universität wien

MASTERARBEIT | MASTER'S THESIS

Titel | Title

Reconstruction of the phylogenetic history of the bc1 complex within prokaryotes

verfasst von | submitted by

Constantin Leitgeb BSc

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2026

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 220
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Joint-Masterstudium Evolutionary Genomics and Systems Biology
Betreut von | Supervisor:	Assoz. Prof. Maria Filipa Baltazar de Lima de Sousa PhD

Abstract

The evolutionary history of the bc_1 complex, also referred to as complex III of the electron transport chain, remains a debated subject. The main discussion is whether the complex was already present in the Last Universal Common Ancestor (LUCA) or if it emerged later in bacterial lineages. The aim of this study was to distinguish between these opposing views by analyzing current microbial diversity.

A large-scale genomic comparison was performed using a dataset of over 35,000 assemblies, The three bc1 subunits: cytochrome *c*, cytochrome *b* and Rieske/ISP, were identified in an inhouse dataset by using HMM-based similarity searches. The resulting hits were subsequently filtered and analyzed. By careful refinement of the threshold values of the KEGG models we were able to also identify subunits within the archaeal sequences. This was followed by phylogenetic reconstructions of each of the subunits and a scenario for the history of the complex in Bacteria and Archaea proposed.

The results showed that complete bc_1 complexes are common among bacterial lineages but absent from archaeal genomes. Based on our analysis, Archaea only contains cytochrome *b* and Rieske/ISP subunits. Furthermore, it appears that the longer version of cytochrome *b* subunit resembles the ancestral state, with the shorter version being found in cyanobacteria as part of the b_6f complex as well as in bc_1 complexes from diverse taxonomic affiliations.

These findings lead us to propose a scenario in which LUCA did not possess a bc_1 complex, with this complex being a bacterial innovation. Later, several inter-domain HGT events are responsible for the presence of the cytochrome *b* and ISP subunits in the archaeal domain.

Abstrakt

Die Evolutionsgeschichte des *bc*$_1$-Komplexes, auch als Komplex III der Elektronentransportkette bezeichnet, ist nach wie vor umstritten. Die Hauptdiskussion dreht sich darum, ob der Komplex bereits im letzten gemeinsamen Vorfahren (LUCA) vorhanden war oder erst später in bakteriellen Abstammungslinien entstanden ist. Das Ziel dieser Studie war es, zwischen diesen gegensätzlichen Ansichten zu unterscheiden, indem die aktuelle mikrobielle Vielfalt analysiert wurde.

Es wurde ein groß angelegter Genomvergleich anhand eines Datensatzes von über 35.000 Assemblierungen durchgeführt. Die drei bc1-Untereinheiten: Cytochrom *c*, Cytochrom *b* und Rieske/ISP wurden in einem internen Datensatz mithilfe von HMM-basierten Ähnlichkeitssuchen identifiziert. Die resultierenden Treffer wurden anschließend gefiltert und analysiert. Durch sorgfältige Verfeinerung der Schwellenwerte der KEGG-Modelle konnten wir auch Untereinheiten innerhalb der Archaea-Sequenzen identifizieren. Darauf folgten phylogenetische Rekonstruktionen jeder der Untereinheiten und ein Szenario für die Geschichte des Komplexes in Bakterien und Archaea.

Die Ergebnisse zeigten, dass vollständige *bc*$_1$-Komplexe in bakteriellen Abstammungslinien häufig vorkommen, in Archaea-Genomen jedoch fehlen. Basierend auf unserer Analyse enthält Archaea nur Cytochrom *b*- und Rieske/ISP-Untereinheiten. Darüber hinaus scheint die längere Version der Cytochrom *b*-Untereinheit dem ursprünglichen Zustand zu entsprechen, während die kürzere Version in *Cyanobakterien* als Teil des *b*$_6$*f*-Komplexes sowie in *bc*$_1$-Komplexen aus verschiedenen taxonomischen Zugehörigkeiten zu finden ist.

Diese Erkenntnisse veranlassen uns zu der Annahme, dass LUCA keinen *bc*$_1$-Komplex besaß, da dieser Komplex eine bakterielle Innovation darstellt. Später sind mehrere domänenübergreifende HGT-Ereignisse für das Vorhandensein der Cytochrom-*b*- und ISP-Untereinheiten im Archaea-Bereich verantwortlich.

Contents

Introduction

The bc_1 complex, also known as Complex III of the electron transport chain (ETC), is located in the mitochondria of aerobic and anaerobic eukaryotes as well as in the inner membrane of aerobic and anaerobic prokaryotes. The ETC consists of multiple protein complexes that together facilitate the formation of proton gradient across membranes, which in turn activates the ATP synthase to regenerate ADP to ATP[1].

A schematic overview of the ETC can be found in Figure 1.

