THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

Enriching Sequence Space with Functional Temperature Data Retrieved from Metagenomic Data

Georgia Wyldbore

BIOX7004: Research Project A
Semester 1, 2024
43235779

Abstract

Background: Enzyme-based industrial processes are gaining increasing prominence in agricultural and pharmaceutical industries, among others, due to their environmental sustainability. They serve as an alternative for the existing processes which have, in the face of increased demand, resulted in increased energy consumption and harmful byproducts. Enzymes, however, often face limitations in industrial applications due to their sensitivity to the required harsh conditions such as high temperatures or extreme pH levels. As such, enzymes able to survive in these conditions are sought after both for use, and study of their properties to assist in protein engineering. Additionally, a wide range of metadata exists for the data stored in the databases of the National Centre for Biotechnology Information (NCBI), where submission templates were designed with the intent of increasing ease of access and facilitating novel research using existing data. Consequently, this research aimed to investigate the viability of using existing annotated metagenomic data to predict the functional temperature of proteins, in the context of identifying potentially thermostable proteins suitable for further study or use.

Methods: A pipeline was developed to retrieve all metagenomic data with associated temperature annotations from the NBCI databases, create a locally available Basic Local Alignment Search Tool (BLAST) database, perform BLAST searches into this database for proteins of interest and filter the output for near-identical matches. By finding an instance of a query protein in the database, a temperature can be assigned to it as a known functional temperature.

Results: Temperature annotations were able to be created for a range of proteins across three protein groups, showing the viability of using existing metagenomic data in this way. Annotations were an average of ~35°C lower than melting point as predicted by the DeepSTABp predictor. No single protein was assigned annotations both below 35°C and above 65°C, suggesting significant differences between the temperature groups which may be pertinent to future study.

Conclusions: The utilisation of existing metagenomic data is a viable process for identifying potentially thermostable proteins. The process would also be easily modifiable to generalise to a broader scope such as pH level, salinity, or colder temperatures. However, more metagenomic data is required to improve annotation frequency on specialised proteins.

Contents

Figures

1 Introduction

Industrial sectors are currently going through a period of rapid, increasing growth. Unfortunately, this growth has put immense pressure on environmental health due to the required increase in energy resources, and the creation of toxic waste and byproducts, or harmful emissions (Sharma et al., 2019). Consequently, there has been a shift towards identifying 'green' industrial technologies which are more sustainable, environmentally friendly, and closer to carbon neutral, without decreasing in efficiency or economic viability (Burkhardt et al., 2023; Sharma et al., 2019; Thomas et al., 2019).

In many cases, a popular option in facilitating this shift has been the use of enzymes to control and speed up certain chemical reactions. Examples of areas benefitting from this are the food, animal feed, textile, pharmaceutical and health industries, among others (Burkhardt et al., 2023). Enzyme-based industrial technologies allow the utilisation of raw materials whiles producing minimal to no waste, and avoiding toxic chemical usage (Sharma et al., 2019).

Enzymes are a type of protein which act as biological catalysts in the cells of living things. Like many aspects of biology, they are evolved to fulfil a role and so tend to have very specific and specialised functions (Modarres et al., 2016). As such, enzymes are often not naturally suited for industrial processes, especially as many of those processes are subject to harsh conditions which are inherently hostile to most enzymes, including extremes of pH, high temperatures, and high salinity. The use of enzymes in many applications is therefore limited to their functionality within these environments (Modarres et al., 2016; Sharma et al., 2019; Thomas et al., 2019). Due to these environmental limits, and the role specificity of natural proteins, many new enzymes are being engineered for customised purposes and environmental conditions. The aim of this engineering tends to be improving certain features of the enzyme, particularly in relation to its performance in harsh conditions (Modarres et al., 2016).

Some organisms, however, have evolved in such a way that their enzymes have naturally occurring features which allow them to perform under extreme conditions of temperature and pH, as well as high pressure, salinity, or solvent concentrations (Burkhardt et al., 2023; Elleuche et al., 2014; Krüger et al., 2018; Singh & Ray, 2021). Those enzymes which are able to function at high reaction temperatures of 45ºC or more specifically are described as thermostable, or alternatively referred to as thermophiles (Kurokawa et al., 2023). Generally, increasing temperature leads to protein unfolding and structural changes, with the internal hydrophobic amino acids becoming exposed to solvents and other hydrophobic amino acids, resulting in the unfolding becoming permanent (Modarres et al., 2016). Despite this, many thermophilic enzymes have evolved in organisms found in high-temperature environments worldwide such as hot springs or other geothermal processes, hot water pipes, and compost, and are able to retain structural stability at temperatures from 50ºC up to 120ºC (Burkhardt et al., 2023; Modarres et al., 2016). In their investigation regarding the metagenomes of hot spring environments, Burkhardt et al. (2023) highlight the likelihood of existing metagenomic data holding information about proteins suited to specific purposes and environments – despite the collection and sequencing being often unrelated to temperature factors – as well as the current under-utilisation of this available data.

The BioProject and BioSample databases were developed to facilitate the availability of metadata within the broader National Centre for Biotechnology Information (NCBI) database, in the hope of promoting reanalysis of data from different perspectives (Barrett et al., 2012). The implementation of these databases has provided a structured and reasonably consistent framework from which to filter relevant data for novel research. Despite the apparent ease of accessing project metadata, which can include temperature, there are currently few available tools providing information or predictions about the thermostability or functional temperature limits of a protein. One such tool, named 'Thermometer', has been developed using machine learning algorithms by Miotto et al. (2022), who acknowledge the lack of research addressing this area. Unfortunately, Thermometer requires the protein of interest to exist in the Protein Data Bank database, thereby requiring the structure of the protein to have been well-researched. Another is 'DeepSTABp', a deep learning predictor for melting point, requiring only a FASTA sequence, growth temperature, and selection of a 'cell' or 'lysate' flag (Jung et al., 2023). While DeepSTABp appears to be the leading predictor available, it requires information regarding protein conditions which may not be known.

Other tools appear to be in research, development, or testing (Kurokawa et al., 2023; Pudziuvelyte et al., 2024), but have not yet been published and as Modarres et al. (2016) stated, any tool which could provide hits about the thermostability of proteins out of the current collection of information would be of great interest (p. 115256).

Despite the validation of functional temperature of a protein requiring specific thermostability tests, the optimal growth temperature for an organism, and therefore its proteins, is easier to determine. The optimal growth temperature of a microorganism typically reflects its natural habitat (Kurokawa et al., 2023), and logically growth temperature will not exceed a maximum functional temperature. As such, habitat temperature can serve as a reasonable approximation of functional temperature. These predictions are then likely to be slightly lower than the true functional temperature, but provide a simple way to identify potentially thermophilic proteins.

While engineering and identifying thermostable enzymes is an active area of research, it appears that one under-researched aspect of the field is whether the thermostability improvement of an engineered protein is dependent on the thermostability of the base protein. While logically, it would seem that beginning from a stronger base would allow the greatest result, this does not appear to have been proven.

The main aim of this research undertaken in this project was to determine the viability of using the annotations of existing metagenomic data to make a prediction around the functional temperature of a protein, in the context of identifying potential thermostable proteins. This was investigated through creation of a pipeline, which was then tested using a subset of the Ketol-acid reductoisomerase (KARI) family of proteins. This family was chosen due to its known industrial uses in the production of biofuels, noted resistance to harsh conditions in certain organisms, and frequent presence in a range of organisms (Chen et al., 2018), and due to it being an active family of study within the research group. As part of the same branched-chain amino acid synthesis pathway, subsets of the dihydroxy-acid dehydratase (DHAD) and acetolactate synthase (ALS) protein families were used as a performance comparison for the pipeline. The accuracy of predictions was then assessed against predictions made by the DeepSTABp tool, visually and with a Wilcoxon rank-sum test.

2 Materials and Methods

A collection of scripts was written to perform each step of the created pipeline to perform annotations on the protein families of interest. An overview of the pipeline is displayed in Figure 1, and each step is explained in greater detail below. For the code used at each stage, refer to Appendix I.

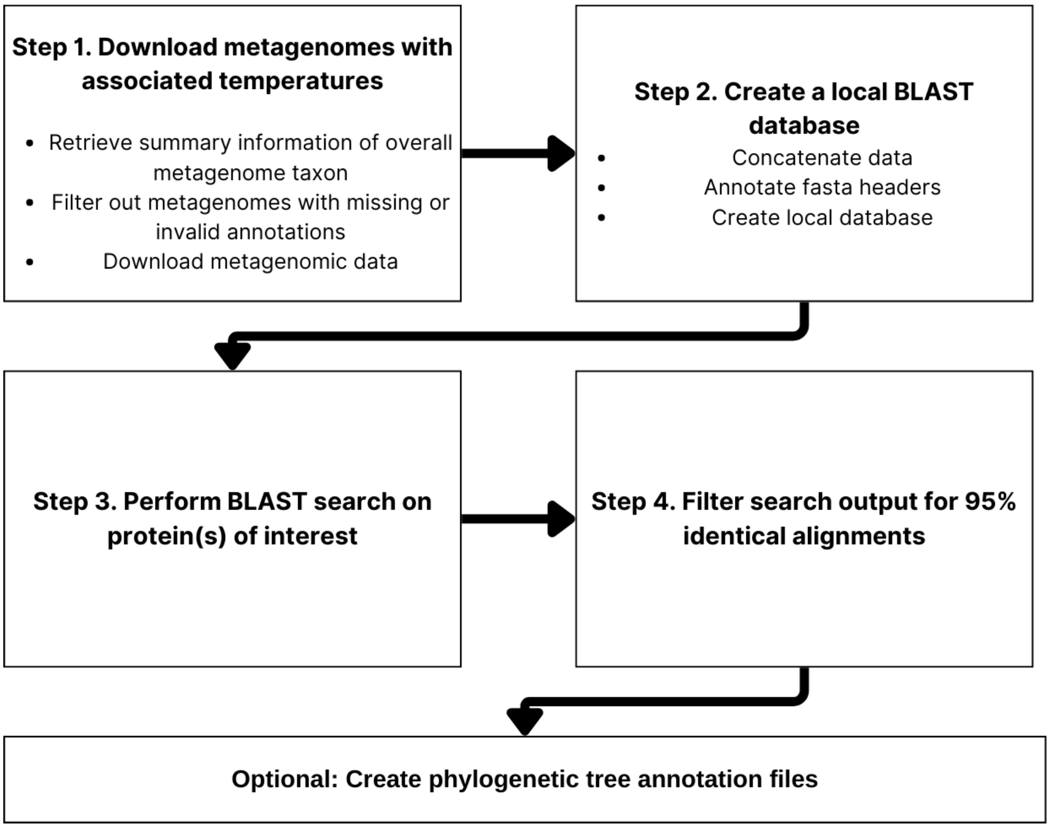

Step 1. Download metagenomes with associated temperatures
Retrieve summary information of overall metagenome taxon
Filter out metagenomes with missing or invalid annotations
Download metagenomic data
Step 2. Create a local BLAST database
Concatenate data
Annotate fasta headers
Create local database
Step 3. Perform BLAST search on protein(s) of interest
Step 4. Filter search output for 95% identical alignments
Optional: Create phylogenetic tree annotation files

