# The Diversity and Evolution of Micro-organisms in Hydrothermal Vents

## Exploring the role of pyrophosphates in the acquisition of energy

University of
St Andrews

**Carlota Bloch Varela**
cbv1@st-andrews.ac.uk

Dept. of Earth and Environmental Sciences
Supervisors: Dr Joanne Boden and Dr Eva Stueeken
August 2023

LAIDLAW
FOUNDATION

# The Diversity and Evolution of Micro-organisms in Hydrothermal Vents

**Carlota Bloch Varela**[1]
1. Dept. of Earth and Environmental Sciences, University of St Andrews

## 1 Introduction

### 1.1 General Overview

Deep-sea hydrothermal vent environments, first discovered in 1977 [1], are characterised by hosting chemosynthetic microbes and being unique ecological niches, with harsh biological parameters, that contain nutrients released from geothermal waters warmed by the volcanic activity present under the seafloor [2]. The emergence of diffuse fluids from the seafloor resulting from high-temperature hydrothermal fluids mixed with cold deep seawater produces a gradation in pH, temperature, and energy sources, generating, depending on said local geological and geochemical conditions present, different habitats and microbial communities [2].

In this research project, the main aim has been to identify the mechanisms these microorganisms use in order to survive and gain energy in such extreme environment by interrogating their metagenomes and DNA architecture. Particularly, we were interested in finding whether they use pyrophosphates instead of ATP, the usual biological fuel, in addition to how and when those genes evolved.

### 1.2 Pyrophospahtes as an energy source

Phosphorous, one of the most abundant chemicals when it comes to living cells, is taken in its dissolved form (Pi) by microorganisms and plants. The reaction of one phosphate anion with another - a dimer - can lead to the production of a pyrophosphate (PPi). The subsequent addition of Pi residues to PPi can lead to the production of polyphosphate (polyP), defined, therefore, as a polymer consisting of Pi residues linked by phosphoanhydride bonds [3][4].

This compound, PPi, during the 1960's, started to be considered a primitive source of energy in early stages of life [5] and an alternative to adenosine triphosphate (ATP) when it was found that photosynthetic bacterial membrane-bound inorganic pyrophosphatase (PPase) was the main pathway to catalyse light-induced phosphorylation of Pi into PPi, which has the capacity to drive energy that entails dark reactions [6][7].

The required membrane-bound PPase, moreover, is also a $Na^+$ and $H^+$ -pump, outlining how PPi in conjunction with the transport of $Na^+$ and $H^+$ are present in the initial evolution of life and the explained mechanism of acquiring energy [6]. Therefore, the abundance of $Na^+$ in contemporary hydrothermal vents, mimicking the Archean Earth bioecological conditions, introduced the question of whether microorganisms found in such environments also employ PPi as their energy source.

Gleaned from everything outlined, we focused on unravelling whether the microorganisms in hydrothermal vents possess the structural genes in their metagenomes for them to have PPi metabolising enzymes, contained in Table 1 [4].

## 2 Methodology: general overview

This research was entirely done with bioinformatics and computational softwares to be able to efficiently process and manipulate the genetic sequences.

Particularly, the project was divided into two stages, with Ubuntu as the main computational platform (Figure 1). In Stage 1, the primary focus was to find whether microorganisms in different hydrothermal vents' communities use pyrophosphates by undertaking HMM and BLAST searches and generating evolutionary trees. Alternatively, the Stage 2 focused on how and when those genes coding for metabolising enzymes evolved through the elaboration of reconciliation trees.

| Polyphosphates | | | | |
|---|---|---|---|---|
| Bacteria | Gene | Enzyme | PID | PN |
| Escherichia coli K12 | ppk1 | Polyphosphate kinase 1 (PPK1) | PF02503 | PP_kinase |
| Pseudomonas aeruginosa PAO1 | ppk2 | Polyphosphate kinase 2 (PPK2) | PF03976 | PPK2 |
| Escherichia coli K12 | surE | SurE | PF01975 | SurE |
| Escherichia coli K12 | gppA | Exopolyphosphatase (PPX) | PF02541 | Ppx-GppA |
| Mycobacterium tuberculosis H37Rv | ppnK | NAD kinase | PF01513 | NAD_kinase |
| Mycobacterium tuberculosis H37Rv | ppgK | Polyphosphate glucokinase | PF00480 | ROK |
| Dictyostelium discoideum | arpABCEFGH | Actin-related protein | PF00022 | Actin |

**Table 1:** Table of the bacterial polyphosphate metabolising enzymes selected to perform the HMM and BLAST searches over the microorganisms' metagenomes. The table includes the bacteria sequenced, the gene, the enzyme produced, and both the Pfam ID (PID) and name (PN) to search for multiple sequence alignments and HMM profiles in the Pfam database [4].

## 3 Stage 1: Use of pyrophosphates

### 3.1 HMM profile and BLAST: Search for homologies

Protein domains designate the basic functional, structural, and evolutionary units in proteins that can fold and evolve independently with respect to the rest of the protein's structure [8][9]. Therefore, protein families group those proteins that have a shared domain, i.e. evolutionary origin reflected in their related functions and similar sequences or structures.

According to the objectives of this research, one of our main aims was to find whether microorganisms in hydrothermal vents employ pyrophosphates as their energy source. To do so, the first step was to find, through either an HMM search or performing a BLAST, homologous sequences among the metagenomes of microorganisms in hydrothermal vents regarding the query sequences of pyrophosphates of interest by aligning them.