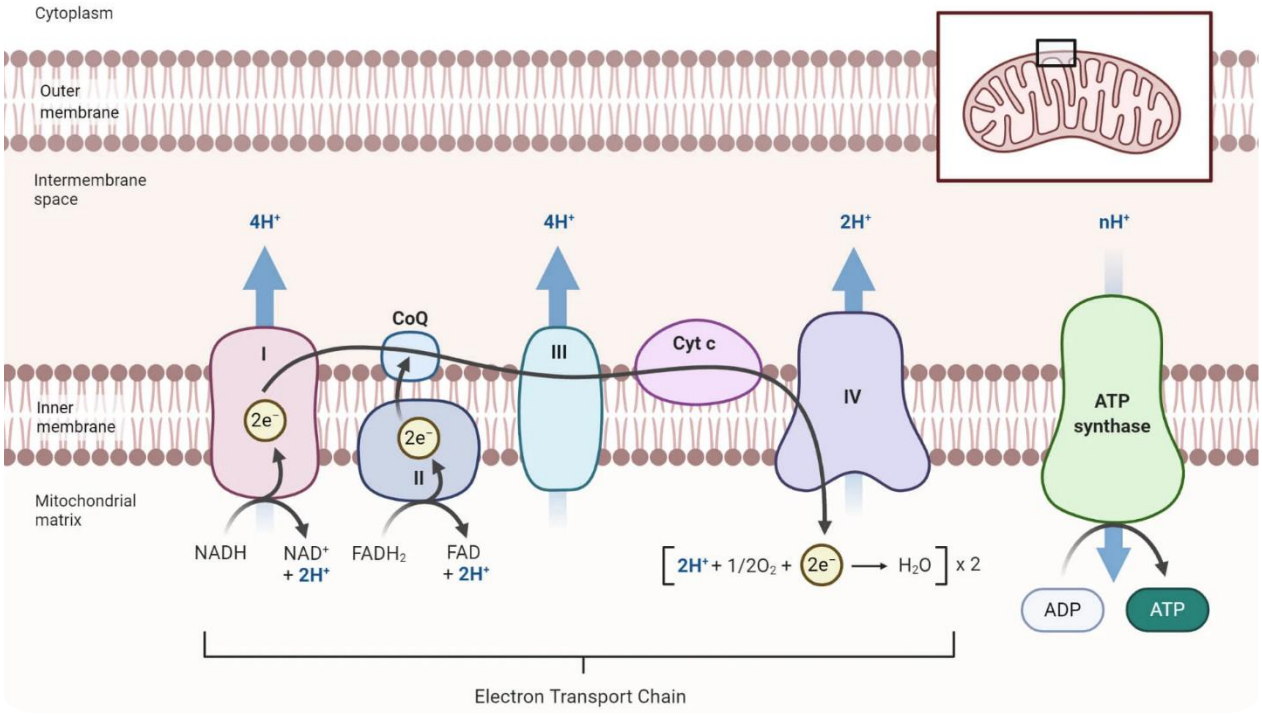

Figure 1: Schematic overview of the canonical electron transport chain in the inner membrane of a mitochondria[2].Created in BioRender[3].

The bc_1 complex operates via the Q-cycle, which uses a bifurcation mechanism to transfer electrons and protons, by coupling an exergonic reaction with an endergonic reaction[4]. To comprehend the function of the bc_1 complex it is essential to understand its structure[5].

This complex has three core subunits: the cytochrome *b* subunit (*cyt b*), the cytochrome c_1 subunit (*cyt* c_1) and the Rieske iron-sulfur cluster subunit (ISP/Rieske)[6]. Several supernumerary subunits may be present, but their function is not fully understood, although they are thought to add structural stability to the complex[6]. The number of subunits can vary from organism to organism, with up to eleven subunits found in humans, but the catalytic core of the bc_1 complex appears to be highly conserved[6].

The Rieske Iron sulfur cluster [2Fe-2S] subunit, which will hereafter be referred to as Rieske/ISP (Iron–Sulfur Protein), is a special form of 2Fe-2S clusters, where instead of the typical four cysteine ligands to coordinate the metal iron atoms, two histidine ligands and

two cysteine ligands coordinate the iron atoms at the center. This replacement results in a higher redox potential of the Rieske center ISP, when compared to canonical 2Fe-2S centers and its specialized role within the bc_1 complex[7].

The cytochrome *b* subunit contains two *b* type hemes. The heme b_H, which has a high redox potential, and the heme b_L, which has a lower redox potential. The cytochrome *b* subunit' transmembrane helices A–D form a four-helix bundle where the two b-type hemes (b_L and b_H) are embedded. Helices B and D supply the conserved histidine ligands that coordinate the heme irons[5].

Cytochrome c_1 acts as an electron transporter to a soluble cytochrome *c*, which then transfers the electron further to complex IV. Both ISP and cytochrome c_1 are membrane-anchored proteins located in the inner membrane of Bacteria and of mitochondria in eukaryotes[5].