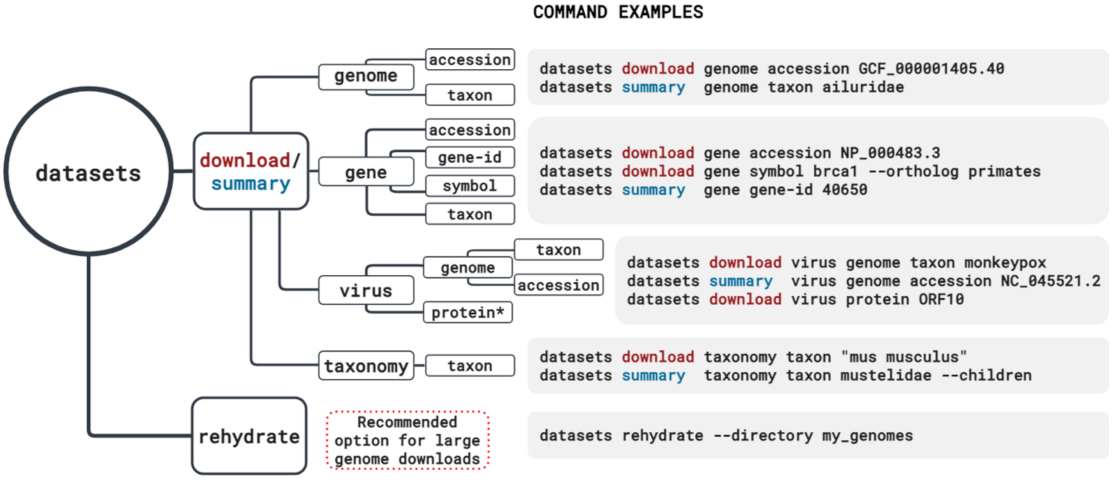
Figure 1 - Created pipeline for determining the predicted functional temperature of a protein family

2.1 Download metagenomes – Step 1

The first major step in the pipeline was to download the metagenome sequences which had associated temperature annotations. To do this, the summary information of all metagenomes of interest had to be retrieved, these had to be filtered for useable data, and then each genome’s sequence file had to be downloaded.

The initial step was to identify and retrieve the metagenome taxon of interest.

On the NCBI website, all metagenomes are stored under one large taxon. This is then further split into ecological, organismal, simulated and synthetic metagenomes. The choice was made to use ecological metagenomes only to allow for reduced scope, while having the highest chance of data having temperature annotations. NCBI provides a tool for downloading information and data from their stores, known as Datasets ((NCBI), 2024a; Sayers et al., 2021), version 16.6.0. This allows users to retrieve summaries or sequences of genomes, genes, viruses and taxonomies, as seen in the examples in Figure 1. By using the taxon ID number – 410657 – the data report for all metagenomes under the ecological metagenome taxon could be retrieved for investigation. The information for 7508 metagenomes were retrieved by this process.



```
datasets download genome accession GCF_000001405.40
datasets summary  genome taxon ailuridae

datasets download gene accession NP_000483.3
datasets download gene symbol brca1 --ortholog primates
datasets summary  gene gene-id 40650

datasets download virus genome taxon monkeypox
datasets summary  virus genome accession NC_045521.2
datasets download virus protein ORF10

datasets download taxonomy taxon "mus musculus"
datasets summary  taxonomy taxon mustelidae --children

datasets rehydrate --directory my_genomes
```

* virus protein: restricted to download of SARS-Cov-2 proteins

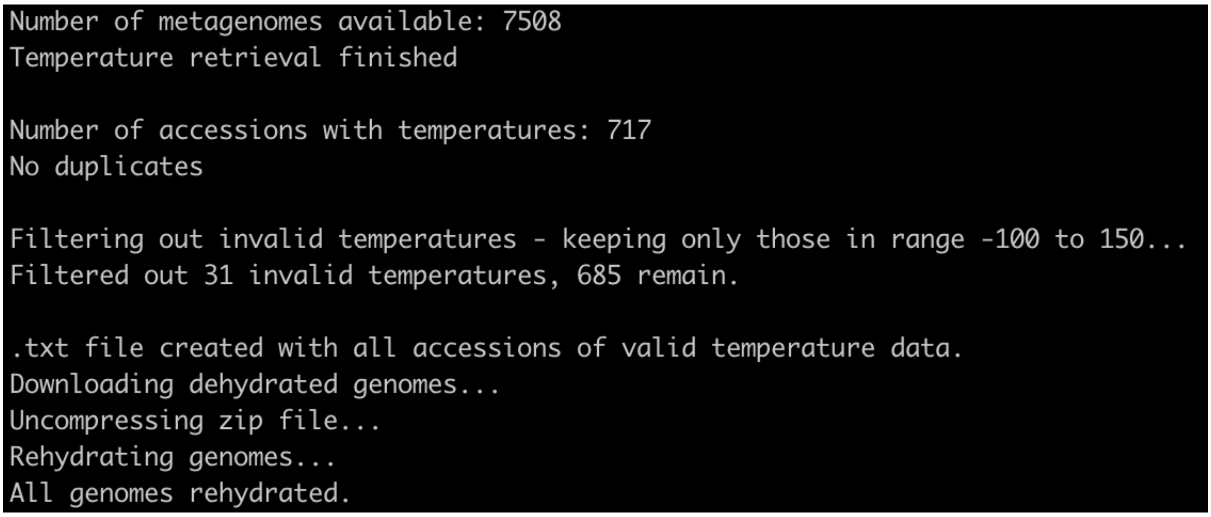
Figure 2 - Usage summary of the NCBI datasets command line tool ((NCBI), 2024a)

With the genome information now contained in a JSON file, the next step was to filter out those metagenomes which had missing or unusable temperature annotations. The information returned through the datasets tool included the BioProject and BioSample information, which is how temperature annotations were accessed. A script was written to loop through each metagenome's information and retrieve these annotations. A visual inspection revealed that temperature was stored in two different ways; the majority had a 'temp' field in the BioSample attributes, but a handful of metagenomes stored temperature and pH values under 'isolation_source', which was taken into account. The metagenome accession number and associated temperature were written to a dictionary, and to a .txt file for inspection and later reference. Of the 7508 metagenomes, 717 had temperature annotations.

Upon inspecting the created .txt file, it was clear that the annotations lacked consistency. The template descriptions provided by BioSample (Barrett et al., 2012)include units, and state that 'missing' or similar should be entered if the value was not recorded. However, in this situation some entries included units, some not, and some contained unreasonable entries such as 99999, -999 and CO2. The assumption was made that all entries were in degrees Celsius, and anything outside a temperature range of -100 to 150 was considered null data. These values were arbitrary but selected for being outside the range of thermal limits for life apparent in Clarke (2014). It is not clear whether a value of 0 would have been considered null or a genuine value as it is a commonly used proxy for having no data. As values either side of zero were found, but without a zero itself, this conflict of use did not require consideration. Of the 717 metagenomes with temperature annotations, thirty-one of them were found to have invalid data, leaving 685 for further use. Another .txt file was created here with the valid remaining metagenome accession numbers and temperatures for later use, and another with the accessions for use in downloading the data.

The final component to Step 1 was to download the FASTA files of contigs for each metagenome. The NCBI datasets tool was once again used for this, using the 'download genome accession' parameters and the file of accession numbers created previously. The guidelines for downloading large genome data packages suggest that for large downloads, the dehydrate flag should be used, which downloads a zip archive containing metadata and the location of the data on the NCBI servers (((NCBI), 2024b). This can then be rehydrated to retrieve the data. As almost seven hundred metagenomes were being downloaded, the dehydrate flag was used, and the commands to unzip the archive and rehydrate the genomes were included in the script, along with progress flags for each step as visible in Figure 2.

This initial script to complete step 1, while specific to this use case, could be easily modified to alter the metagenomes searched to a more specific range of the ecological metagenomes if required.



```
Number of metagenomes available: 7508
Temperature retrieval finished

Number of accessions with temperatures: 717
No duplicates

Filtering out invalid temperatures - keeping only those in range -100 to 150...
Filtered out 31 invalid temperatures, 685 remain.

.txt file created with all accessions of valid temperature data.
Downloading dehydrated genomes...
Uncompressing zip file...
Rehydrating genomes...
All genomes rehydrated.
```

Figure 3 – Terminal progress output from running the script written to perform Step 1 of the pipeline.

2.1 Create a local BLAST database – Step 2

Step two of the pipeline was to create a Basic Local Alignment Search Tool (BLAST) database locally with the identified metagenomes.
The makeblastdb tool ((NCBI), 2008; Camacho et al., 2009) (ver. 2.9.0) allows the creation of a local database using specified FASTA sequences. The commands for makeblastdb appear to only allow one file to be specified as an input, so the FASTA files for each genome were concatenated into one large FASTA file to allow for input.
When testing Step 3 of the pipeline, and attempting to interpret outputs of a BLAST search, it was discovered that several processing steps were required to link a contig containing a matched sequence back to its accession number and temperature annotation. However, it was a simple step to retrieve the header of the contig containing the match. As such, the decision was made to annotate the headers of each contig with the metagenome accession and temperature, where possible, to allow all this information to be contained in one place.