Firstly, HMMsearch - Hidden Markov Model - is a tool with a statistical basis that considers the potential alignments according to their relative likelihood [10] and is employed to detect protein homologies with the use of an HMM profile, a probabilistic position-specific evolutionary scoring model accounting for the likeliness of observing specific symbols and the frequency of finding indels in each position of a multiple sequence alignment; consequently, HMM technology is used to compare said HMM profiles to single or multiple sequences. For it, the following line of code was employed, were X stands for the name of the pyrophosphate, the Profile.HMM document contains the HMM profile, and the H.faa document the metagenomes of interest:

```
hmmsearch -o X_hmm_hits.out -E
0.1 --tblout X_hmm_hits.tab --
domtblout X_hmm_hits.domtab -A
Alignment.out --noali --notextw
Profile.HMM H.faa
```

HMM search was conducted with metagenomes from the Mid-Cayman Rise, located in the Caribbean Sea, a region encompassing two divergent hydrothermal vents: the mafic Piccard vent field and the ultramafic Von Damm vent field. The former, the deepest hydrothermal vent site discovered (4960 m of depth), is characterised by it having high temperatures and containing several metals such as sulphide. The latter, located at the top of a massif at a depth of 2350 m, is characterised by cooler temperatures and being rich in hydrocarbons such as hydrogen or methane [1]. The allegedly phylogenetic distinct microorganisms in these two sites, despite the divergent biogeological conditions, present a metabolic functional equiva-

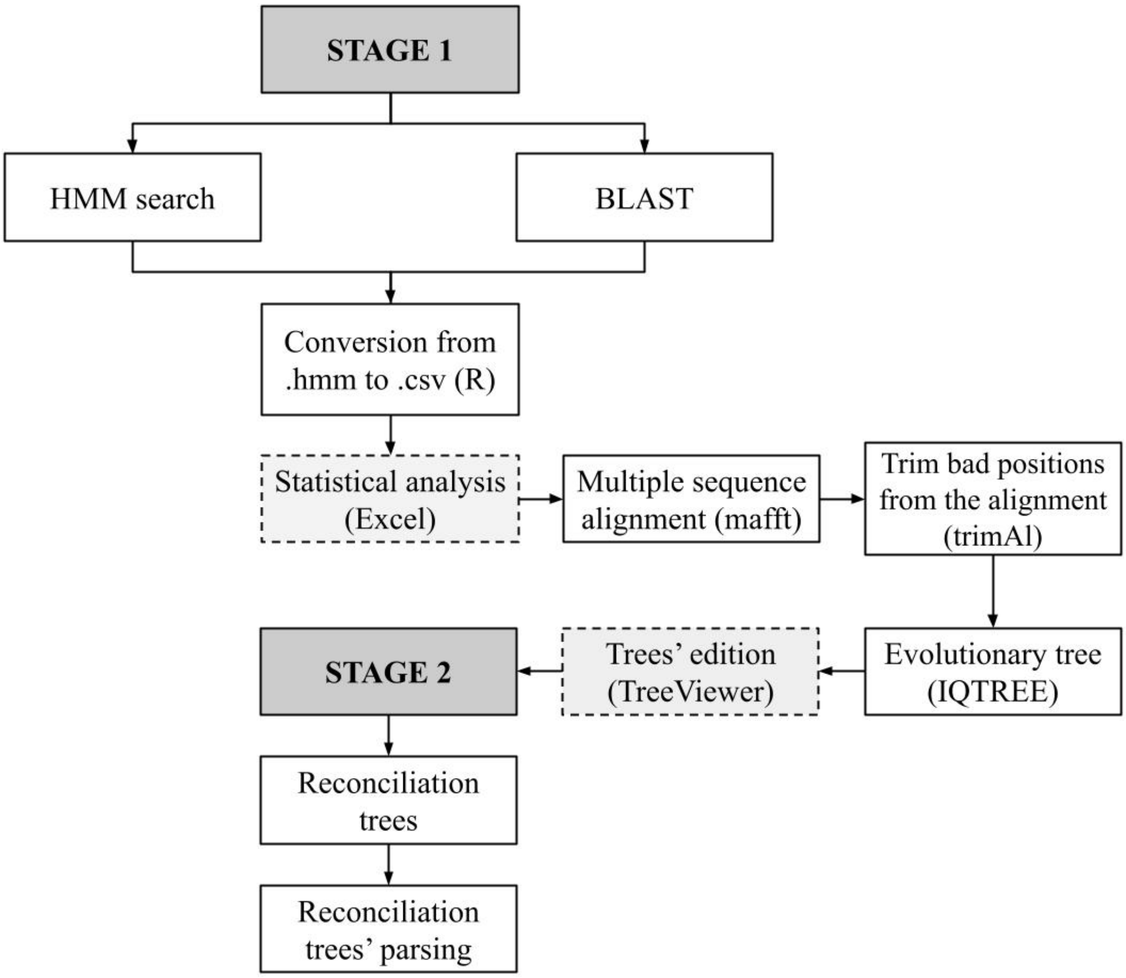

**Figure 1:** Methodology employed divided in two stages. Stage 1: Finding, through homologous search, multiple sequence alignment, and evolutionary trees, if microorganisms in hydrothermal vents employ PPi as their source of energy. Stage 2: Elaboration of reconciliation trees. The boxes with discontinuous lines indicate steps in the procedure that entail the manipulation of the already obtained results to analyse them.

lence, something that we interrogated in this research focusing on the metabolisation of polyP (Table 2).

Secondly, whereas HMMsearch uses site-specific probabilistic models, BLAST - Basic Local Alignment Search Tool - was also used in this project, an alternative algorithmic tool that uses site-independent substitution score matrices to find homologies among the input sequences [10]. For it, the following line of code was employed, were C.faa contains the pyrophosphate query sequences for the search, D.csv is the excel output of the search, and the H.faa document contains the metagenomes of interest:

```
tblastn -query C.faa -out D.csv
-evalue 1e-25 -subject H.faa
-outfmt "6 qseqid sseqid sstart
send evalue bitscore pident length
mismatch gapopen qstart qend sseq"
```

A BLAST search has been conducted over metagenomes collected in hydrothermal vents at Axial Seamount, one of the most active volcanoes in the NE Pacific ocean located on the Juan de Fuca Ridge [2]. These genomes (Table 3) were interrogated using BLAST because its an appropriate method when the sequences are already assembled into contigs, defined as a set of overlapping DNA sequences or segments that conform a contiguous genomic region [11].