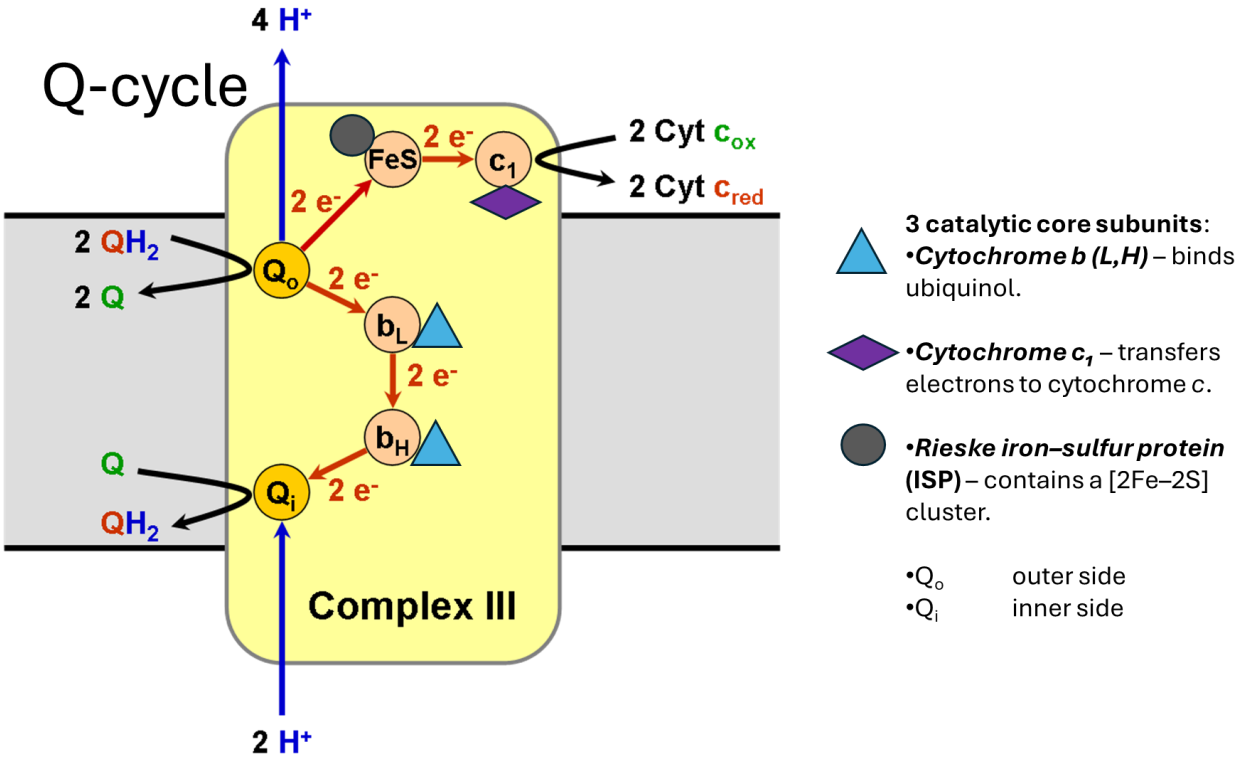

Figure 2: Overview of the Q-cycle. Cytochrome b *is represented as a triangle, cytochrome* c *as a square and Rieske/ ISP as a circle. Image adapted from Hoffmeier(2006)* [8].

Overview of the Q-cycle

The Q-cycle, Q for the involved quinol, describes the mechanism that allows complex III to transfer energy by coupling electron transfer with proton movement across the membrane. It was first described in the 1970s by Peter D. Mitchell[9]. Since then, the model has been revised and updated by several authors, including Antony R. Croft and colleagues[10], who added more details about the reaction kinetics, structural information, and the role of semiquinone intermediates[9,11,12]. Later, high-resolution structural studies provided further insight into the dimeric organization of the bc_1 complex[6]. Currently the

modified Q-cycle is the version that is generally accepted to explain how the energy-transfer mechanism in complex III works.

The *bc₁* complex's quinol oxidation (Q_o) site is located between cytochrome *b* and the Rieske ISP/cytochrome c_1 chain on the membrane's P-side. The two *b*-type hemes link Q_o to the quinone reduction site (Q_i) at the N-side. The two electrons of quinol (Q_H2) are separated at the Q_o site; one is passed through heme b_L to heme b_H, and the other is transferred through the Rieske [2Fe-2S] cluster and cytochrome c_1 to a soluble cytochrome *c*. The quinone (Q) reduction process uses the electron in heme b_H at the Q_i site. To effectively convert a quinone to quinol at the Q_i site, two turnovers of the Q_o site are necessary because only one electron is delivered to the Q_i site per Q_{H2} oxidized. Overall, two protons are taken up from the N-side (matrix/cytoplasm) and four protons are released into the P-side (intermembrane space/periplasm) as a result of the net oxidation of one Q_{H2} and the transfer of two electrons to two molecules of cytochrome *c*[5].