Within each metagenome's FASTA files, the header contained information about the contig, generally including an identifier number, the name of the metagenome, the contig number and information about the genome. For example, a header may be formatted as:

>LCWZ01112965.1 Anaerobic digester metagenome contig112932, whole genome shotgun sequence

The identifier number is linked to a Whole Genome Shotgun (WGS) project, with the first four letters identifying a project, the first two numbers the version of that project, and the remainder identifying that individual contig ((NCBI), 2020). Hence, the example above would belong to version one of the project LCWZ. This WGS identifier was available in the original JSON file downloaded in Step 1, and a mapping was created between accession and WGS identifier using this, as well as between accession and temperature using the text file created in Step 1. With the assistance of BioPython's (Cock et al., 2009) (ver. 1.78) SeqIO, the metagenome accession and temperature were added to the sequence headers. Initially, this caused several KeyErrors, and to avoid this a try-except block was used which printed the identifier for which a map could not be found. The script was run with the output being written to an error file for later sorting.

Two main incorrect assumptions were the source of the KeyErrors. First, the assumption was incorrectly made from inspection that all genomes would have the four-letter project identifier; any genome added to the WGS project database after 2019 actually uses a six-letter identifier. Second, the assumption that all contigs were formatted the same way. A small collection of metagenomes associated with hot springs projects did not include the WGS identifier in the FASTA header at all.

While it turned out that these hot springs files did in fact include the genome accession number, at the time it was unclear how they could be mapped back and so they were left unannotated. Those where the WGS key was the new, longer version were stored in a file separately with their key, accession and temperature, for processing later. This error handling stage of the pipeline could have been optimised further, as will be discussed later, however this was not possible be due to time constraints.

With the FASTA headers annotated, the BLAST database could be made using makeblastdb, ensuring the 'nucl' and 'parse_seqids' flags were used to identify that it was a nucleotide database, and that the header information of each sequence was correctly interpreted.

2.3 Perform a BLAST search – Step 3

The third step of the pipeline was to perform a BLAST search, in order to identify sequences in the database with high similarity to proteins of interest. As the database contains nucleotide sequences, but proteins are used as queries, a translated nucleotide ‘tblastn’ search was performed.

BLAST allows for a high level of customisation in searches and their outputs, with around twenty-five different flags available to specify, and twenty-nine options for output fields ((NCBI), 2008; Camacho et al., 2009). The specific parameters used for this investigation were:

- -outfmt “7 qseqid sseqid pident ppos evalue bitscore salltitles”
- -evalue 1e-100
- -max_target_seqs 10000
- -num_threads 4

The outfmt flag specifies the format of the output file and the fields to include within it. In this case, 7 represents a tabular format, and the following parameters represent respectively the query sequence ID, subject sequence ID, percentage identity, percentage of positive-scoring matches, e-value for the alignment, bit score for the alignment, and the title of the subject sequence – i.e. the FASTA sequence header annotated in Step 2. As these searches were returning a very large number of results for each query sequence, an e-value cutoff was specified using the evalue flag. The max_target_seqs flag was set to 10,000 after trial and error on the KARI protein family, as a value which was not reached for any query sequence. The num_threads flag allowed for each search to be run using multiple threads, to aid in speeding up processing time.

Initially, the file containing all of the 716 KARI proteins was used as the input for the BLAST search. This took approximately five days to run. As such, the choice was made to split the file into multiple smaller FASTA files, to enable parallelisation. A tool for this purpose, aptly named FASTA splitter, was found online in the form of a Perl script (Kryukov, 2017). The number of files to split into was chosen arbitrarily to be a multiple of five which allowed for less than 100 sequences in each file.

To allow for generalisability of the pipeline in future, a Bash script was written to perform the BLAST searches on one or more specified input FASTA files. To attempt to optimise performance and runtime of these searches, the script was set up to allow up to five searches to be running in parallel. As soon as one search finished, the next was able to begin, and a progress statement was delivered to the terminal with which file was being processed. This was achieved by using an array of process ID numbers, with a loop continually checking whether a process was finished and removing it from the array if so. This script also ensured that output files were placed in the directory from which the script was run, not where the script was located, ensuring that protein family information could be kept within its own folders and the script did not need to be linked into that folder each time.

2.4 Filter search output – Step 4

The final step of the basic pipeline was to filter the hits returned by the BLAST search for those which have a high identity match to a query sequence. This was to ensure that temperatures were not being assigned to queries unless the two sequences were almost identical, increasing the chances that the match is representing the same protein in the database. This increased chance in turn represents an increase in the likelihood that the query protein could be found at the same temperature as the metagenome containing the match, and that temperature can be assigned as a functional temperature for the query. 95% was chosen as a value close to 100%, but with a small margin of error for insertions, deletions, mutations or substitutions.

Once again, a script was written to perform this step. As with Step 3, the desire was to maximise the generalisability of this script. In this case, the user specifies first an output file name, and one or more input files. The output will then contain informative lines beginning with a # character, including the query name, number of hits the sequence had in total, and number that are above 95% identity. This is followed by the information from Step 3's BLAST output for that database hit on a new line, separated by tabs. If a query has no hits above the 95% threshold, it will not be included in the output file. The threshold of 95 is also easily updatable within the file by changing one line of code.

While the output file of this script is human readable, a further script was written to perform an extra step of parsing. This extra step provided a format which included only the query sequence identifier and the temperatures at which it had high identity matches, and which could be used to pass into an existing script to produce an annotation file in the optional Step 5. Once again, for generalisability, this final script allowed specification of a csv output file name and one or more input files to parse, with the expectation these input files were formatted in the way the first script of this step provides. It also expected a flag of 'highest', 'lowest', or 'all', specifying whether only the highest temperature, lowest temperature, or all possible temperatures should be written to the output for each query. Additionally, this step handled accessing the temperatures of those contigs which caused KeyErrors during annotation. The exception is for those which did not provide a WGS project identifier in the FASTA header, and in these cases a warning message was provided to the user with relevant information to perform a manual search.

2.5 Create tree annotation files

While not part of the pipeline, the final step in the process for this investigation was to annotate the phylogenetic tree for the KARI, DHAD and ALS protein families. This was done through modifying existing code, which makes use of Interactive Tree of Life (iTOL) templates (Letunic & Bork, 2024), to take in a custom colour map which better allowed visualisation of the functional temperature assigned to each protein. A further small file was written to reformat the identifiers of the DHAD protein family as they were formatted slightly differently within their tree.

Given the focus of this research on thermophilic proteins in particular, and the distribution of temperatures in the database (see Figure 6), a binning approach was used to assign colours to lower temperatures, while higher temperatures each received their own colour. As the colder temperatures ranged from around -3 to 35, this was able to split neatly into eight temperature bins, in shades of blue (Figure 4). One representing the negative values, then one for each group of five degrees. An additional bin was added to cover the large gap between the cold and hot temperatures, though in this instance nothing was mapped to it as no temperatures *t* existed such that $35 < t < 65$. Having these nine bins allowed for enough variation in colour between them to be easily distinguishable by eye. For the hotter temperatures, which ranged from 65 to 82, each of the seven values was assigned its own colour in a yellow to red gradient (Figure 5).

The choice to bin the colder temperatures but not the hotter was backed by the context of this research. A difference in functional temperature of five degrees when in a cooler range is unlikely to affect a protein much, as this is variation seen in temperatures day-to-day in many climates. However, a small difference at very hot temperatures may be the catalyst for a protein to begin unfolding, and so grouping these together may obscure some of the data. The annotated trees can be seen in Section 3, Figures 11, 12 and 13.

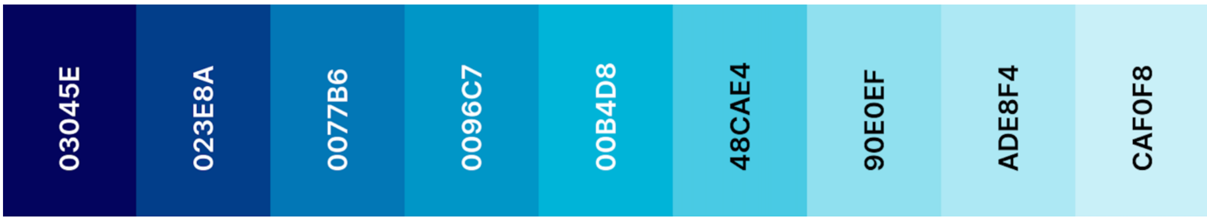


Figure 4 - Colour palette with HEX codes for cold (<65º) temperatures. Darkest blue represents -3 - 0, then increasing in 5 degree increments, with the final colour representing 36 – 65. Bins are inclusive of lower temperature, and exclusive of upper.

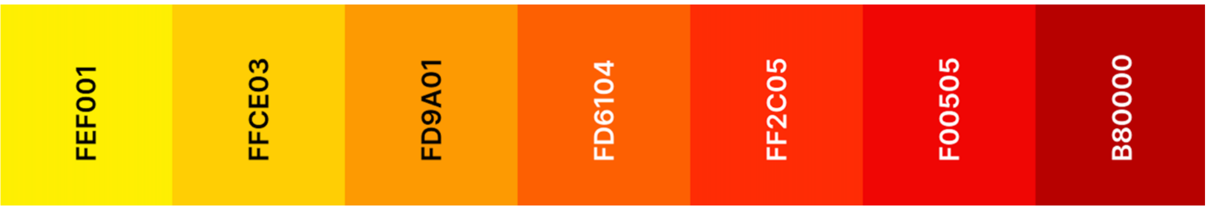


Figure 5 - Colour palette with HEX codes for hot (≥65º) temperatures. Lightest yellow represents 65, followed respectively by 70, 71, 75, 80, 81 and 82.

2.6 Validation

The accuracy of predictions made by the created pipeline was tested against predictions made by the DeepSTABp tool (Jung et al., 2023), using the pipeline prediction as the growth temperature parameter and selecting the ‘cell’ flag. The difference in resulting temperature predictions was assessed using a Wilcoxon rank-sum test due to the non-normal distribution of the data, and the data was visualised using scatterplots which can be found in Section 3, Figures 14 and 15.

3 Results

As the focus of this research was primarily on assessing the viability of using existing data annotations to make predictions about the functional temperatures of proteins, there are relatively few experimental results to report. However, statistics pertaining to the distribution of temperature data, and the BLAST search results are presented here, along with the annotated phylogenetic trees for the three investigated protein families.