The results from both HMM and BLAST searches were converted into a .csv file to statistically analyse the hits obtained and annotate which

| Mid-Cayman Rise | | | | |
|---|---|---|---|---|
| Sample name | Sample number | Vent field | Tmax (°C) | Year |
| Old Man Tree | FS841 | Von Damm | 114 | 2012 |
| Ravelin #2 | FS842 | Von Damm | 86 | 2012 |
| Shrimp Hole | FS844 | Von Damm | 50 | 2012 |
| Ginger Castle | FS848 | Von Damm | 47 | 2012 |
| Main Orifice | FS849 | Von Damm | 109 | 2012 |
| Hot Chimlet, BVM | FS851 | Piccard | 106 | 2012 |
| Shrimp Canyon, BVM | FS852 | Piccard | 44 | 2012 |
| X-19 at BV #4, BVM | FS854 | Piccard | 18 | 2012 |
| Shrimp Gulley #2, BSM | FS856 | Piccard | 108 | 2012 |
| Near Main Orifice | FS866 | Von Damm | 130 | 2013 |
| Shrimp Hole | FS872 | Von Damm | 30 | 2013 |
| Twin Peaks | FS874 | Von Damm | 140 | 2013 |
| Shrimp Buttery | FS877 | Von Damm | 131 | 2013 |
| Hot Cracks #2 | FS879 | Von Damm | 29 | 2013 |
| Old Man Tree | FS881 | Von Damm | 114 | 2013 |

**Table 2:** Data of the samples taken from the Mid-Cayman Rise hydrothermal vents used for the HMM search of pyrophosphates among their metagenomes. The table includes the samples' names, identification number, from which vent they were taken and its temperature, and the sampling year [1].

ones were higher than the cut-offs, defined as a statistical value that demarcates the hits considered significant.

### 3.2 Relatedness between hits: Multiple sequence alignment (mafft), trimming (trimAl, and Evolutionary tree

As the National Centre of Biotechnological Information does to identify the proteins stored in GenBank and RefSeq, the hits above the cut-offs obtained from the searches could be used exclusively as already conclusive, but in this research we decided to check our results by only extracting those sequences closely related to the query sequences by producing an evolutionary tree.

To do so, the query sequences and the metagenomes sequences were aligned using the following mafft command, were HitsQueries.faa is the assembled document with the search hits and the query sequences:

```
mafft --mpi --large --genafpair --
thread 5 HitsQueries.faa >>
HitsQueries.genprog.faa
```

After trimming the poor sections with gaps in 85% or more in the sequences, the evolutionary tree was built with the following IQTREE command, were HitsQueries.trim85.faa document contains the trimmed and assembled sequences:

```
iqtree -s HitsQueries.trim85.faa -
bb 1000 -alrt 1000 -pre
<ProteinName> -ntmax 5 -nt AUTO
-m MFP -mrate E,I,G,I+G,R -madd
C10,C20,C30,C40,C50,C60,EX2,EX3,
EHO,UL2,UL3,EX_EHO,LG4M,LG4X,CF4,
LG+C10,LG+C20,LG+C30,LG+C40,
LG+C50,LG+C60 -nm 5000
```

The final stage was to take the previous step's output and, by highlighting in the tree the query sequences, detect the metagenomes more closely related to them using TreeViewer platform.

### 3.3 Results of the Stage 1

#### 3.3.1 HMM and BLAST searches: statistical analysis

For this first stage, one of the two main outputs was the statistical analysis of the significant hits

| Juan de Fuca Ridge | | |
|---|---|---|
| Sample Port | Tmax (°C) | Year |
| Port 1 | 52 | 2020 |
| Port 2 | 28 | 2020 |
| Port 5 | 48 | 2020 |
| Port 6 | 38 | 2021 |
| Port 8 | 34 | 2021 |
| Port 9 | 23 | 2021 |
| Port 10 | 20 | 2021 |
| Port 11 | 12 | 2021 |
| Port 12 | 11 | 2021 |
| Port 13 | 10 | 2021 |
| Port 14 | 14 | 2021 |
| Port 15 | 11 | 2021 |
| Port 17 | 10 | 2021 |
| Port 20 | 12 | 2021 |
| Port 21 | 7 | 2021 |
| Port 22 | 8 | 2021 |

**Table 3:** Data of the samples taken from the Juan de Fuca Ridge hydrothermal vents used for the BLAST search of pyrophosphates among their metagenomes. The table includes the samples' port, its temperature, and the sampling year. Note: Temperatures in red indicate temperatures that have not been validated [2].

obtained in the HMM and BLAST searches.

The chart plotted in Figure 2 summarizes the number of hits obtained after the HMM search. The first noticeable element from the chart is that in all of the sites from which samples were taken there have been hits between the metagenomes collected and the HMM profile with the exception of the site FS841. This consequently indicates that from almost every sample drawn from multiple sites in the Mid-Cayman Rise there are microorganisms employing enzymes to metabolise pyrophosphates for the acquisition of energy.

But beyond the number of hits per site per pyrophosphate, a negative correlation between hits and temperature was qualitatively found when looking at the chart, as with a higher temperature there were less hits in HMM search, and with lower temperature, more hits. This qualitative observation was quantitatively interrogated by performing a statistical correlation test with JASP; based on the performance of a prior test, the Shapiro-Wilk test, Pearson's test for normally distributed data had to be employed. The results showed a statistically significant negative correlation between the total number of hits per site and its maximum temperature, $r(13) = -0.669, p = .006$.