Alternatives of *bc₁* complex

Within plants and *Cyanobacteria*, no bc_1 complex has been reported but rather they possess a homologous complex called the *b₆f* complex, which has a cytochrome b_6 subunit homologous to the cyt *b* subunit and also an Rieske/ISP subunit. Only Cytochrome *f* is not homologous to cyt c_1, but even though their structure differs, it performs an analogous role of its counterpart. A schematic comparison can be seen in Figure 3. The cytochrome *b* of bc_1 has a size of 42kDa, which is twice as large as cytochrome b_6 that only has 23kDa. A similar pattern can be seen in the number of transmembrane helices, while the larger *cyt b* consist of eight transmembrane helices, the b_6 version only has four[13].

Until around 2010, the cytochrome *bc₁* and *b₆f* complexes were the only known complexes capable of transferring electrons from reduced quinones to *cytochrome c*[14]. However, research on *Rhodothermus marinus* revealed an alternative electron-transfer complex[15]. This system, now named the alternative complex III (ACIII), is composed of modules previously known from other respiratory complexes and operates in a distinctly different way from the canonical *bc₁* complexes[16,17]. The most notable differences are its unique subunit composition and the absence of Rieske/ISP and cytochrome *b* subunits. Consequently, ACIII is thought to use a different electron-transfer mechanism and to function without the electron-bifurcation–type Q-cycle characteristic of the *bc₁* and *b₆f*

complexes. Structural and biochemical research on the ACIII is still ongoing[18,19].

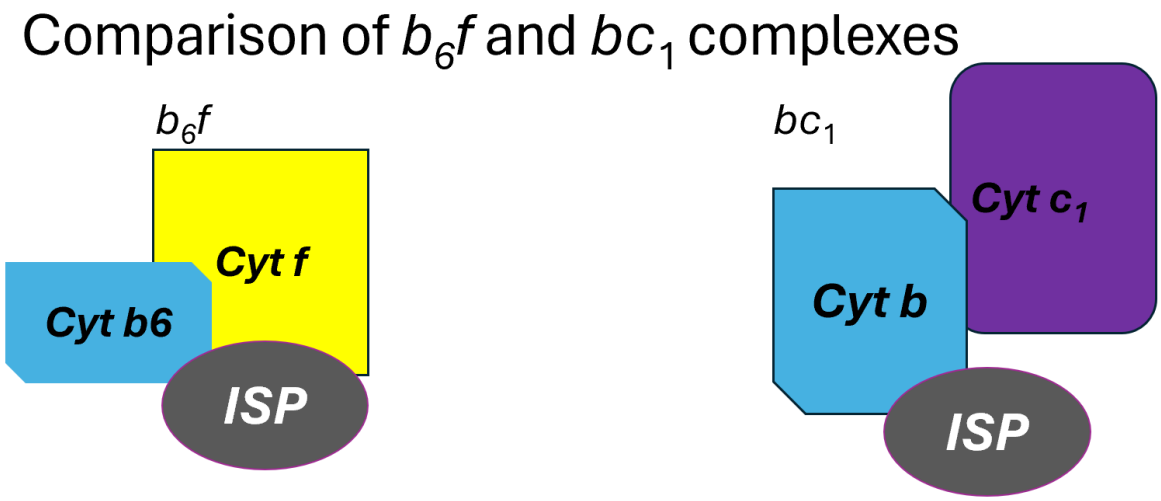

Figure 3:Schematic Comparison of the subunit composition of the bc1 and b6f complexes. Cytochrome b subunits are shown in light blue, ISP in grey and cytochrome f in yellow and cytochrome c1 in purple. Subunits that are shown in the same color are homologous proteins.

Evolutionary History of the bc_1 complex

Considering that, bc_1 and b_6f are homologous complexes, the question arises: which form represents the ancestral state? Over the years, different scenarios about the origin of these complexes have been proposed. However, no clear answer has yet been found and the debate remains. Even the history of the individual subunits is still being debated. For example, the cytochrome *b* and the shorter version cytochrome b_6 subunits could both represent the ancestral state according to different authors. Dibrova et al [20] presented a theory in which the b_6 subunit more likely resembles the ancestral state and that the longer version of cytochrome b was created by multiple independent fusion events of the smaller b_6 (Figure 5A). Furthermore, Dibrova et al questioned the case of LUCA having a Rieske/cytochrome *b* based on the environment of the last universal common ancestor, as the organism would have needed high potential electron acceptors from its environment, which only became available after the emergence of oxygenic photosynthesis[21]. Dibrova's proposed theory not only argues that the b_6f version more likely represents the ancestral state but also that LUCA did not possess such a complex, but rather the complex emerged later within the bacterial domain as seen in the schematic Figure 4 scenario 1 [20,22].