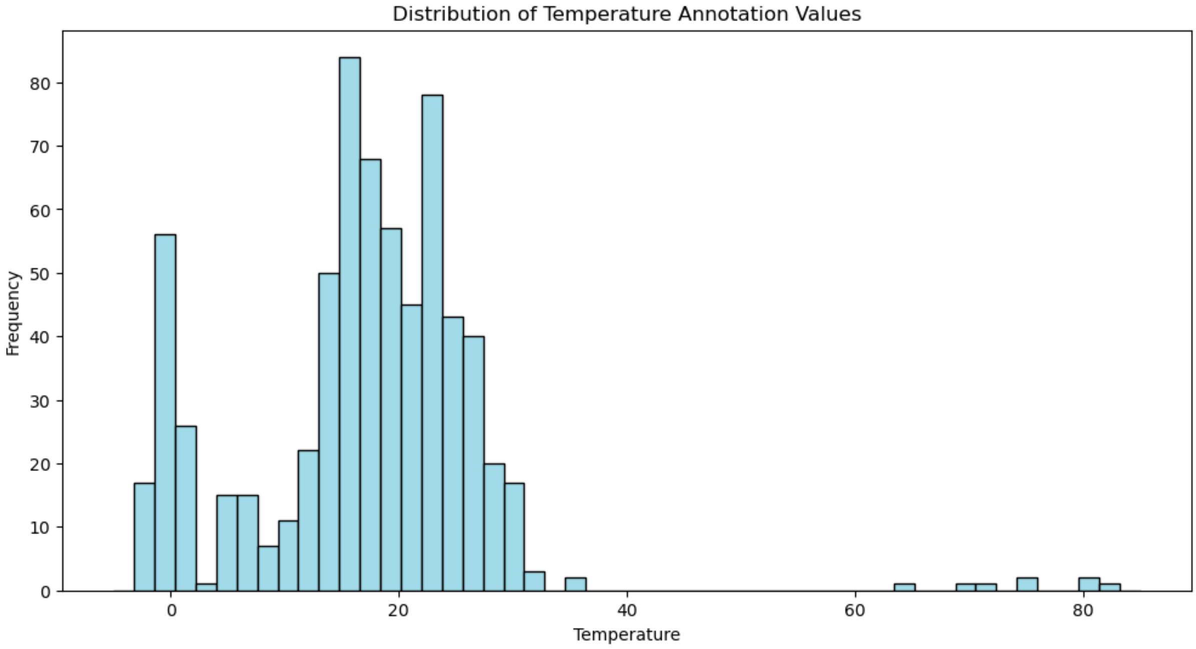


Figure 6 - Distribution of temperature annotations, in ºC, of ecological metagenomes after filtering for invalid values

Figure 6 here shows the distribution of temperature values for the ecological metagenomes which had associated temperature annotations, after removing the presumed invalid values. The temperatures were presumed to be in degrees Celsius, though units were not stated for many values. This figure clearly shows that a high proportion of the metagenomes were taken from environments with temperatures below 40º, with only seven metagenomes having values above 60º.

	KARI	DHAD	ALS
Number of extant proteins known to be in the family	716	1658	1990
Number of proteins with no BLAST search hits	43	0	2
Highest number of total BLAST search hits for a single protein	7980	10073	9835
Number of proteins with ≥1 high identity BLAST search hits	45	101	98
Number of proteins with exactly one high identity BLAST search hit	10	21	63
Highest number of high identity BLAST search hits	85	88	64
Highest number of total BLAST search hits for proteins with ≥1 high identity result	7980	10070	1225
Number of times a protein's only BLAST search hit was of a high identity	0	0	0

Table 1 - Summary of notable values from results, highlighting proportions of total and high identity hits for each protein family

Table 1 gives a summary of some of the notable values which are not immediately apparent from the figures presented here. Notably, it provides context as to how large each protein family is, the proportion of proteins which had hits of at least 95% identity – referred to interchangeably as 'high identity' results, and the ranges of how many search results were assigned to various proteins. This allows a comparison of how the BLAST database performed across the three protein families, highlighting the fact that the number of proteins given temperature annotations is not directly proportional to the size of the protein family.

Figures 7 and 8 represent the distribution of hit counts returned in the BLAST searches for individual query sequences, grouped by protein family. Figure 7 looks at the number of hits returned for all proteins queried, that is, each protein sequence in each of the three families. Figure 8 then looks specifically at those query proteins which had at least one alignment returned with an identity match of over 95%, referred to interchangeably as a 'high identity' result.

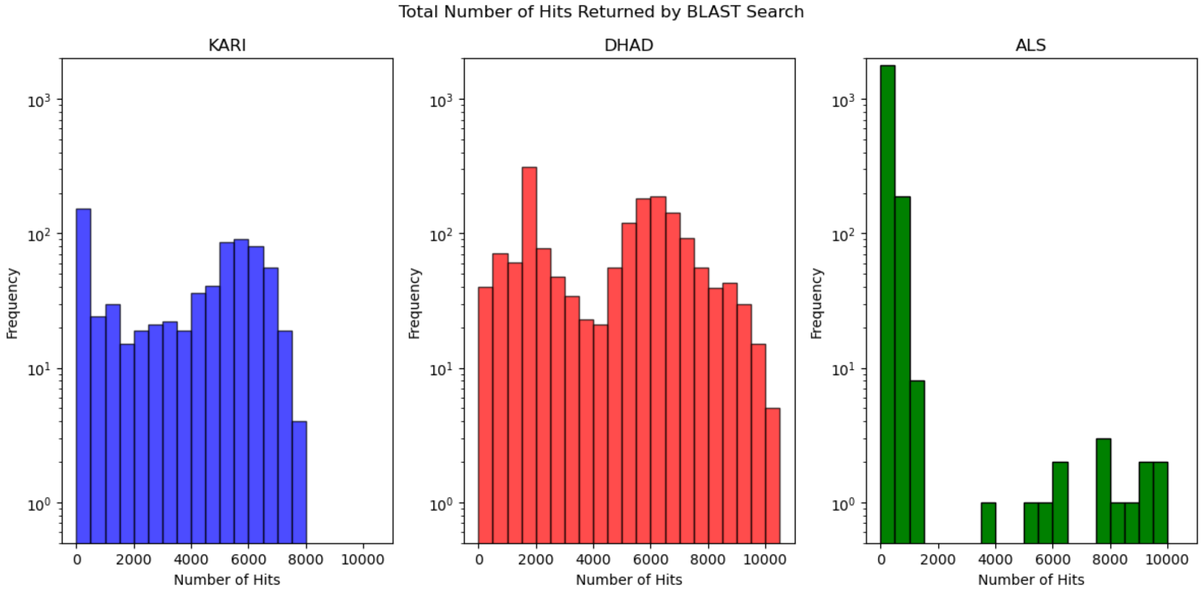


Figure 7 - Distribution of the total *number of BLAST search hits returned for individual protein queries, grouped by protein family*

Due to the significantly higher frequency of ALS proteins returning very few hits in BLAST searches when compared to the other two families, the y-axis frequency is presented on a log10 scale in Figure 7 to aid readability. In Figure 8, the x-axis scale is limited to a maximum of 1300 for the ALS family, as no proteins with high identity matches had a total number of matches exceeding this. Using a log10 transformation on this x-axis was considered, but ultimately discarded due to the resulting decreased readability; hence a reduced scale was used instead. Note the similarity in distribution shape between Figures 7 and 8, which suggests that the total number of hits does not indicate the likelihood of a protein having a high identity hit.

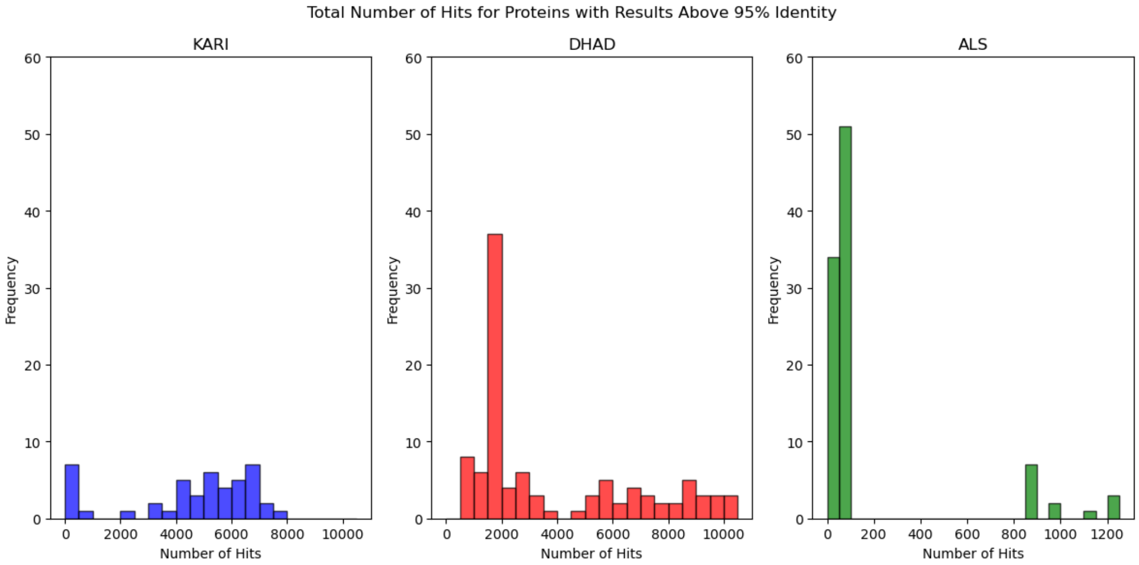


Figure 8 - Distribution of the total number of BLAST search hits for individual protein queries which had at least one high identity result, grouped by protein family
Note: ALS x-axis has a reduced scale for improved readability of results

Figure 9 then shows the distribution of the number of high identity hits a protein received in its BLAST search. This highlights the difference between the distributions of the number of high identity hits and the total number of hits shown in Figure 8, as these two distributions are made with the same proteins. Specifically, this dissimilarity between Figures 8 and 9 suggests that the likelihood of finding a high identity hit does not increase proportionally with the total number of hits. The positive skew of each family's plot in Figure 9 also shows that very few high identity matches are likely to occur for a given protein, which may indicate the existence of a singular 'best match' in the database in many cases.