With these statistical results we could have concluded that microorganisms in hydrothermal vents with higher temperatures employ less enzymes that metabolise pyrophosphates to acquire energy compared to the microorganisms in hydrothermal vents with lower temperatures, where microorganisms tend to use more pyrophosphates metabolising enzymes. Nevertheless, before concluding this, we decided to test the correlation between non-pyrophosphate genes against the temperature in each site to see if the correlation found is exclusively present with the pyrophosphates tested or a general trend.

To do so, a common gene, the ribosomal protein S8, was used to perform an additional HMM search among the Mid-Cayman Rise metagenomes. With the statistically significant hits and the temperatures per site a Pearson's r correlation for normally distributed data was per-

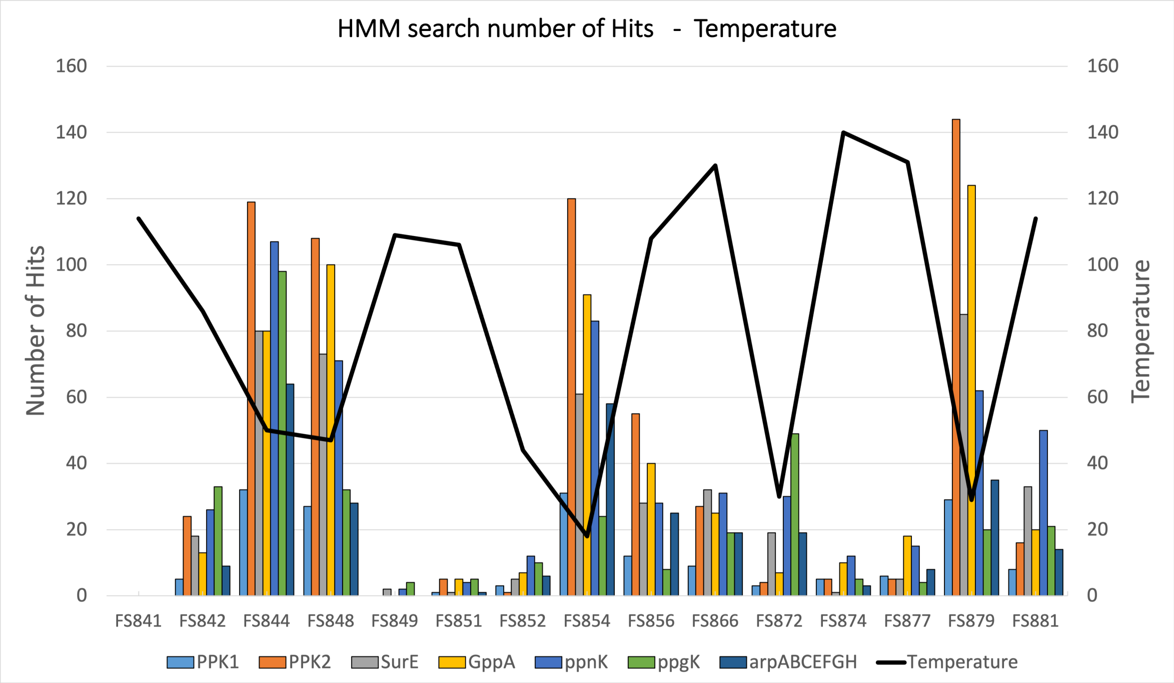

**Figure 2:** Bar graph representing the amount of hits per polyphosphate after the HMM search for each of the Mid-Cayman Rise samples, and the temperature of each site represented by a line.

formed, and the results likewise showed that there was a statistically significant negative correlation between the total number of hits per site and its maximum temperature, $r(13) = -0.573, p = .025$. Consequently, it was found that the previously stated correlation with pyrophosphates' metabolising enzymes is not exclusively found with them and its rather a general trend among genomes in this region. A potential explanation for this could be that it is because of a lower survival rate in higher temperatures and consequently there are less microorganisms in the samples.

Another statistical analysis was also performed with the statistically significant hits obtained from the BLAST search, summarised in Table 4; a chart was not plotted because of the huge difference among hits between pyrophosphates. The first noticeable element, like in the HMM search results, is how in all of the ports used for the BLAST search have been hits for each of the pyrophosphates selected. Therefore, the microorganisms found in hydrothermal vents at Axial Seamount do possess genes that codify for enzymes that metabolise those Pi compounds for the acquisition of energy.

Moreover, from the total number of hits presented in Table 4, it is also noticeable how SurE is way more abundant among all sampled microorganisms compared to the other pyrophosphates and how ppnK is the least abundant one. This difference could be due to the specific biochemical and geological parameters of the hydrothermal vents at Axial Seamount, the extent to which was not analysed in this project but that could be interrogated in future research.

Furthermore, after having questioned the correlation between HMM searches' hits and temperature, it was also interrogated if a similar pattern could be found with the hits in BLAST searches. To do so, since the Shapiro-Wilk test was non-significant, a Pearson's r correlation test was conducted, leading to a non-significant result: $r(14) = -0.414, p = .111$. This could be so because the range of temperatures in this site is small, meaning that the difference between the max (52°C) and min (7°C) temperatures is not as big as the one found in Mid-Cayman Rise hydrothermal vents (max = 140, min = 18). Consequently, since the temperature range is small, there is not a significant effect on the number of hits obtained from the BLAST search.

