Group Project: New substitution matrices and compare with BLOSUM62

Michael Hammel, Daniel Giesel, Stephan Liu

20th January, 2023

Introduction

Throughout evolution, biological mutations occur which are able to alter and adjust the function of many parts of a cell, from regulatory process, structural features and many more. Due to the inherent complexity of a ground truth model to evolution, simpler models have been devised to aid in evolutionary biology [1]. One such model is the substitution matrix. Substitution matrices are stochastic matrices which try to model evolution by placing a "cost" on a certain mutation happening, from one character state to another e.g between the amino acids A->C in proteins. This is achieved by taking a multiple sequence alignment (MSA) and computing the observed and background frequencies (happens due to chance) of the amino acids.[2].

As a result of evolution, organisms need to and have adapted to many different environments. The adaptation can be based on a variety of different factors: nutrition, water availability and one in particular - temperature [3]. As such, various organisms can and have been classified according to their optimum growth temperatures and the temperature of the environment they live in. An example of such classes are psychrotroph, mesophiles and thermophiles. [4]. As a result of these environmental adaptations come modifications and alterations in the DNA (and therefore protein) sequences [5]. This is especially visible in proteins with highly important cellular functions, regions of which tend to be conserved across species. Due to redundancies in the sequences (codons) which encode amino acids, as well as a preference to conserve structure rather than sequence, the actual composition of these proteins can vary [6].

As such, our project explores the effect of the conserved protein PcrA in prokaryotes of different temperature classes. Specifically, we look at its affect on the amino acid composition and in turn the resulting substitution matrix. We compare our temperature-specific substitution matrices to one of the very often BLOSUM matrices (in particular, BLOSUM62) .

Materials & Methods

Calculation of Substitution Matrix

BLOSUM(BLOcks SUbstitution Matrix) matrices [7] are used to score similarity of amino acids in a multiple sequence alignment. To calculate the substitution score for an amino acid pair BLOSUM uses blocks within the sequences of homologous proteins. The log-odds values are obtained using the following formula:

$$S_{ij} = \left(\frac{1}{\lambda}\right) log \left(\frac{p_{ij}}{q_i \times q_j}\right), \tag{1}$$

with p_{ij} indicating the expected occurrence for each i,j pair in the alignment and q_i, q_j being the frequency of the amino acids in general.

Obtaining sequences

We decided to use PcrA, a helicase involved in plasmid roilling-circle replication and DNA-repair, as our protein to calculate a substitution matrix. PcrA is located inside of the cell, which reduces the impact of external factors other than temperature to a minimum. PcrA is also well studied and a large number of sequences for this protein can be found in the corresponding databases. Using ProtDB, we queried the database for the PcrA homologs in various bacterial species. For each sequence, we looked up the species optimal growth temperature and ordered them into the three different groups. For each group, a FASTA file was created which includes 30 Sequences of PcrA-homologs.
Additionally, we utlisied the GSHT dataset providing information of organisms and their optimum growth temperatures [8] to obtain species names for each group. We then used the species name as a parameter for the `entrez_get_fasta.sh` script (via the command `entrez_get_fasta.sh <Organism Name>`) located in the Supplementary Materials to obtain a FASTA file of PcrA (if available) for that organism.