The evolutionary scenario described above was later questioned by Kao and Hunte (2014) and by Ducluzeau and Nitschke (2016). Kao and Hunte analyzed the evolution of these proteins using conserved sequence motifs. They observed that the long version of cytochrome b is present in most archaeal lineages, with the exception of Haloarchaea, which appear to have acquired their cytochrome *b* through a horizontal gene transfer (HGT) event from Actinomycetota. In addition, the short cytochrome *b* variant is absent from archaea. These observations support the hypothesis that the Last Universal Common Ancestor (LUCA) already possessed a *bc*-like complex, as illustrated in

Scenario 2 of Figure 4. Under this assumption, Archaea would be expected to retain some split sequences. In this scenario, LUCA contained the long version of cytochrome *b*, which was later divided into two separate forms in bacterial lineages, each representing either the N- or C- terminus of the cytochrome b as seen in Figure 5B. Kao et al addressed also the arguments of Dibrova et al on the environment of LUCA, stating that aerobic respiration did evolve before oxygenic photosynthesis and also pointed out that a newly discovered methane oxidation pathway of Candidatus Methylomirabilis hints that oxygen might already have been available for metabolism before the emergence of the oxygenic photosynthesis[23].

The case of the two different types of subunits for the different complexes seems more resolved at this point. *Cytochrome f* is not a homolog of the cytochrome c_1, rather it represents a highly specialized version of a cytochrome that is exclusively found in plants, algae and Cyanobacteria[24,25]. The common and widely accepted view is that cytochrome *c* predates the cytochrome *f* [25].

However, the question about the ancestral version of the cytochrome *c* family remains. Cytochromes *c* can be categorized in four classes, each having a different number of heme-binding motifs[26]. During the early 2000s, Baymann et al. proposed that mitochondrial cytochrome c_1 and the di-heme cytochromes found in ε-proteobacteria (now classified within the phylum Campylobacterota) share a common ancestral origin[24]. Previously, these proteins were thought to be unrelated; however, the discovery of a "missing link" protein in the thermophilic bacterium *Aquifex aeolicus* provided key insights into the evolutionary relationship of the cytochrome *c* of the cytochrome c_1 subunit. This protein appears to be a collapsed di-heme cytochrome, representing an intermediate form resulting from the structural and sequence transition from ancestral di-heme cytochromes to the mono-heme cytochrome c_1. Specifically, the data suggest that the shortened "collapsed" di-heme structure—lacking the C-terminal heme-binding motif—gave rise to the mono-heme cytochrome c_1. This evolutionary pathway supports a model in which an ancestral short or partial di-heme cytochrome (or "progenitor" form) underwent mutational events leading to the loss of one heme-binding site, ultimately resulting in the mono-heme form observed in mitochondria, as shown in schematic Figure 5C.

To summarize, all current theories share the fact that of the ancestral complex would have at least a cytochrome b, either the short or the long version, and an Rieske/ISP subunit. Therefore, some authors attempted to establish a different, more coherent name, “the Rieske/cytochrome *b* complexes”[27,28]. In this thesis, I will continue to use the bc_1 name

instead of Rieske/cytochrome *b* for the sake of clarity and continuity.