### 3.3.2 Evolutionary trees

Regarding the latter outcome in this first stage, evolutionary trees for each search had to be produced. Since not all trees finalised running due to time constraints, two HMM-based trees - PpgK

| Sample | ppk1 | ppk2 | SurE | gppA | ppnK | ppgK | arpABCEFGH |
|---|---|---|---|---|---|---|---|
| Port 1 | 5611 | 618 | 25012 | 88 | 8 | 39 | 0 |
| Port 2 | 9232 | 1150 | 66615 | 197 | 40 | 215 | 0 |
| Port 5 | 6950 | 1079 | 55579 | 184 | 54 | 185 | 8 |
| Port 6 | 5546 | 1003 | 44206 | 137 | 53 | 136 | 0 |
| Port 8 | 9173 | 1161 | 56074 | 194 | 51 | 191 | 70 |
| Port 9 | 6355 | 887 | 47493 | 141 | 49 | 172 | 0 |
| Port 10 | 10633 | 1492 | 64678 | 186 | 36 | 170 | 0 |
| Port 11 | 5322 | 861 | 42754 | 122 | 23 | 97 | 12 |
| Port 12 | 6638 | 1029 | 45010 | 132 | 24 | 90 | 53 |
| Port 13 | 7523 | 1030 | 46421 | 127 | 13 | 115 | 71 |
| Port 14 | 6950 | 1027 | 46826 | 129 | 14 | 119 | 77 |
| Port 15 | 9682 | 1586 | 70299 | 189 | 31 | 173 | 70 |
| Port 17 | 8576 | 1216 | 61321 | 172 | 20 | 148 | 70 |
| Port 20 | 7990 | 1337 | 57574 | 152 | 20 | 142 | 30 |
| Port 21 | 9293 | 1469 | 72482 | 178 | 30 | 242 | 82 |
| Port 22 | 7981 | 1205 | 54004 | 147 | 20 | 152 | 90 |
| Total | 123455 | 18150 | 856348 | 2475 | 486 | 2386 | 633 |

**Table 4:** Number of hits per polyphosphate and in total after the BLAST search for each Juan de la Fuca Ridge's sample.

and PpnK - will be included in this essay as an example of how all the evolutionary trees were analysed to select, out of all the hits, the ones considered true hits or actual orthologues from a phylogenetic viewpoint.

Two pieces of information included in the trees are the high support scores and cut-offs values, relevant for determining the veracity of the hits. Regarding the former, support values over 95% - represented with white circles - indicate a high confidence in the tree branches actually representing a true phylogenetic signal in the data. If we look at both trees, most of the branches do have a high support value, reinforcing that the evolutionary tree does indeed represent a faithful evolutionary relationship among the hit and query sequences, and, therefore, those microorganisms probably have pyrophosphates' metabolising enzymes. Concerning the latter, sequences scores above the cut-offs - represented with blue circles - demarcate those sequences that reach the minimum HMM search score to be considered a true hit. Based on this, by looking at the trees, most of the branches do pass the minimum cut-off score, indicating how those metagenomes are homologous sequences to the query ones, and, consequently, they have potentially incorporated in their genome genes that codify for pyrophosphates' metabolising enzymes.

Having established the veracity of the trees' evolutionary signal, the main outcome that can be noticed by looking at both trees is how the query sequences, represented by a yellow star, are spread around the tree and not condensed in one area of it, entailing that the rest of the sequences, i.e. the HMM search hits, are closely related to those query sequences. Nevertheless, some of the branches are far from the query sequences, meaning that the hits present in those branches are not that closely related to the pyrophosphates' sequences or that are distant relatives, putting into question the actual veracity of the HMM search hits and entailing the possibility of those metagenomes not including genes for metabolising the pyrophosphate in question. Consequently, those branches were not included as actual orthologues despite them being considered hits after the HMM search, showing the importance of running evolutionary trees to check the search results.