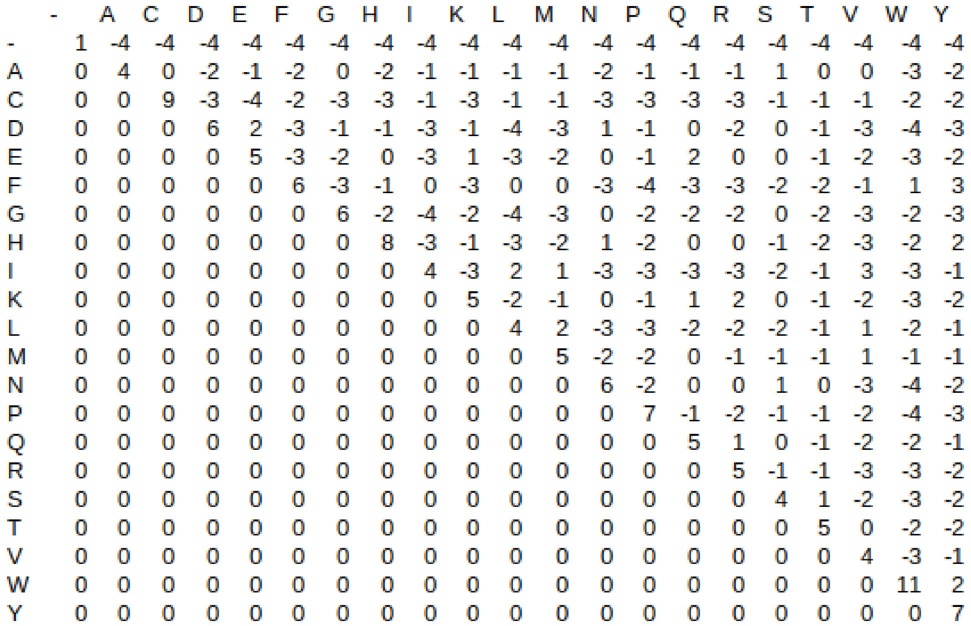
Multiple Sequence Alignment

Using MEGA11 [9] (**v11.0.11**, Homepage), we generated a multiple sequence alignments using the 30 sequences for each temperature group, resulting in 3 `".fas"` files.

Quality control, similarity and filtering

To prevent duplicates of sequences and impact of proteins which the database falsely showed as PcrA homologs, we filtered out sequences that were to similar to each other or too distant from all other sequences in the set. By creating a UPGMA-tree using MEGA11 (**v11.0.11**, Homepage), we calculated the phylogenetic distance of the individual samples and reduced groups of sequences which are too close to each other to a single sequence to represent that group. Very few sequences which had very long inserts (300bp) were also removed.