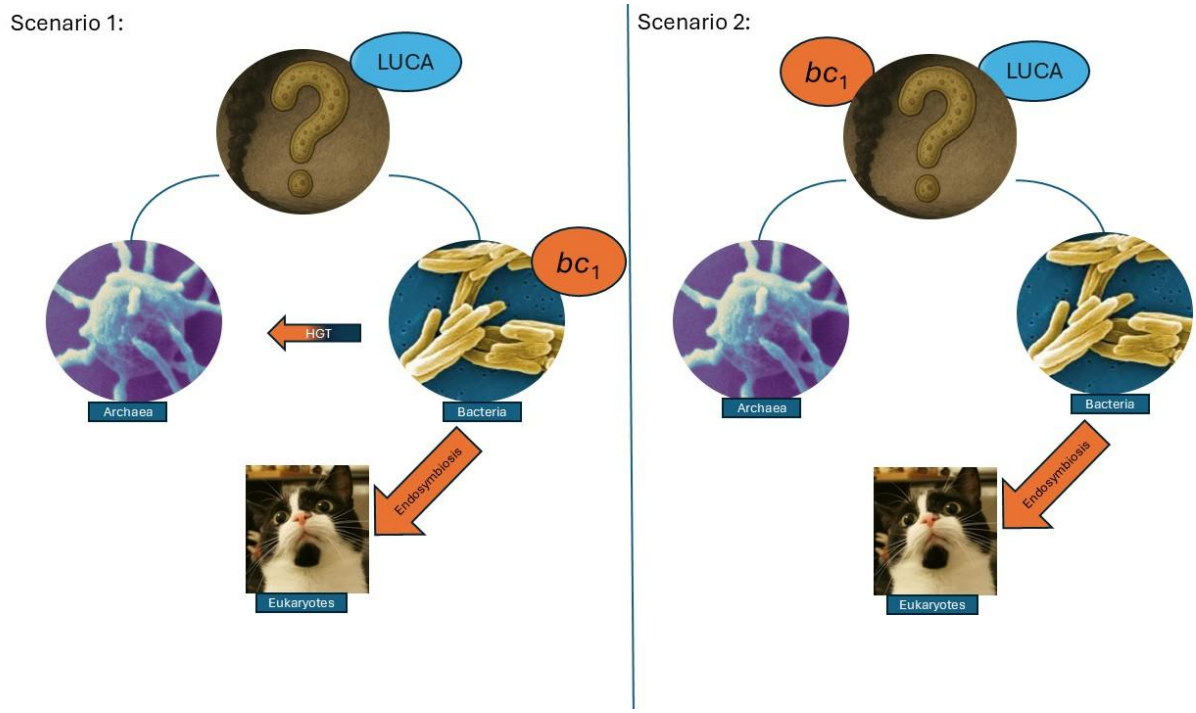

Figure 4:Schematic overview of the two scenarios presented of the possible case of origin of the bc_1 *complex. Archaea are represented in the figure by Lokiarchaeum ossiferum [29], bacteria are represented by Mycobacterium tuberculosis [30]. Eukaryotes are represented by a cat [20,22].*