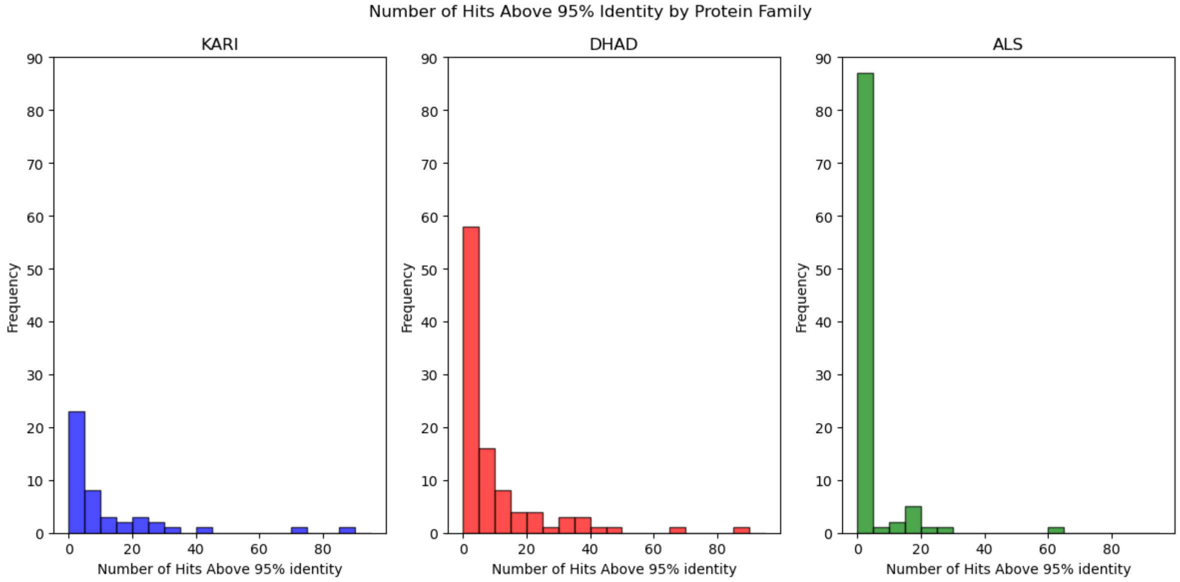


Figure 9 - Distribution of the number of BLAST search results which had ≥95% identity to an individual protein query, grouped by protein family

Figure 10 represents the same data as Figure 9, but with each protein classified based on the temperature value of its hits. For each protein, a temperature category was assigned to it based on whether the BLAST search returned high quality hits where all the temperatures were <65º, ≥65º, or a combination of the two. These categories are represented on the graph by blue, orange and green respectively.

What this illustrates is that the matches with a high identity returned for any given protein can be grouped as either high temperature, or low temperature, as no protein had matches in the green 'mixed' category. This is further demonstrated in Table 2, which provides a numerical summary of the totals within each category, as well as the ranges displayed in Figures 11, 12 and 13.

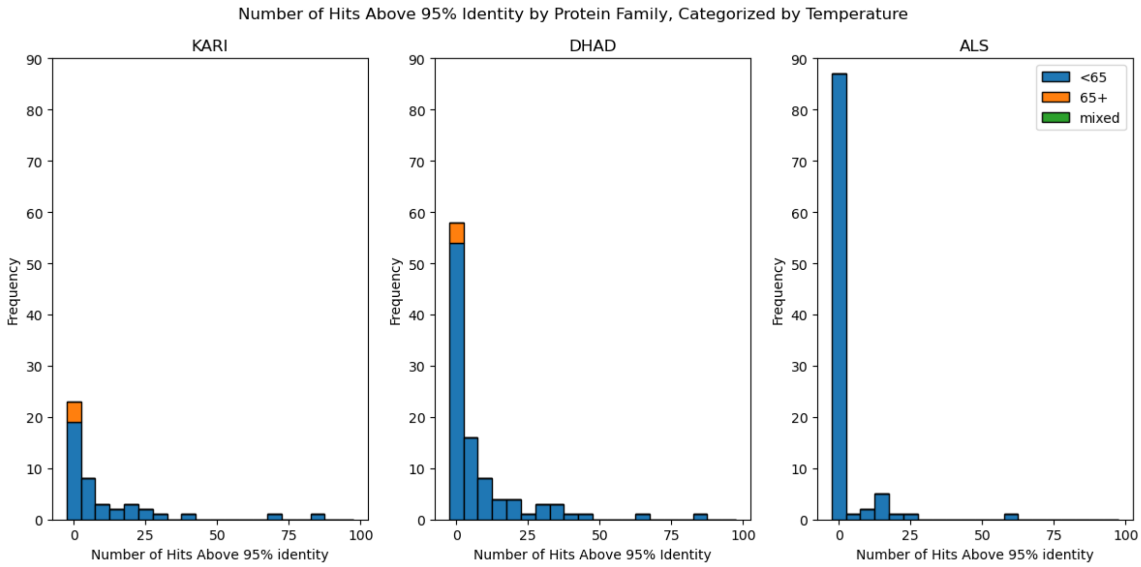


Figure 10 - Distribution of the number of high identity BLAST search results for a protein query, classified by the temperature of those results, and grouped by protein family

	KARI	DHAD	ALS
Number of proteins with only 'cold' (<65º) high identity BLAST search results	41	97	98
Number of proteins with only 'hot' (≥65º) high identity BLAST search results	4	4	0
Number of proteins with both 'hot' and 'cold' high identity BLAST search results	0	0	0

Table 2 - Summary of number of proteins with results in each temperature category - hot (≥65º), cold (<65º) or mixed - grouped by protein family

Figures 11, 12 and 13 show the annotated phylogenetic trees for each of the three protein families. The colours leading from the node to the outside, along with the inner section of the strips on the outside, are coloured by the temperature of the hottest high identity hit found in the BLAST search for that protein. The outer strip on the outside is then coloured by the lowest temperature of the high identity hits for that protein. The highest temperature was used for the strip between the node and the identifier on the outside, as given the initial context of thermophilic proteins for this research, the hotter temperature was of more interest. Initially, only annotations for the highest temperatures for each protein were created for the same reason, however the addition of the coldest temperature allows for an easy visual representation of the range of temperatures at which any protein is predicted by the pipeline to exist. Links to downloadable versions of Figures 11, 12 and 13 are available in Appendix II, alongside a link to all three interactive trees on the iTOL website (Letunic & Bork, 2024).

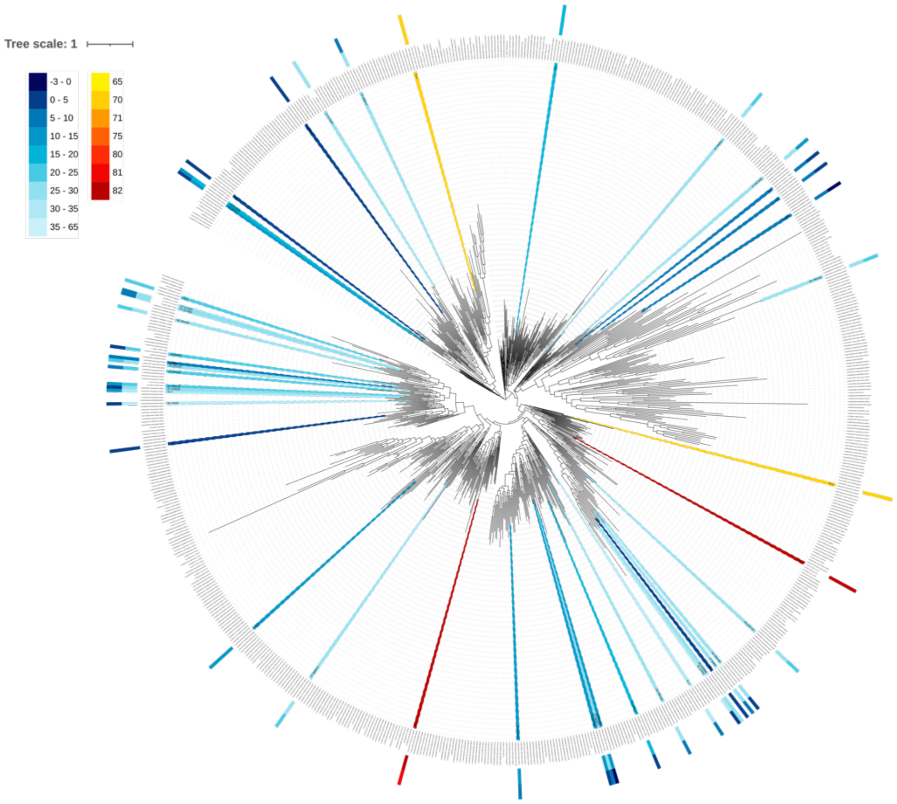


Figure 11 - KARI family phylogenetic tree, with coloured annotations for the hottest (inside) and coldest (outside) temperatures associated with high identity BLAST hits for each protein

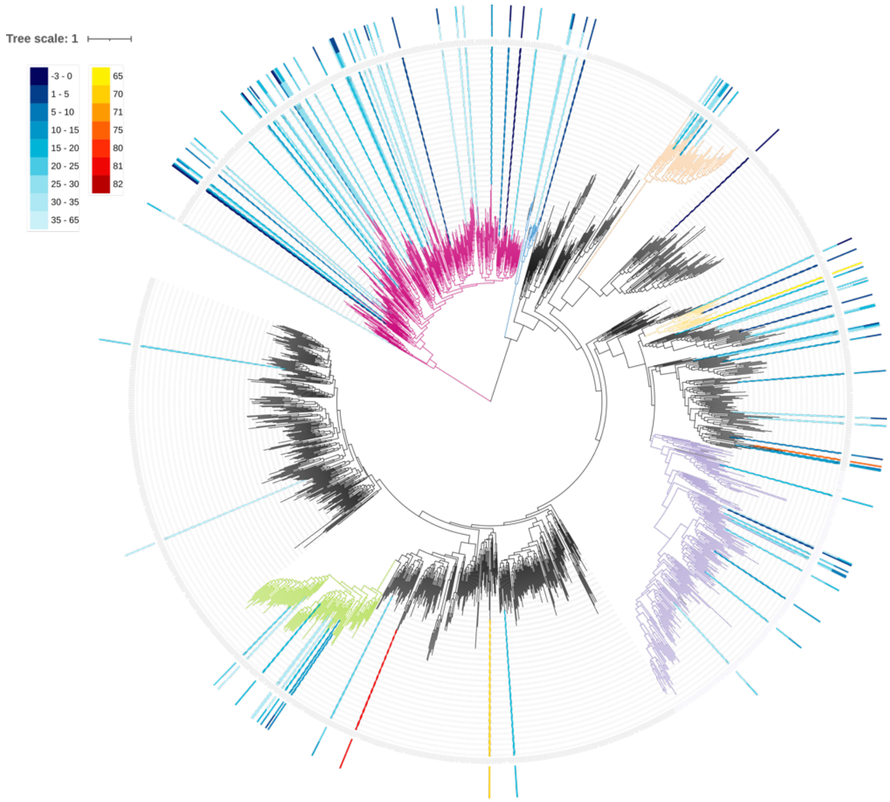


Figure 12 - DHAD family phylogenetic tree, with coloured annotations for the hottest (inside) and coldest (outside) temperatures associated with high identity BLAST hits for each protein