Results



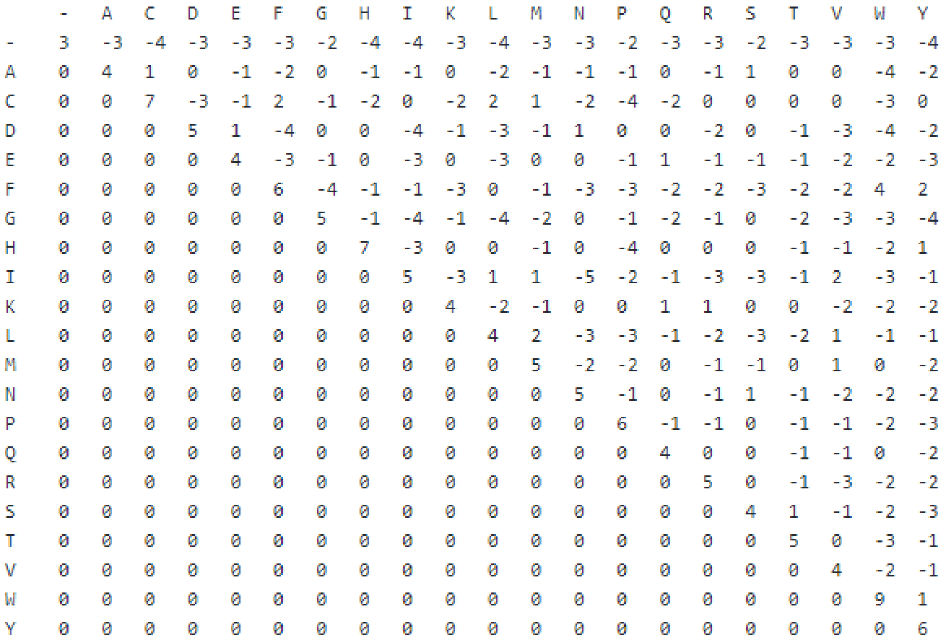
	-	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
-	1	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4
A	0	4	0	-2	-1	-2	0	-2	-1	-1	-1	-1	-2	-1	-1	-1	1	0	0	-3	-2
C	0	0	9	-3	-4	-2	-3	-3	-1	-3	-1	-1	-3	-3	-3	-3	-1	-1	-1	-2	-2
D	0	0	0	6	2	-3	-1	-1	-3	-1	-4	-3	1	-1	0	-2	0	-1	-3	-4	-3
E	0	0	0	0	5	-3	-2	0	-3	1	-3	-2	0	-1	2	0	0	-1	-2	-3	-2
F	0	0	0	0	0	6	-3	-1	0	-3	0	0	-3	-4	-3	-3	-2	-2	-1	1	3
G	0	0	0	0	0	0	6	-2	-4	-2	-4	-3	0	-2	-2	-2	0	-2	-3	-2	-3
H	0	0	0	0	0	0	0	8	-3	-1	-3	-2	1	-2	0	0	-1	-2	-3	-2	2
I	0	0	0	0	0	0	0	0	4	-3	2	1	-3	-3	-3	-3	-2	-1	3	-3	-1
K	0	0	0	0	0	0	0	0	0	5	-2	-1	0	-1	1	2	0	-1	-2	-3	-2
L	0	0	0	0	0	0	0	0	0	0	4	2	-3	-3	-2	-2	-2	-1	1	-2	-1
M	0	0	0	0	0	0	0	0	0	0	0	5	-2	-2	0	-1	-1	-1	1	-1	-1
N	0	0	0	0	0	0	0	0	0	0	0	0	6	-2	0	0	1	0	-3	-4	-2
P	0	0	0	0	0	0	0	0	0	0	0	0	0	7	-1	-2	-1	-1	-2	-4	-3
Q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	1	0	-1	-2	-2	-1
R	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	-1	-1	-3	-3	-2
S	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	1	-2	-3	-2
T	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	-2	-2
V	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	-3	-1
W	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11	2
Y	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7

Figure 1: Computed substitution matrix using BLOSUM62. Modified from [10] to exclude the non-canonical amino acid letters B X, and rearrangement of columns/rows to fit the order of the other substitution matrices.

Table 1: Scores for same-state amino acid transitions i.e from Alanine(A)→Alanine(A). These are found on the diagonal of the matrices

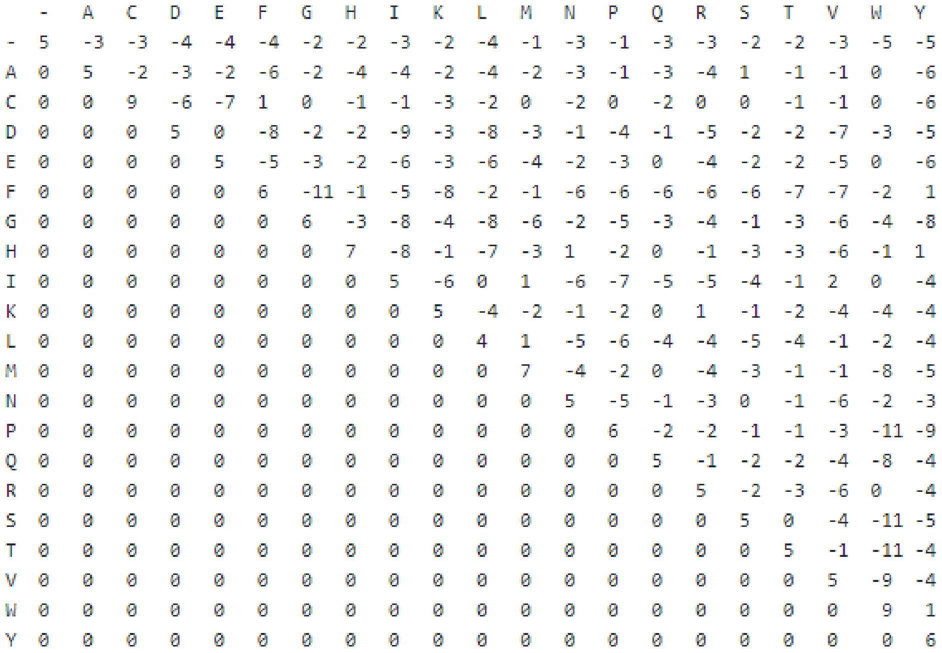
	-	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
Mesophiles	3	4	7	5	4	6	5	7	5	4	4	5	5	6	4	5	4	5	4	9	6
Psychrophiles	5	5	9	5	5	6	6	7	5	5	4	7	5	6	5	5	5	5	5	9	6
Thermophiles	3	4	8	4	3	5	5	6	4	3	3	5	5	5	5	4	4	5	4	7	5