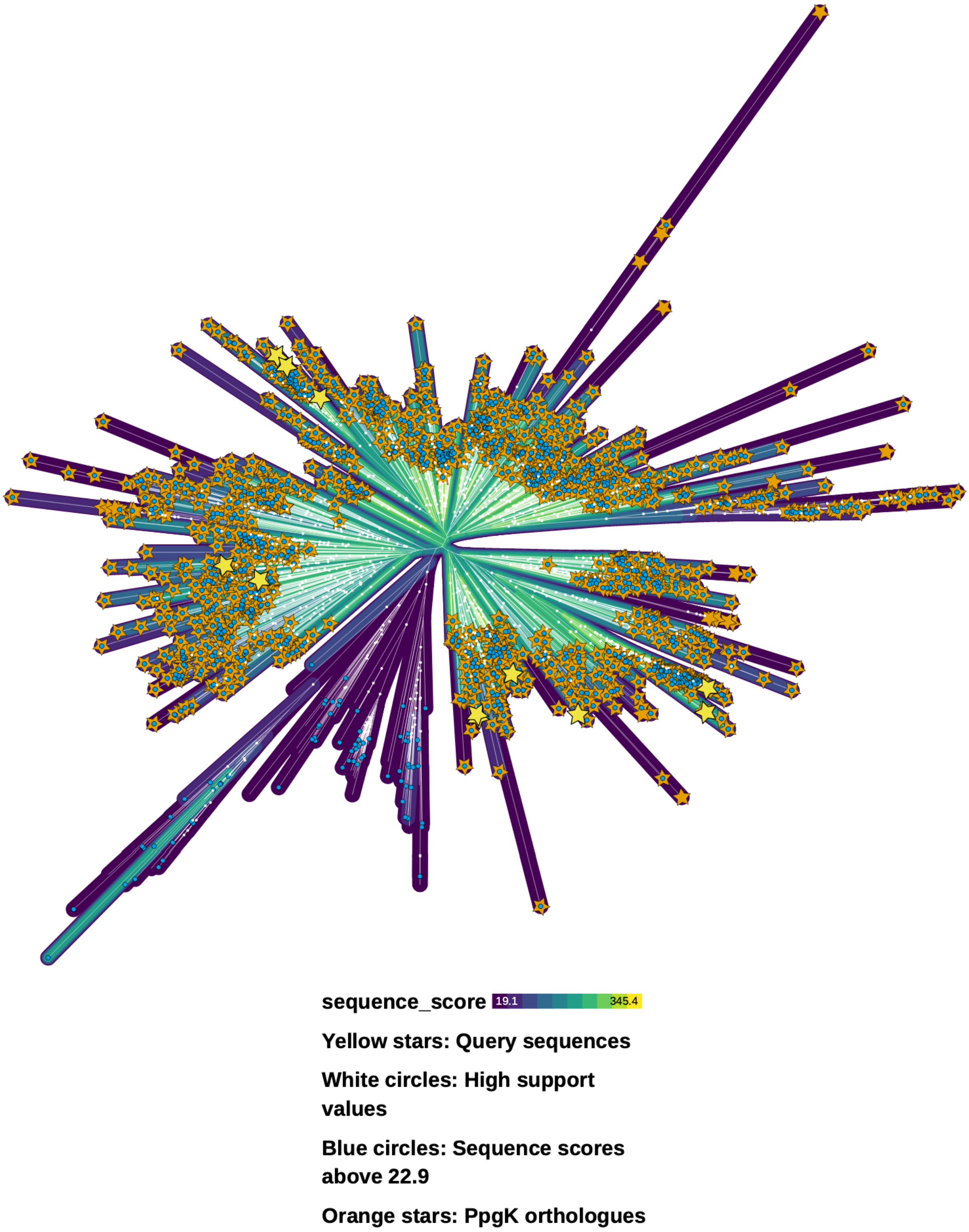

**Figure 3:** Evolutionary tree elaborated based on the phylogenetic signals among the hydrothermal vents' metagenomes and the pyrophosphate PpgK query sequences (yellow stars) to validate the hits obtained after the HMM search.

Gleaned from everything mentioned, the evolutionary trees helped verify the HMM and BLAST search outcomes and identify the true hits between the hydrothermal vents' metagenomes and the query sequences.

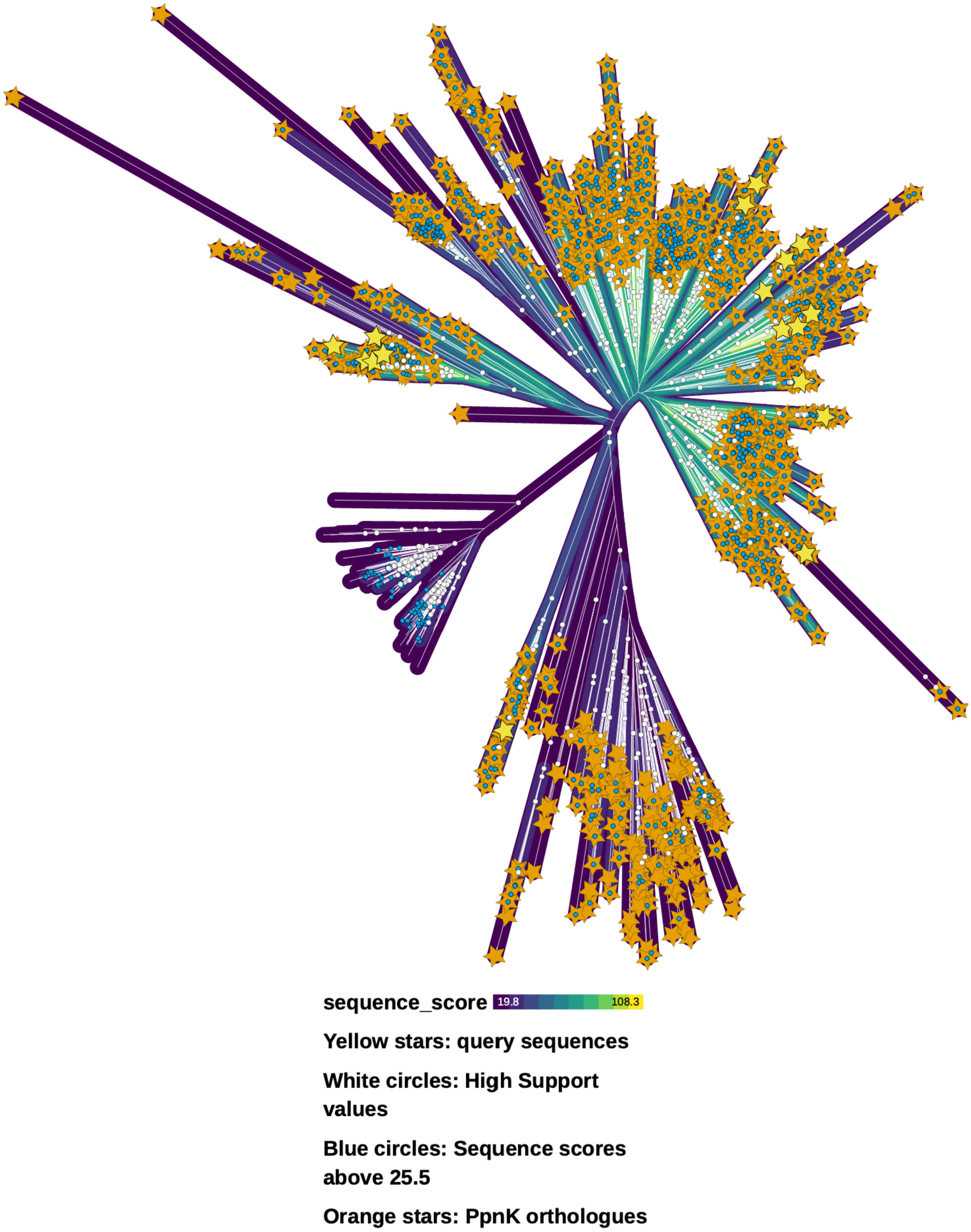

**Figure 4:** Evolutionary tree elaborated based on the phylogenetic signals among the hydrothermal vents' metagenomes and the pyrophosphate Ppnk query sequences (yellow stars) to validate the hits obtained after the HMM search.

