., Spoerl, R., Narbad, A., and Meijers, R. (2011). Catalytic domain of CD27L endolysin targeting Clostridia Difficile.

60. Mayer, M.J., Garefalaki, V., Spoerl, R., Narbad, A., and Meijers, R. (2011). Structure-Based Modification of a Clostridium difficile-Targeting Endolysin Affects Activity and Host Range. J. Bacteriol. *193*, 5477–5486. 10.1128/jb.00439-11.

61. Ho, J., Adeolu, M., Khadka, B., and Gupta, R.S. (2016). Identification of distinctive molecular traits that are characteristic of the phylum “Deinococcus-Thermus” and distinguish its main constituent groups. Syst. Appl. Microbiol. *39*, 453–463. 10.1016/j.syapm.2016.07.003.

# 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. 24 structures (AmiA includes) (PDBeFOLD)
4. Gram-negative
5. Gram-positive

## Appendix D: All validation scores for the multiple sequence alignment

**Heatmaps demonstrating A: the True Positive Rate (TPR) and B: the Positive Predictive value (PPV) for each pairwise alignment for 5 candidate AmiC structures**, where TPR is the ratio of sequence pairs aligned in both the sequences and structures compared to the total number of residues aligned in the structures, and PPV is the ratio of aligned sequence and structure residue pairs to the total number of residues aligned in the sequences.

**B**

**A**

A graph of numbers and a number

Description automatically generated with medium confidenceA graph of numbers and a bar graph

Description automatically generated with medium confidence

## Appendix E: Full Logo for the final Multiple Sequence Alignment

**WebLogo displaying conservation across the entire multiple sequence alignment, including conserved and less-conserved regions.** X-axis represents column of alignment. Y-axis represents probability of observing residue in that column (always sums to 1.0).

A chart of different colored letters

Description automatically generated

## Appendix F: Occupancy plots for each insertion region

Raw counts of per-species occupancy of regions I-1 to I-8, annotated by gram staining status. A=I-1, B=I-2, C=I-3, D=I-4, E=I-5, F=I-6, G=I-7, H=I-8.

**Bhhhhhhhhhhhhh**

A graph of numbers and columns

Description automatically generated A graph of different colored bars

Description automatically generated

**D**

**C**

**A**

A graph of numbers and columns

Description automatically generated A graph of numbers and columns

Description automatically generated

**F**

**E**

A bar graph with numbers and symbols

Description automatically generated A graph of numbers and graphs

Description automatically generated with medium confidence

**H**

**G**

A graph of numbers and a number of non gap columns

Description automatically generated A graph of numbers and columns

Description automatically generated

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>