The Figures 1,2,3 and 4 show the computed substitution matrices for the corresponding group of sequences. On an initial observation, we are already able to spot differences in the values when comparing each matrix to another. An initial glance over the different substitution matrices shows that values between each do differ. In particular, we observe a much higher proportion of larger negative numbers within the Psychrophiles when compared to the rest. Three out of the four occurrences of the lowest number (-11) are within the column of Tryptophan (W→S; W→T; W→P). Looking at the diagonals (i.e A→A no change in state) in Table 1, we do not see any significant changes, in most cases only changing by an order of 1, and in some cases 2.



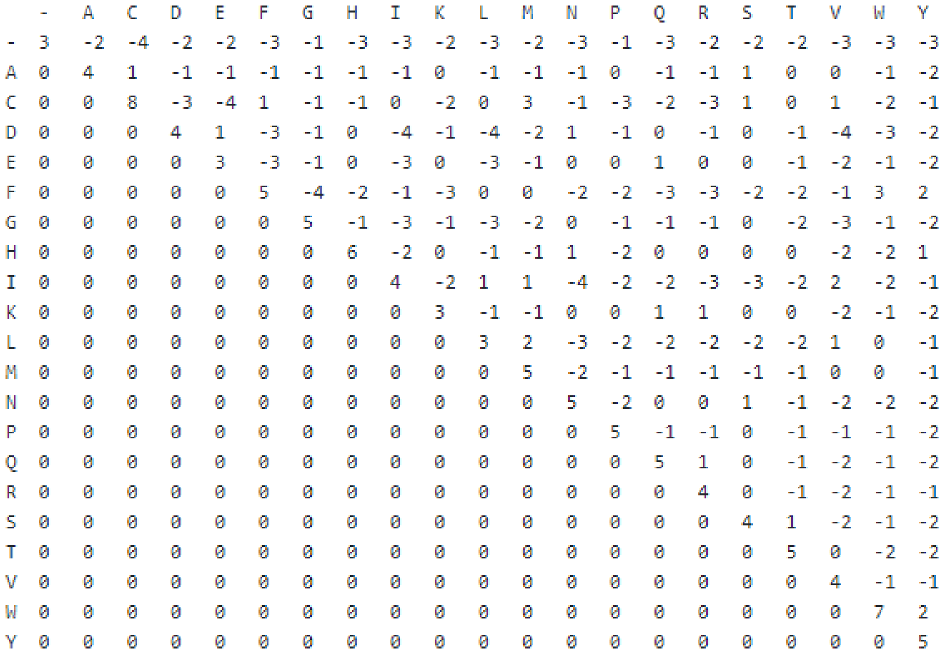
	-	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
-	3	-3	-4	-3	-3	-3	-2	-4	-4	-3	-4	-3	-3	-2	-3	-3	-2	-3	-3	-3	-4
A	0	4	1	0	-1	-2	0	-1	-1	0	-2	-1	-1	-1	0	-1	1	0	0	-4	-2
C	0	0	7	-3	-1	2	-1	-2	0	-2	2	1	-2	-4	-2	0	0	0	0	-3	0
D	0	0	0	5	1	-4	0	0	-4	-1	-3	-1	1	0	0	-2	0	-1	-3	-4	-2
E	0	0	0	0	4	-3	-1	0	-3	0	-3	0	0	-1	1	-1	-1	-1	-2	-2	-3
F	0	0	0	0	0	6	-4	-1	-1	-3	0	-1	-3	-3	-2	-2	-3	-2	-2	4	2
G	0	0	0	0	0	0	5	-1	-4	-1	-4	-2	0	-1	-2	-1	0	-2	-3	-3	-4
H	0	0	0	0	0	0	0	7	-3	0	0	-1	0	-4	0	0	0	-1	-1	-2	1
I	0	0	0	0	0	0	0	0	5	-3	1	1	-5	-2	-1	-3	-3	-1	2	-3	-1
K	0	0	0	0	0	0	0	0	0	4	-2	-1	0	0	1	1	0	0	-2	-2	-2
L	0	0	0	0	0	0	0	0	0	0	4	2	-3	-3	-1	-2	-3	-2	1	-1	-1
M	0	0	0	0	0	0	0	0	0	0	0	5	-2	-2	0	-1	-1	0	1	0	-2
N	0	0	0	0	0	0	0	0	0	0	0	0	5	-1	0	-1	1	-1	-2	-2	-2
P	0	0	0	0	0	0	0	0	0	0	0	0	0	6	-1	-1	0	-1	-1	-2	-3
Q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	-1	-1	0	-2
R	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	-1	-3	-2	-2
S	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	1	-1	-2	-3
T	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	-3	-1
V	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	-2	-1
W	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	1
Y	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6