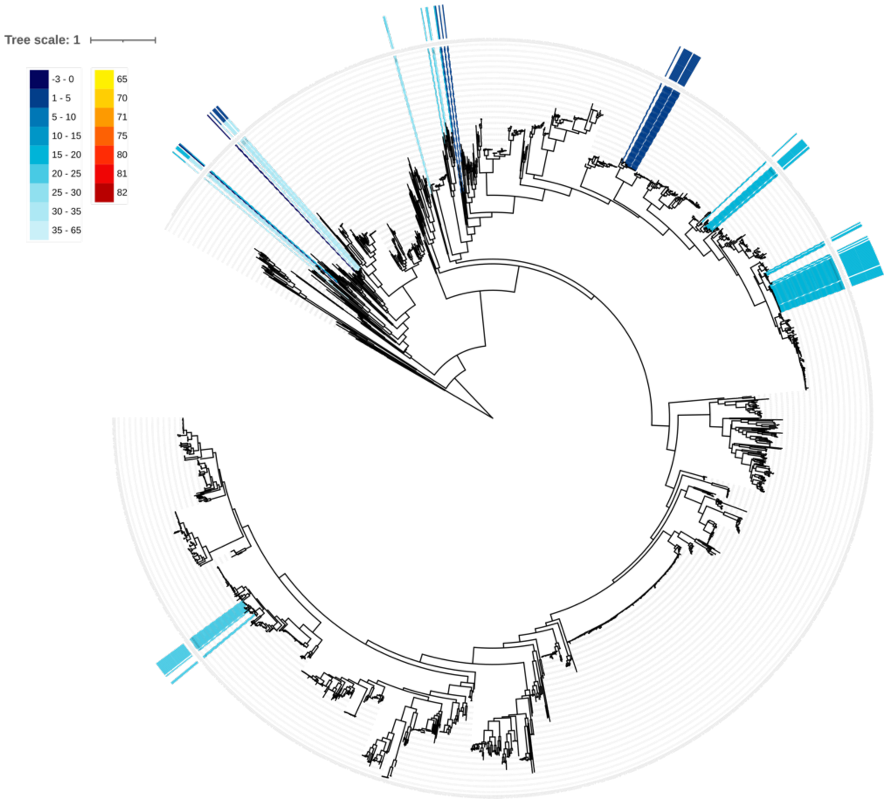


Figure 13 - ALS family phylogenetic tree, with coloured annotations for the hottest (inside) and coldest (outside) temperatures associated with high identity BLAST hits for each protein

A point of interest to note in these phylogenetic trees are the reinforcement of the point made by Figure 10 and Table 2, that the highest and lowest annotated temperatures for any protein are both within the same low or high range. Additionally, and perhaps surprisingly, there is relatively little clustering by temperature in either the KARI proteins in Figure 10, or the DHAD proteins in Figure 11, with the few high temperature annotations being scattered somewhat sporadically. Meanwhile, the ALS proteins in Figure 13 display very clear clustering of annotations, with all the annotations in a cluster coloured the same shade, so indicating an annotation range of at most five degrees for that cluster. The precise temperature values for the maximum or minimum temperature annotation for any protein can be seen on the full, interactive versions of each tree.

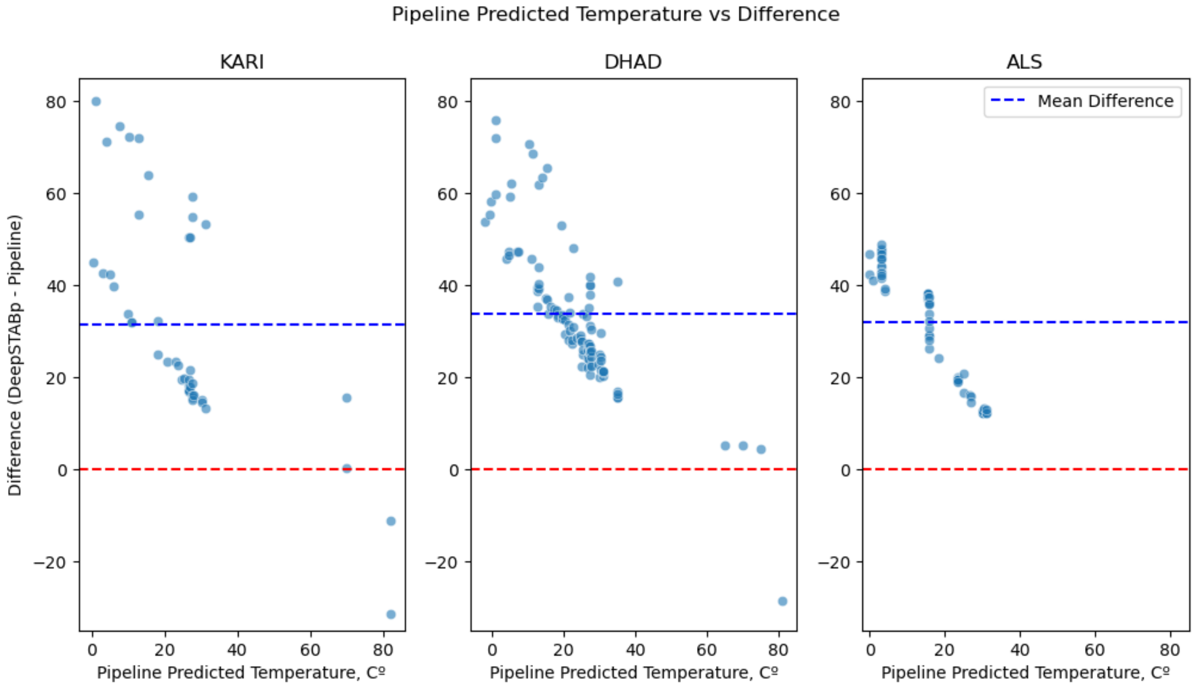


Figure 14 - Temperature predicted by pipeline against difference in prediction between pipeline and DeepSTABp methods

Figure 14 takes the temperatures predicted by the pipeline presented in this report, and by the DeepSTABp tool (Jung et al., 2023), and plots the pipeline predictions against the difference between the two, for each family. This clearly shows that the pipeline has an average temperature prediction around 35º lower than DeepSTABp's prediction of melting point, with temperatures towards the extremes of the scale having the highest margin of error. Figure 15 displays the relationship between the temperatures of the two predictive methods, highlighting the narrower range of predictions made by DeepSTABp as well as the lack of correlation between the two predictive methods.

The results of the Wilcoxon rank-sum test can be seen for each protein family in Table 3, validating the results seen in Figure 14. The relatively low test statistic for all three families indicates the differences are predominantly in one direction, with the p-value being less than 0.05 in all cases, indicating the differences are statistically significant.

	KARI	DHAD	ALS
Test statistic	27.0	43.0	0.0
p-value	7.17e-11	9.55e-18	8.33e-18

Table 3 - Results of the Wilcoxon rank-sum test on predicted temperature differences

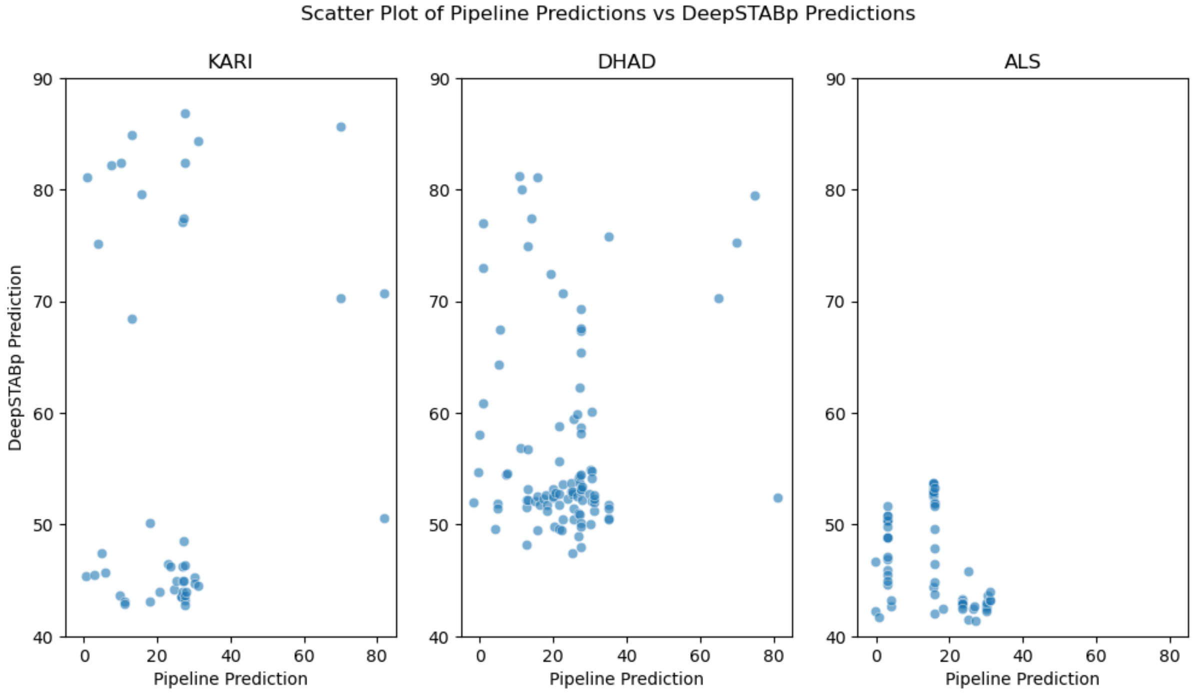


Figure 15 - Scatterplot of the temperature values predicted, in ºC, by the pipeline and DeepSTABp methods for each annotated protein

4 Discussion

The main aim of this research, as stated earlier, was to determine the viability of using existing data to predict an approximation of the functional temperature of a given protein. While this was intended to specifically be in the context of identifying potentially thermostable proteins, the results have implied applications to a broader scope of uses.