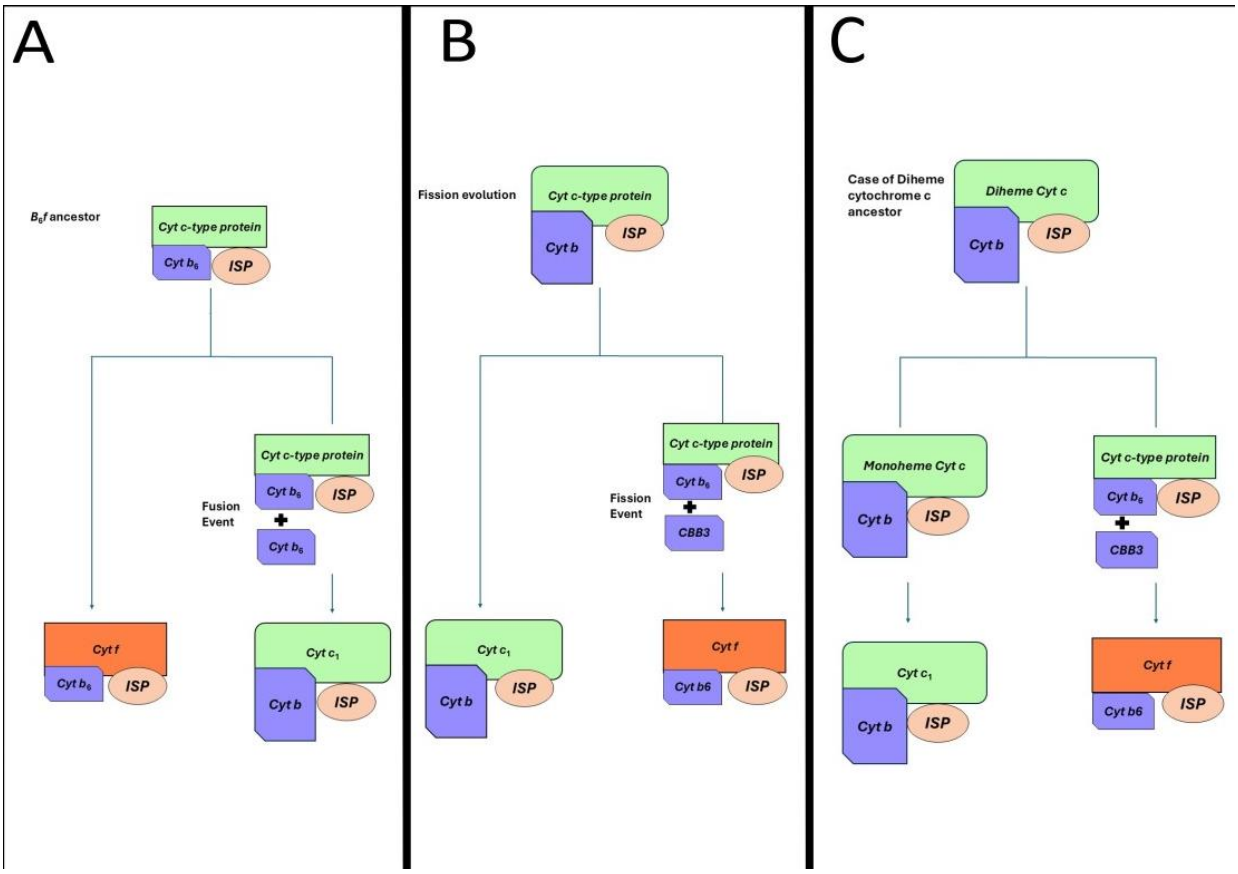

Figure 5 : Evolutionary scenarios for the evolution of the bc_1 and b_6f (Rieske/b) complexes. A: Possible scenario representing the ancestry of b_6f type complex with bc_1 complex evolving later via replacement of the f subunit and fusion of the b_6 subunit[20,22] . B: Ancestry of the bc_1 complex with a fission event and replacement of the c subunit giving rise to the b_6f complex[23]. C: Evolution of cytochrome c_1 from a diheme version to a monoheme version[24]. Cytochrome c subunits are represented in green, ISP is shown in peach, blue are the cytochrome b subunits and cytochrome f is shown in orange colors. Subunits that share a color are homologs.

Data and Methods

The computational results of this work have been achieved using the Life Science Compute Cluster (LiSC) of the University of Vienna. The research utilized Python v.3.12.0[31] alongside the Linux operating system for data processing and analysis. To enhance the clarity and quality of the written content, AI tools such as ChatGPT[32] and Perplexity[33] were employed. Additionally, portions of the code were developed with assistance from GitHub Copilot [34], an AI-powered code completion tool.

Data Processing and Manipulation

To streamline data handling and enhance computational efficiency, the Python libraries Pandas v.2.1.3 [35] and NumPy v.1.26 [36] were employed. These tools facilitated structured data management and efficient processing throughout the project.

Dataset

The dataset that was used in this analysis was downloaded from NCBI in November 2019

[37]. It comprises 15,518 proteome entries from both Archaea and Bacteria. As an initial step, the genomes were sorted into their respective domains of life, facilitating a more structured and efficient downstream analysis.

Visualization of Results

To enhance result interpretation, data was visualized using Matplotlib v.3.8.0 [38] in Python. Scatterplots and line plots illustrated the data distribution and assisted in refining threshold values. Experimentally confirmed entries from the BRENDA database were highlighted in bright red for clarity. These visualizations are included in the results section and supplementary information.

Determination of KEGG Model Cut-Offs

Identification of Functional Orthologs of the bc1 Complex

The two most important concepts for this thesis are homology and orthology. Homology describes the relationship between two points of a tree that have a common ancestor. Both points originated from the same node, but this does not imply that they share a common function. To be able to reconstruct the relationships of sequences, a certain degree of sequences similarity must be observable. Orthologous is a term for describing homologous genes that arose by a speciation event[39]. Paralogs, on the other hand, are the homologs resulting from the gene duplication in a common ancestor. Another example would be xenology, in which a gene is passed on by horizontal gene transfer (HGT) from one distantly related species to another[39].

Here, the focus was on finding and identifying functional orthologous protein sequences, namely to find protein sequences of the bc_1 complex that share a common origin through a speciation event and still retain the same role or function[40].

To address this issue, profile Hidden Markov Models (HMM) were employed. They can represent a wide range of sequences, including DNA and proteins, and account for variations among the sequences to which they are compared. This is achieved by using transitions, match states, insert states, and delete states. In the sequence family, match states are conserved positions. At specific locations, insert states may permit additional residues. Certain positions can be left empty in delete states. Additionally, translations are the likelihood of moving from one residue to another. Profile HMMs have an advantage over pairwise alignment searches in that the various model states capture subtle differences and patterns, particularly for distant homologs[41]. To identify homologous sequences, an Hidden Markov Models (HMMs) search strategy was employed. The models used in this study correspond to version 2022-11-01 from the KEGG [42] release 104.0 and were downloaded on 2.12.2024.

HMM models were run with HMMER v.3.4[43] to identify functional orthologs within the dataset. These HMMs were sourced from the Kyoto Encyclopedia of Genes and Genomes

(KEGG) database via the website [44]. KEGG models provide predefined threshold values and concise descriptions of the orthologous groups of proteins they represent. Each KEGG model is assigned a unique KO (KEGG Orthologs) number. For this study, ten KEGG models were selected, each corresponding to a subunit of the bc_1 complex. Table 1 provides a detailed overview of these models, including relevant annotations and characteristics.

Table 1:KEGG models used in this project collected from the KOFAM database including the initial threshold values.