Figure 2: Substitution matrix for Mesophiles.



	-	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
-	5	-3	-3	-4	-4	-4	-2	-2	-3	-2	-4	-1	-3	-1	-3	-3	-2	-2	-3	-5	-5
A	0	5	-2	-3	-2	-6	-2	-4	-4	-2	-4	-2	-3	-1	-3	-4	1	-1	-1	0	-6
C	0	0	9	-6	-7	1	0	-1	-1	-3	-2	0	-2	0	-2	0	0	-1	-1	0	-6
D	0	0	0	5	0	-8	-2	-2	-9	-3	-8	-3	-1	-4	-1	-5	-2	-2	-7	-3	-5
E	0	0	0	0	5	-5	-3	-2	-6	-3	-6	-4	-2	-3	0	-4	-2	-2	-5	0	-6
F	0	0	0	0	0	6	-11	-1	-5	-8	-2	-1	-6	-6	-6	-6	-6	-7	-7	-2	1
G	0	0	0	0	0	0	6	-3	-8	-4	-8	-6	-2	-5	-3	-4	-1	-3	-6	-4	-8
H	0	0	0	0	0	0	0	7	-8	-1	-7	-3	1	-2	0	-1	-3	-3	-6	-1	1
I	0	0	0	0	0	0	0	0	5	-6	0	1	-6	-7	-5	-5	-4	-1	2	0	-4
K	0	0	0	0	0	0	0	0	0	5	-4	-2	-1	-2	0	1	-1	-2	-4	-4	-4
L	0	0	0	0	0	0	0	0	0	0	4	1	-5	-6	-4	-4	-5	-4	-1	-2	-4
M	0	0	0	0	0	0	0	0	0	0	0	7	-4	-2	0	-4	-3	-1	-1	-8	-5
N	0	0	0	0	0	0	0	0	0	0	0	0	5	-5	-1	-3	0	-1	-6	-2	-3
P	0	0	0	0	0	0	0	0	0	0	0	0	0	6	-2	-2	-1	-1	-3	-11	-9
Q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	-1	-2	-2	-4	-8	-4
R	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	-2	-3	-6	0	-4
S	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	-4	-11	-5
T	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	-1	-11	-4
V	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	-9	-4
W	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	1
Y	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6

Figure 3: Substitution matrix for Psychrophiles.