Foremost, the results show that it is possible to use the existing metagenomic data from the NCBI database to assign a temperature to a variety of proteins using the constructed pipeline. This provides proof of concept for using metadata to perform novel investigations. Further, given the proteins in the database are shown to exist at their respective temperatures and that a protein of interest must be at least 95% identical to those in the database, it is not unreasonable to assume that the protein of interest can be functional at that temperature as a minimum. While this does not necessarily provide the maximum functional temperature, it does provide a temperature at which the protein is assumed to be functional, as well as an estimation of optimal growth temperature(Kurokawa et al., 2023). While not the focus of this research, this secondary estimation may have other use cases within research.

Further, the created pipeline has a wide scope for generalisability and further development. While originally written only to test the KARI proteins, it was successfully able to generalise to other families of different sizes, and would be functional for single proteins, whole families, or files of unrelated protein sequences, due to the scripts being written in such a way as to allow input variation[1]. It would also be easily adaptable to search within a more specialised set of ecological metagenomes, to annotate information about proteins found at colder temperatures, to adjust key filters or thresholds, or to investigate other metadata relevant to industrial enzymes such as pH or salinity. The ability of the pipeline to have been constructed in such a way indicates that this approach could be developed into a

[1] The author wished to bring particular attention to this point, as developing a script with such input flexibility while retaining efficiency was something not previously attempted. As such, this aspect of the research and pipeline development provided a significant learning opportunity.

bioinformatic tool for wider use, providing applications for research in a broader range of areas.

While the focus of this research was not a comparison of the performance of the temperature annotation method across the KARI, DHAD and ALS protein groups, but rather to determine method viability, the comparison between the three does provide insight into the pipeline as it currently stands. This in turn gives an indication of the use limits requiring attention.

Across all three families, the total number of hits returned for an individual protein was commonly low, though with a small increase round 6000 results – both peaks are visible in Figure 7. When this same total hit distribution is shown for only those proteins which had high identity hits (Figure 8), the shape of the distribution remains generally consistent with the exception of the ALS proteins, having a maximum number of hits around 1250. This distribution similarity suggests having a high number of lower-identity hits does not increase the chance of a high identity hit being found. Rather, the early peak and apparent frequency decline towards the upper limit of hits implies that many queries were identifying specific matches in favour of generic regions of conservation, increasing the likelihood of the identified matches yielding valid predictions. The positive skew of the number of high identity matches shown in Figure 9 further supports this implication of quality annotations, as a high number of matches here may have simply indicated a region of conversation rather than a true instance of a similar protein.

Figures 11, 12 and 13 illustrate the potential for differences in data availability for closely related proteins through the clustering of annotations. While the KARI protein family in Figure 11 shows relatively minimal clustering by temperature, and the DHAD proteins (Figure 12) cluster only slightly more in some clades, the ALS proteins occur almost only in clusters (Figure 13). This may suggest several things; perhaps KARI and DHAD proteins are more dissimilar to their close neighbours on the tree than ALS proteins, perhaps the ALS proteins are more specialised and have evolved to work in a specific environment, or perhaps there are some other functional constraints which affect the ALS proteins more heavily than the other families. As the initial step in branched-chain amino acid synthesis, it is not unreasonable to speculate that ALS proteins must specialise to individual environmental conditions(Takano et al., 2023), with the suggestion of ALS being more highly specialised supported by several other factors. A higher number of ALS proteins were included than in either of the DHAD or KARI groups, but proportionally the fewest high identity hits and resulting temperature annotations were returned. Figure 13 also shows the reduced variation in maximum and minimum annotations for ALS proteins in comparison to KARI or DHAD. Additionally, a majority of the ALS proteins had very few hits in the BLAST search both overall, and for those with high identity. While the clustered annotations displayed for the ALS proteins suggests that nodes within a clade may be very similar to each other, it could also suggest that clades are more distinct from each other.

If groups of, or individual, proteins are more specialised and distinct, the likelihood of finding a matching instance of that protein in the assembled database decreases, and with it the chance of assigning a temperature annotation. To informally test this, three query proteins were searched in the database. These were copies of heat shock 90, reverse gyrase, and an antifreeze protein, taken from prokaryotic organisms and representing proteins which are commonly found, specialised to hot temperatures, and specialised to cold temperatures respectively. Aligning with expectations, the heat shock protein found around 6000 matches overall, and several high identity matches, while reverse gyrase and antifreeze found 60 and 0 matches respectively overall, and no high identity matches. Also as expected, the reverse gyrase protein had the majority of its lower identity matches within the metagenomes from high temperatures. While not rigorous or proof of statistical significance, these informal tests support an overall suggestion that the pipeline is potentially ineffective for specialised proteins, as they are less likely to have near-identical matches within the database. Despite this, any temperature annotation which can be found for a specialised protein may carry more certainty with it, as it is then more likely to be a true match.

Given the nature of the three protein groups' related purposes in branched-chain amino acid synthesis, it is reasonable to expect all three to be identified in a singular metagenome in the instance that any one is identified. Figures 11 through 13 show that this is not the case, as the ALS proteins received no annotations of high temperatures. While this raises questions about the accuracy of the temperature annotations, it may also be a result of the suggested specialised nature of the ALS group, or indicate that the particular ALS protein which works with the high temperature KARI and DHAD pairs identified was not included in the group used in this research. With very few high temperature annotations overall, it is difficult to isolate a reason for this discrepancy.

The considerably smaller set of data available for the high temperature genomes is likely to be significantly impacting the proportion of high temperature annotations. The distribution of temperatures available in the data, shown in Figure 6, is heavily skewed towards lower values. In all three families the majority of annotations being made are between -3ºC and 35ºC - much the same range as demonstrated by the peak in Figure 6. If the assumption that ALS proteins are more specialised is upheld, then it is possible that there is a trio of proteins found at high temperatures, but that not enough high temperature data was available to find a suitably close match for the ALS protein. In contrast, under this same assumption, the KARI and DHAD proteins are more generalisable and so more readily able to find matches within the high temperature metagenome. The addition of more high-temperature metagenomes to databases would aid in removing ambiguity in results, both existing and future. While the existing data was abundant enough to demonstrate validity of the pipeline approach, only approximately ten percent of the available metagenomes contained annotation, with this figure reducing to nine percent after invalid values were removed. The BioProject and BioSample databases (Barrett et al., 2012) have made annotating and accessing data easier for instances such as this, but perhaps researchers need to be encouraged to record as much metadata as possible when collecting their project samples, to better facilitate the re-examination of data intended by the BioProject database. Increasing the frequency of annotations may also allow the temperature distributions to become somewhat less skewed, which in turn would increase the reliability of annotations produced by the pipeline.

While possible that the KARI and DHAD high temperature annotations were the result of a lower temperature misclassification, the clear split between temperature groups makes this seem unlikely. Figure 10 and Table 2 illustrate clearly that when a protein has any high identity BLAST search result, every high identity result for that protein will fall in the same temperature category. This is also displayed by the trees in Figures 11, 12 and 13, where the two bars on the outside never display colours of both groups for a single protein.

While some results do show a wide range of colder temperatures, for example the upper left group on the ALS tree in Figure 13, this was not investigated further due to the thermophilic context of this research.

The clear temperature grouping into high and low groups suggests that the proteins found at hot temperatures are likely to be sufficiently different to those from lower temperatures. This increases confidence in predictions, and indicates viability in using the pipeline as an initial method to identify potentially thermostable proteins. One application of this may be to narrow down large families to a small number of proteins for initial research, reducing research expenses. As the temperature difference presumably has an identifiable cause in the sequences themselves, this may also be of further use to research in predicting protein temperature based on sequence alone.

In an effort to validate the accuracy of the predictions, the DeepSTABp melting point prediction tool was utilised(Jung et al., 2023). The variation in predictions between the pipeline and DeepSTABp values can be seen in Figures 14 and 15, showing a consistent and statistically significant under-prediction by the pipeline of around 35ºC across all three protein groups. This under-prediction was expected, as mentioned in Section 1, as the temperature at which a protein is found is unlikely to exceed the melting point of that protein. The prediction differences were greatest at the extremes of those provided by the pipeline, with more stability around the 10ºC to 30ºC range of pipeline predictions. This coincides with the highest density of metagenome temperatures, so may indicate simply that the higher abundance of proteins found around these temperatures allowed DeepSTABp to predict with more certainty on proteins existing around this range. The only instance in which the pipeline predicts a higher temperature than DeepSTABp is above 80ºC, which calls the melting point predictor accuracy at this range into question. As the pipeline is presumed to have identified an instance of the query protein existing at its annotated temperature, considering the earlier discussion on the improbability of misclassifying a high temperature, the predicted melting point here seems unreasonable.

As these two methods predict slightly different things, and as the accuracy of DeepSTABp is not fully known on proteins outside its training set, it is hard to use this validation as more than a guideline. However, the overall results align with those expected of the pipeline.

Once again, addition of more annotated metagenomes to the database would aid in removing the ambiguity around accuracy here.

Several limitations were identified with the processes in the pipeline, which could be improved upon in the future. The most major of these was that alignment length was not considered within the BLAST search results. One of the initial stretch goals of this research was to look at creating a tree from the high identity hits for several proteins to identify whether results were grouping by temperature. Once the temperature clustering of hits within the KARI family became apparent, this became less necessary and so was omitted in favour of investigating the generalisability of the pipeline to other families. However, while performing local tests on a database subset, it was discovered that the length of the local alignments in results was varying between around 10 and 300 amino acids. It is questionable whether a result which only aligns a very small alignment with a high identity could be considered as valid a prediction as one which matches almost the whole protein, though the argument could be made that perhaps only a small, conserved region is important in some instances. The discovery of this limitation unfortunately occurred too late in the research process for the pipeline scripts to be altered at the various affected steps. Future research with this pipeline should consider this factor, and potentially include an investigation into reasonable lengths and conserved regions.