# 4 Stage 2

## 4.1 Origin of genes: Reconciliation tree and its parsing

A reconciliation tree is a phylogenetic approach that aims at combining two or more co-evolving biological events in order to unravel their interdependence and how they relate from an evolutionary viewpoint [12]. For this research, the evolu-

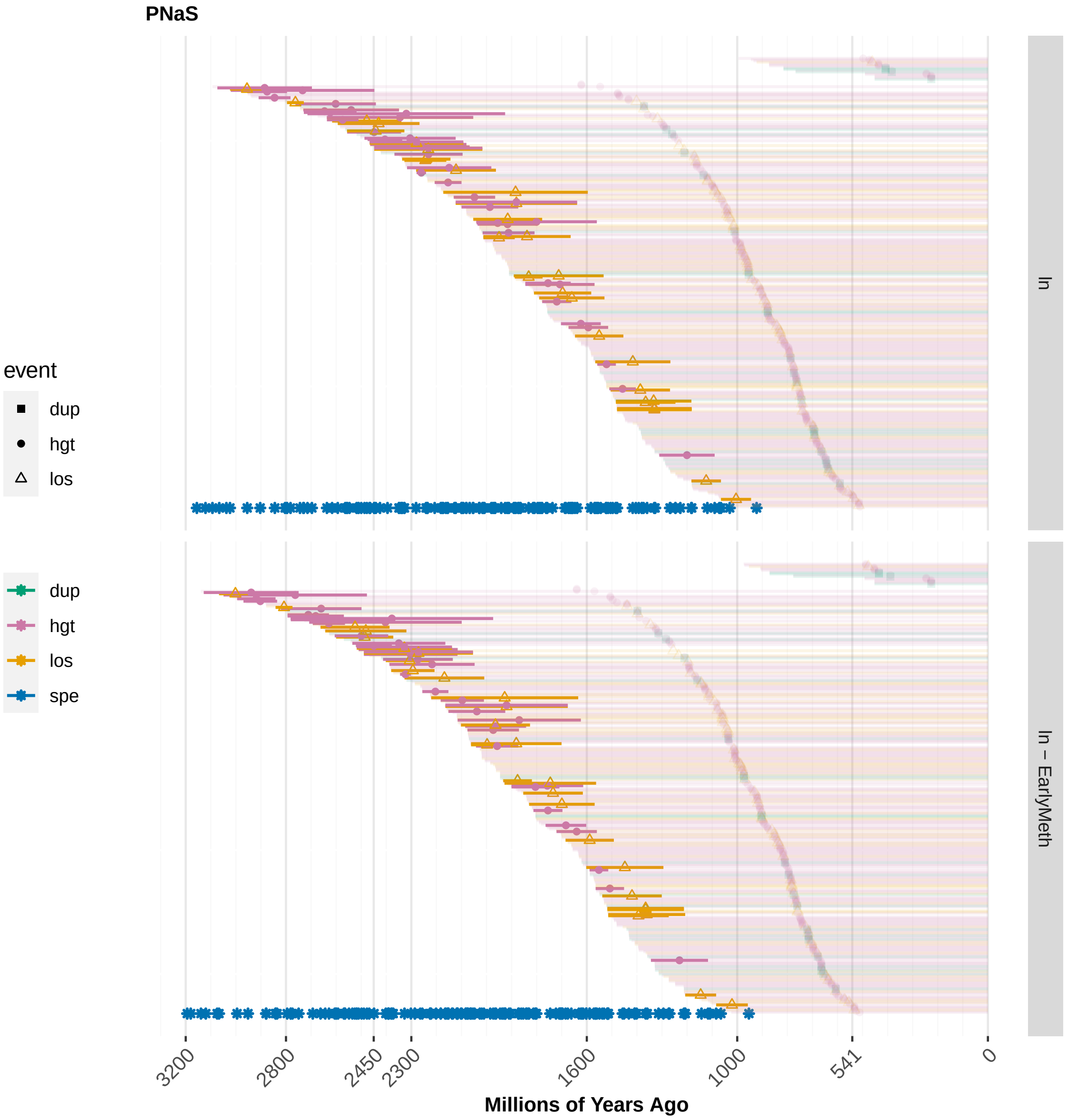

**Figure 5:** Dot plot comparing the reconciliation trees' results from both molecular clocks (In, In-EarlyMeth) regarding how the PNaS gene evolved by including the events of duplication, horizontal transference, loss, and speciation.

tionary trees run in the Stage 1 were combined with two tree-based molecular clocks, defined as the rates of genetic change across the tree of life with an absolute timescale [13], to see when a particular gene evolved, when it was horizontally transferred from one organism to another or vertically inherited, when it was lost for certain lineages, when it was duplicated, and see if its origin is close to the origin of life.

The problem that was faced during this research project was time constraints, as certain trees took more time than expected. Consequently, it was not possible to finish all of them or have them running. Therefore, I was given an already made Bayesian Gene Tree from different phosphate porters - PNaS - and a molecular clock that assumed that microorganisms (particularly methanogens) evolved more than 3.46 billions of years (Ga) to compare it with another molecular clock that assumed microorganisms evolving more than 2.7 Ga.

To compute the reconciliation trees and produce a dot representation of the genes' evolution, the following line of code (ecceTERA), in conjunction with an R script, was employed, where C.chronogram stands for the molecular clock and X.ufboot for the phosphorous cycling genes' Bayesian tree:

```
ecceTERA species.file= C.chronogram
dated=2 gene.file=X.ufboot
amalgamate=true verbose=true
print.reconciliations=1
recPhyloXML.reconciliation=true
char.sep=_ >> X.ML.EarlyMeth.log
```

### 4.2 Results of the Stage 2

For this stage and, as stated, due to the time constraints, a reconciliation dot plot from the phosphate porters PNaS will be included and analysed to see when the coding gene evolved and started to be used by microbes.