	-	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
-	3	-2	-4	-2	-2	-3	-1	-3	-3	-2	-3	-2	-3	-1	-3	-2	-2	-2	-3	-3	-3
A	0	4	1	-1	-1	-1	-1	-1	-1	0	-1	-1	-1	0	-1	-1	1	0	0	-1	-2
C	0	0	8	-3	-4	1	-1	-1	0	-2	0	3	-1	-3	-2	-3	1	0	1	-2	-1
D	0	0	0	4	1	-3	-1	0	-4	-1	-4	-2	1	-1	0	-1	0	-1	-4	-3	-2
E	0	0	0	0	3	-3	-1	0	-3	0	-3	-1	0	0	1	0	0	-1	-2	-1	-2
F	0	0	0	0	0	5	-4	-2	-1	-3	0	0	-2	-2	-3	-3	-2	-2	-1	3	2
G	0	0	0	0	0	0	5	-1	-3	-1	-3	-2	0	-1	-1	-1	0	-2	-3	-1	-2
H	0	0	0	0	0	0	0	6	-2	0	-1	-1	1	-2	0	0	0	0	-2	-2	1
I	0	0	0	0	0	0	0	0	4	-2	1	1	-4	-2	-2	-3	-3	-2	2	-2	-1
K	0	0	0	0	0	0	0	0	0	3	-1	-1	0	0	1	1	0	0	-2	-1	-2
L	0	0	0	0	0	0	0	0	0	0	3	2	-3	-2	-2	-2	-2	-2	1	0	-1
M	0	0	0	0	0	0	0	0	0	0	0	5	-2	-1	-1	-1	-1	-1	0	0	-1
N	0	0	0	0	0	0	0	0	0	0	0	0	5	-2	0	0	1	-1	-2	-2	-2
P	0	0	0	0	0	0	0	0	0	0	0	0	0	5	-1	-1	0	-1	-1	-1	-2
Q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	1	0	-1	-2	-1	-2
R	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	-1	-2	-1	-1
S	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	1	-2	-1	-2
T	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	-2	-2
V	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	-1	-1
W	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	2
Y	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5

Figure 4: Substitution matrix for Thermophiles.