Another limitation identified was in the process of the pipeline itself. The errors which occurred during Step 2 of the pipeline were handled by creating a separate file, allowing a mapping between various features of each metagenome such as genome accession number, WGS project identifier, or temperature. This mapping was then used to assign temperatures during Step 4. The pipeline could be refined to deal with this mapping during the annotation of the FASTA headers before creating the database, or could be used as an alternative to annotating the FASTA sequence headers at all. The annotation of the headers with the genome accession and temperature allowed for human readability in the process of creating the pipeline, but would be unnecessary if this were to be developed into a tool as users would not generally require access to that information at that stage of the pipeline. Some errors were also not able to be accounted for even in the parsing step and required manual lookup of the temperature due to a lack of clarity in the existing header information. This is

likely rectifiable but would need a re-evaluation of how the genome-temperature lookup is performed.
The final limitation is that organism was not able to be considered. While possible that some matches identified were based on a common organism between query and database, it is unclear how this may affect the results. It is possible that the identity threshold may need to be adjusted if organism is a driving factor behind protein similarities to allow for matching irrespective of organism. However, a matching organism would also result in a high degree of accuracy in terms of a protein existing at its annotated temperature. As such, organism distribution is something which should be considered in future research.

With the viability of this pipeline temperature annotation method demonstrated, along with the potential broader uses, many opportunities for further research present themselves. One possibility would be to expand upon and validate the current results by determining the accuracy of the predictions when compared to the true stable temperature of a protein, as determined in a lab. Thermostability predictor tools using machine learning are also an option for this validation. These tools could not be used in this instance as due to the somewhat under-researched nature of the protein families used in this project, the correct format of protein sequence was unavailable to perform those tests.
This leads to a possible extension to this research: using the differences between the sequences annotated with different temperature values through this pipeline to predict annotations – and as such approximate thermostability – for those proteins without annotations. The defined split between the high and low temperature annotations suggests an identifiable variation between sequences which could be leveraged to make predictions. Another opportunity could be to investigate how this pipeline could be adapted to cater for ancestral sequences. Ancestral sequences tend to be more thermostable(Thomas et al., 2019), but it seems likely that as the ancestral sequences will not appear in the database, finding a match to them may require alteration of the identity threshold. The pipeline could also be adapted to investigate other extreme environments such as pH or salinity, along with any other field included in the submitted metadata.

5 Bibliography

(NCBI), N. C. f. B. I. (2020). *Whole Genome Shotgun Submissions*. https://www.ncbi.nlm.nih.gov/genbank/wgs/

(NCBI), N. C. f. B. I. (2024a). *datasets*. https://www.ncbi.nlm.nih.gov/datasets/docs/v2/reference-docs/command-line/datasets/

(NCBI), N. C. f. B. I. (2024b). *Download large genome data packages*. https://www.ncbi.nlm.nih.gov/datasets/docs/v2/how-tos/genomes/large-download/

(NCBI), N. C. f. B. I. U. (2008, (2008 Jun 23) (Updated 2021 Jan 7)). *BLAST® command line applications user manual: Building a BLAST database with your (local) sequences*. https://www.ncbi.nlm.nih.gov/books/NBK569841/

Barrett, T., Clark, K., Gevorgyan, R., Gorelenkov, V., Gribov, E., Karsch-Mizrachi, I., Kimelman, M., Pruitt, K. D., Resenchuk, S., Tatusova, T., Yaschenko, E., & Ostell, J. (2012). BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. *Nucleic Acids Research*, *40*(Database issue), D57-D63. https://doi.org/10.1093/nar/gkr1163

Burkhardt, C., Baruth, L., Neele, M.-H., Klippel, B., Margaryan, A., Paloyan, A., Panosyan, H. H., & Antranikian, G. (2023). Mining thermophiles for biotechnologically relevant enzymes: evaluating the potential of European and Caucasian hot springs. *Extremophiles*, *28*(1), 5. https://doi.org/10.1007/s00792-023-01321-3

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics*, *10*, 421. https://doi.org/10.1186/1471-2105-10-421

Chen, C. Y., Ko, T. P., Lin, K. F., Lin, B. L., Huang, C. H., Chiang, C. H., & Horng, J. C. (2018). NADH/NADPH bi-cofactor-utilizing and thermoactive ketol-acid reductoisomerase from Sulfolobus acidocaldarius. *Sci Rep*, *8*(1), 7176. https://doi.org/10.1038/s41598-018-25361-4

Clarke, A. (2014). The thermal limits to life on Earth. *International Journal of Astrobiology*, *13*(2), 141-154. https://doi.org/10.1017/S1473550413000438

Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., & de Hoon, M. J. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*(11), 1422-1423. https://doi.org/10.1093/bioinformatics/btp163

Elleuche, S., Schroder, C., Sahm, K., & Antranikian, G. (2014). Extremozymes--biocatalysts with unique properties from extremophilic microorganisms. *Curr Opin Biotechnol*, *29*, 116-123. https://doi.org/10.1016/j.copbio.2014.04.003

Jung, F., Frey, K., Zimmer, D., & Muhlhaus, T. (2023). DeepSTABp: A Deep Learning Approach for the Prediction of Thermal Protein Stability. *Int J Mol Sci*, *24*(8). https://doi.org/10.3390/ijms24087444

Krüger, A., Schafers, C., Schroder, C., & Antranikian, G. (2018). Towards a sustainable biobased industry - Highlighting the impact of extremophiles. *N Biotechnol*, *40*(Pt A), 144-153. https://doi.org/10.1016/j.nbt.2017.05.002

Kryukov, K. (2017). *FASTA Splitter (Version 0.2.6) [Computer software]*. In http://kirill-kryukov.com/study/tools/fasta-splitter/

Kurokawa, M., Higashi, K., Yoshida, K., Sato, T., Maruyama, S., Mori, H., & Kurokawa, K. (2023). Metagenomic Thermometer. *DNA Res*, *30*(6). https://doi.org/10.1093/dnares/dsad024

Letunic, I., & Bork, P. (2024). Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. *Nucleic Acids Res*. https://doi.org/10.1093/nar/gkae268

Miotto, M., Armaos, A., Di Rienzo, L., Ruocco, G., Milanetti, E., & Tartaglia, G. G. (2022). Thermometer: a webserver to predict protein thermal stability. *Bioinformatics*, *38*(7), 2060-2061. https://doi.org/10.1093/bioinformatics/btab868

Modarres, H. P., Mofrad, M. R., & Sanati-Nezhad, A. (2016). Protein thermostability engineering. *RSC Advances*, *6*(116), 115252–115270. https://doi.org/10.1039/C6RA16992A

Pudziuvelyte, I., Olechnovic, K., Godliauskaite, E., Sermokas, K., Urbaitis, T., Gasiunas, G., & Kazlauskas, D. (2024). TemStaPro: protein thermostability prediction using sequence representations from protein language models. *Bioinformatics*, *40*(4). https://doi.org/10.1093/bioinformatics/btae157

Sayers, E. W., Beck, J., Bolton, E. E., Bourexis, D., Brister, J. R., Canese, K., Comeau, D. C., Funk, K., Kim, S., Klimke, W., Marchler-Bauer, A., Landrum, M., Lathrop, S., Lu, Z., Madden, T. L., O'Leary, N., Phan, L., Rangwala, S. H., Schneider, V. A., . . . Sherry, S. T.

(2021). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, *49*(D1), D10-D17. https://doi.org/https://doi.org/10.1093/nar/gkaa892

Sharma, S., Vaid, S., Bhat, B., Singh, S., & Bajaj, B. K. (2019). Thermostable enzymes for industrial biotechnology. In *Advances in Enzyme Technology* (pp. 469–495). https://doi.org/10.1016/b978-0-444-64114-4.00017-0

Singh, K. K., & Ray, L. (2021). Extremozymes: biocatalysts from extremophilic microorganisms and their relevance in current biotechnology. In B. B. Mishra, S. K. Nayak, S. Mohapatra, & D. Samantaray (Eds.), *Environmental and Agricultural Microbiology: Applications for Sustainability*. https://doi.org/10.1002/9781119525899.ch14

Takano, H. K., Benko, Z. L., Zielinski, M. M., Hamza, A., Kalnmals, C. A., Roth, J. J., Bravo-Altamirano, K., Siddall, T., Satchivi, N., Church, J. B., & Riar, D. S. (2023). Discovery and Mode-of-Action Characterization of a New Class of Acetolactate Synthase-Inhibiting Herbicides. *J Agric Food Chem*, *71*(47), 18227-18238. https://doi.org/10.1021/acs.jafc.3c03858

Thomas, A., Cutlan, R., Finnigan, W., van der Giezen, M., & Harmer, N. (2019). Highly thermostable carboxylic acid reductases generated by ancestral sequence reconstruction. *Commun Biol*, *2*, 429. https://doi.org/10.1038/s42003-019-0677-y

Appendix I – Pipeline Code

All code used for the pipeline can be accessed on github at

https://github.com/gwyldbore/BIOX7004_report

The commands to run the pipeline are provided below.

```
$ python retrieve_all_genomes.py
$ python make_blast_db.py
$ python proj_acc_temp.py > error_all_info.txt
$ perl fasta-splitter.pl --n-parts <n> --nopad <input_file>
$ python get_identities_above_95.py <output_file> <filename1> [<filename2> ...]
$ python parse_results.py -highest|lowest|all <output_file> <filename1> [<filename2>..]
```

To create iTOL annotation files, use cells within the relevant Jupyter Notebook.

If working with DHAD proteins, format_dhad.py may be required.

Appendix II – iTOL Trees

Image download links

KARI

https://itol.embl.de/export/2031211985680931716939560

DHAD

https://itol.embl.de/export/2031211985673801716939375

ALS

https://itol.embl.de/export/2031211985682971716939602

Link to access interactive trees

https://itol.embl.de/shared/2moAfhE4RQqVp