The dot plot represented in Figure 5 is comparing the reconciliation trees' output between the two molecular clocks previously discussed with different evolutionary assumptions. Nevertheless, as the dot plot shows, despite the difference in parameters and calibrations among said molecular plots, evidence coming from them suggest that the microbes started employing PNaS to collect phosphate around 3.200 million years ago, near the origin of life [14].

These results suggest, therefore, that microorganisms in hydrothermal vents employ metabolising enzymes that emerged in early stages of life, highlighting the use of this primitive source of energy also employed by microbes that existed, as stated, close to the origin of life.

## 5 General Conclusions

In this research, it has been interrogated the employment of PPi by microorganisms in hydrothermal vents by searching genes coding for metabolising enzymes in their metagenomes. Based on the statistical results obtained from the HMM and BLAST searches, considering the time constraints experienced, most of the microorganisms in the Mid-Cayman Rise and Juand de Fuca Ridge employ pyrophosphates as their source of energy instead of ATP, having those results been verified by the trees compiled and the evolutionary relationship found between those microorganisms' metagenomes and the query sequences.

On top of that, through the elaboration of two reconciliation trees, it was observed how the microorganisms sampled from hydrothermal vents employ as their source of energy PPi due to having encoded genes for metabolising enzymes that emerged in early stages of life and were employed by microbes that existed close to the origin of life.

Gleaned from everything outlined, it can be stated that this research has positively contributed to evolutionary studies regarding what fuels and delivers energy to microorganisms by adding comprehensive information concerning metabolic development, functionality, and metagenomic studies. Moreover, it has also contributed to understanding and unravelling evolutionary patterns due to the possession of said metabolising genes by microorganisms in early evolutionary stages of life, delivering a more general overview of broader ecological mechanisms and energy acquisition.

*I would like to thank my supervisors for their unconditional support throughout this project, and Lord Laidlaw and the Laidlaw Foundation for the opportunity to conduct this research*

## References

[1] D. Galambos, R. E. Anderson, J. Reveillaud, and J. A. Huber, "Genome-resolved metagenomics and metatranscriptomics reveal niche differentiation in functionally redundant microbial communities at deep-sea hydrothermal vents," *Environmental Microbiology*, vol. 21, no. 11, pp. 4395–4410, 2019. DOI: `10.1111/1462-2920.14806`.

[2] A. Opatkiewicz, D. Butterfield, and J. Baross, "Individual hydrothermal vents at axial seamount harbor distinct subseafloor microbial communities. fems microbiol ecol 70: 413-424," *FEMS microbiology ecology*, vol. 70, pp. 413–24, Jul. 2009. DOI: `10.1111/j.1574-6941.2009.00747.x`.

[3] J. Pérez-Castiñeira, R. Docampo, T. Ezawa, and A. Serrano, "Editorial: Pyrophosphates and polyphosphates in plants and microorganisms," *Frontiers in Plant Science*, vol. 12, Mar. 2021. DOI: `10.3389/fpls.2021.653416`.

[4] L. Wang, J. Yan, M. J. Wise, *et al.*, "Distribution patterns of polyphosphate metabolism pathway and its relationships with bacterial durability and virulence," *Frontiers in microbiology*, vol. 9, Apr. 2018. DOI: `10.3389/fmicb.2018.00782`.

[5] A. Baykov, A. Malinen, H. Luoto, and R. Lahti, “Pyrophosphate-fueled na+ and h+ transport in prokaryotes,” *Microbiology and molecular biology reviews : MMBR*, vol. 77, pp. 267–276, Jun. 2013. DOI: `10.1128/MMBR.00003-13`.

[6] N. Holm and H. Baltscheffsky, “Links between hydrothermal environments, pyrophosphate, na+, and early evolution,” *Origins of life and evolution of the biosphere : the journal of the International Society for the Study of the Origin of Life*, vol. 41, pp. 483–93, Apr. 2011. DOI: `10.1007/s11084-011-9235-4`.

[7] F. Lipmann, “Projecting backward from the present stage of evolution of biosynthesis,” *The Origin of Prebiological Systems and of their Molecular Matrices*, pp. 259–80, 1965.

[8] Y. Wang, H. Zhang, H. Zhong, and Z. Xue, “Projecting backward from the present stage of evolution of biosynthesis,” *Computational and Structural Biotechnology Journal*, vol. 19, pp. 259–80, 2021. DOI: `10.1016/j.csbj.2021.01.041`.

[9] M. F. Aziz and G. Caetano-Anollés, “Evolution of networks of protein domain organization,” *Sci Rep*, vol. 11, pp. 259–80, 2021. DOI: `10.1038/s41598-021-90498-8`.

[10] S. R. Eddy and the HMMER development team. “Hmmer user’s guide: Biological sequence analysis using profile hidden markov models.” (), [Online]. Available: `http://eddylab.org/software/hmmer/Userguide.pdf`. (accessed: 07.07.2021).

[11] E. Green. “Contig.” (), [Online]. Available: `https://www.genome.gov/genetics-glossary/Contig`. (accessed: 09.07.2021).

[12] H. Menet, V. Daubin, and E. Tannier, “Phylogenetic reconciliation,” *PLoS computational biology*, vol. 18, 2022. DOI: `10.1371/journal.pcbi.1010621`.

[13] M. Lee and S. Ho, “Molecular clocks,” *Current Biology*, vol. 26, R399–R402, May 2016. DOI: `10.1016/j.cub.2016.03.071`.

[14] B. Golding, “Evolution: When was life’s first branch point?” *Current Biology*, vol. 6, pp. 679–682, 1996. DOI: `10.1016/S0960-9822(09)00448-5`.