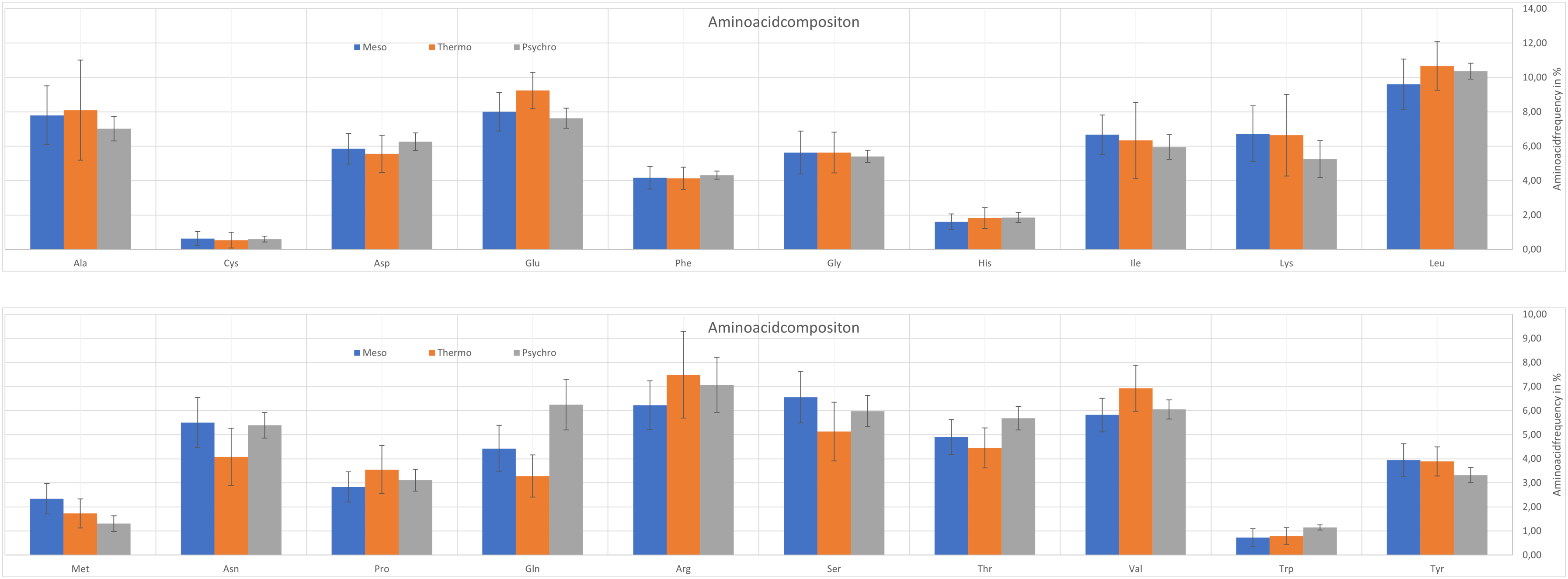


Figure 5: Amino acid composition of Mesophiles (blue), Psychrophiles (grey) and Thermophiles (orange).

Figure 5 shows the amino acid composition of mesophiles, thermophiles and psychrophiles. The bar chart includes the x-axis which shows the different amino acids and the y-axis their frequencies within the sequences.
It is clear to see that some bars have a higher variance than others. Ala(2.90), Ile(2.21) and Lys(2.37) in thermophiles are particularly noticeable when compared to the other two groups respectively. The bar chart also shows groups with very small to hardly any variance that can be seen for Cys(0.18) and Trp(0.11) as well as Phe(0.23) in psychophiles. Apart from these outliers, the distribution of amino acids is the same in all three groups with Cys being the rarest, while Leu is the most abundant. However, there is still the trend that those have fairly low variance.

Discussion

Previous publications showed differences in amino acid composition of proteins from thermophillic and psychrophillic organisms[11]. The following results are for the comparison between the thermophillic and the psychrophillic matrices. For all for amino acids with polar side chains(Ser,Thr,Asn and Gln) the composition changed to a higher content of polar side chains on lower optimum growth temperatures (OGTs). For the other amino acids, changes in composition can be seen, but are not statistically significant.

One notable result was the higher proportion of negative numbers for Tryptophan in the Psychrophile substitution matrix. To investigate this, literature searches regarding the relationship between psychrophiles (i.e lower temperatures) and the reduction in transitions to and from Trpytophan were carried out. We discovered there to be a lack of information regarding possible direct connections between Tryptophan and lower temperatures. Tryptophan is one of the more (if not most) expensive amino acids to synthesise due to its complex composition requiring multiple involved biosynthetic reactions [12]. It is also likely that this is why it is usually obtained through nutrition rather than synthesis. The expensive price to pay for its synthesis, and the likely lower amount of nutritional availability in cold regions, could be a factor in the lower occurence of tryptophan in psychrophiles.

Our project does come with it's limitations. BLOSUM62 has a filtering step which filters out sequences with a >62% similarity to each other, whereas in ours we conducted a more general filtering process of excluding any too similar or too distant sequences. By removing too dissimilar sequences, it could hinder the application of our substitution matrices on organisms that exist in the same temperature niche, but are very distantly related. BLOSUM also selects a small, highly-conserved, gapless regions from which it calculates the scores. In comparison, our example takes the entire length of the sequences into account. It differs from the BLOSUM62 matrix as gaps are also taken into consideration. Additionally, further investigation into which categories of amino acid (e.g polar, charged) most of the more drastic differences in occur in.

Further explorations could explore the use of these substitution matrices in other processes such as alignments and see their effect. Preliminary testing using ClustalW [13] showed no differences in clustering of the sequences after application of any of the substitution matrices.

Subsequent analysis could also be carried out within the stated temperature groups. Thermophiles, for example, can be further sub-classified into Simple-, extreme- or hyperthermophiles. Exploring whether inter-class or within-class differ more would be a next step forward.

As we try to implement a substitution matrix considering the optimal growth temperatures of the organisms, we mentioned the possibility to modify scoring formula (1) to account for variation in the difference to the optimal (or mean) temperature for each temperature group.
The closer the targeted microorganism is to the optimum temperature the lower the penalty and vice versa. Therefore the new formula could look something like this:

$$S_{ij} = \left(\frac{1}{\lambda}\right) log \left(\frac{p_{ij} \times (1 - \frac{|deviation\ from\ the\ temperature|}{average\ optimal\ temperature})}{q_i \times q_j}\right) \tag{2}$$

This formula suggests that the closer the organism is to its optimal temperature the closer the values are to the BLOSUM matrix on the other hand it shows how significant the temperature would affect a substitution matrix when taken into consideration.

Code avaiability

Our solution was implemented using Python. The code we used and the dataset we worked with can be found inside the Supplementary Material folder.

References

1. Arenas, M. Trends in substitution models of molecular evolution. *Frontiers in Genetics* **6** (Oct. 2015).
2. "Henikoff, S. & Henikoff, J. G. "Amino acid substitution matrices from protein blocks". "en". *"Proc Natl Acad Sci U S A"* **89,** "10915–10919" (Nov. 1992).
3. Berry, E. D. & Foegeding, P. M. Cold Temperature Adaptation and Growth of Microorganisms. *Journal of Food Protection* **60,** 1583–1594 (Dec. 1997).
4. "Kim, T. D., Ryu, H. J., Cho, H. I., Yang, C. H. & Kim, J. "Thermal behavior of proteins: heat-resistant proteins and their heat-induced secondary structural changes". "en". *"Biochemistry"* **39,** "14839–14846" (Dec. 2000).
5. Chu, X.-L., Zhang, B.-W., Zhang, Q.-G., Zhu, B.-R., Lin, K. & Zhang, D.-Y. Temperature responses of mutation rate and mutational spectrum in an Escherichia coli strain and the correlation with metabolic rate. *BMC Evolutionary Biology* **18** (Aug. 2018).
6. Sousounis, K., Haney, C. E., Cao, J., Sunchu, B. & Tsonis, P. A. Conservation of the three-dimensional structure in non-homologous or unrelated proteins. *Human Genomics* **6** (Aug. 2012).
7. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences* **89,** 10915–10919 (Nov. 1992).
8. *Melnikov GSHT: a database of organisms with defined optimal growth temperatures* http://melnikovlab.com/gshc/. Accessed: 12-01-2023.
9. Tamura, K., Stecher, G. & Kumar, S. MEGA11: Molecular Evolutionary Genetics Analysis Version 11. *Molecular Biology and Evolution* **38** (ed Battistuzzi, F. U.) 3022–3027 (Apr. 2021).
10. https://ftp.ncbi.nih.gov/blast/matrices/BLOSUM62 (2023).
11. Yang, L.-L., Tang, S.-K., Huang, Y. & Zhi, X.-Y. Low Temperature Adaptation Is Not the Opposite Process of High Temperature Adaptation in Terms of Changes in Amino Acid Composition. *Genome Biology and Evolution* **7,** 3426–3433 (Nov. 2015).
12. Barik, S. *The Uniqueness of Tryptophan in Biology: Properties, Metabolism, Interactions and Localization in Proteins* Nov. 2020.
13. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* **22,** 4673–4680 (1994).