Subunit	KO number	Description	Threshold	Lowered Threshold
Cytochrome *b*	K00412	ubiquinol-cytochrome c reductase cytochrome b subunit	288.80	210
	K03887	menaquinol-cytochrome c reductase cytochrome b subunit	286.33	210
	K03891	ubiquinol-cytochrome c reductase cytochrome b subunit	489.80	290
ISP (Rieske subunit)	K03886	menaquinol-cytochrome c reductase iron-sulfur subunit [EC:1.10.2.-]	115.90	90
	K00411	ubiquinol-cytochrome c reductase iron-sulfur subunit [EC:7.1.1.8]	81.03	50
	K03890	ubiquinol-cytochrome c reductase iron-sulfur subunit	202.87	50
Cytochrome c_1	K00413	ubiquinol-cytochrome c reductase cytochrome c1 subunit	39.43	20
	K03888	menaquinol-cytochrome c reductase cytochrome b/c subunit	171.53	90
	K03889	ubiquinol-cytochrome c reductase cytochrome c subunit	91.87	70
cytochrome *b*/ cytochrome c_1	K00410	ubiquinol-cytochrome c reductase cytochrome b/c1 subunit	715.67	490

Similarity searches

HMM models describe observable events dependent on internal factors that cannot be directly observed [45]. For this study, the hmmsearch function was chosen over hmmscan due to its efficiency. The -domblout flag generated a detailed output, which was subsequently filtered to retain only relevant taxonomic and sequence information [46].

Significant hits were determined by filtering HMMER results, as the program reports all potential hits. Hits were considered as functional orthologs if their bit-scores exceeded threshold values (as presented in Table 1). The process of filtering bacterial sequences involved identifying hits that had multiple annotations corresponding to different KEGG models. In such cases, the hit with the higher bit score was retained; if the bit scores were identical, the E-value was compared, and the hit with the lower E-value was kept.

Comparison with the BRENDA Database

To evaluate and potentially refine model thresholds, results were compared with the manually curated BRENDA database [47], a primary repository for enzyme functional data. Each functional enzyme is assigned an Enzyme Commission (EC) number. In this study, models corresponded to either [EC:7.1.1.8] [48] or [EC:1.10.2.2] [15], both representing quinol-cytochrome-*c* reductase were used. The comparison involved cross-referencing HMMER results with BRENDA's dataset, obtained from the 2.12.2023 release version 2023.129 [49], to determine whether threshold adjustments were necessary. This was done by highlighting the entries that were downloaded from BRENDA that had experimentally confirmed cases of bc_1 complexes.

Presence-Absence Taxonomic Distribution

A presence-absence matrix was generated to provide a comprehensive overview of model occurrences and to show the taxonomic distribution. This matrix, visualized as a heatmap, helped to fine-tune the threshold values for KOs. Seaborn v.0.13.1 [50] in Python was used to create these plots. Each KEGG model's occurrence across genomes was normalized by phylum, yielding values between zero and one for easier interpretation. Absence of subunits in a phylum was denoted by white cells in the heatmap.

Rerun of HMMER

Following threshold adjustments, the analysis was repeated with an expanded set of HMM models. Instead of the initial 10 models, all models from the same KEGG release were included. This ensured accurate assignment of sequences, as some were reassigned based on improved scoring criteria. The workflow otherwise remained unchanged, and the revised threshold values are presented in Table 1.

Additional Functional Annotations

Information regarding domain regions within the proteins was obtained from the PFAM [51] database, which is part of the InterPro framework. The specific InterPro release version 33.1 was downloaded in October 2023 from their website.

Based on the existing PFAM annotation table from the in-house frozen set of genomes, the hits were mapped to PFAM-provided domain information, including position and overlap detection. A script checked each query for multiple domain annotations. If a sequence was a fusion of two different domains it was divided into separate FASTA files, extending the cut location by 15 amino acids (AA) before and after the PFAM-reported position to preserve contextual information.

Entries containing two or more different cytochrome c domains were excluded from further analysis to ensure the reconstruction of a single phylogenetic gene tree. Two approaches were applied: the PFAM-based approach described above and a motif-based approach checking for multiple occurrences of the CXXCH [52] motif.

Local and Global Alignments

Hits were sorted into three groups, each representing a subunit of Complex III. All-versus-all local alignments were performed using DIAMOND v.2.1.10.164 [53] in sensitive mode for faster performance compared to BLAST. Results were filtered by identity score and E-value, retaining pairs with an identity ≥25% and an E-value ≤10^{-10}. These results were used as input for global alignment with EMBOSS Needle v.6.6.0.0, implementing the Needleman-Wunsch algorithm[54]. Here, the threshold was set at identity > 25%, but E-values were not considered as they are not reported from the program.

Clustering

Clustering was performed using the Markov Cluster Algorithm (MCL) [55]. This is an approach to group more similar entries within a network of nodes and edges, where nodes are entries and edges are the relationship between the entries. As seen in Figure 6 each dot represents a sequence, and each line represents the similarity values between each sequence. By simulating a flow or random "walk" between the entries a pattern appears as the sequences with higher similarly tend to have more possible lines or paths between them and a cluster is formed. Since the number of clusters to be generated cannot be directly specified, the inflation parameter regulates the granularity of the clustering. The smaller the values of inflation the larger the clusters will be and thus the total number of clusters will be smaller and vice versa for bigger values of the inflation. With an inflation value of 1.2 for optimal granularity intra- and inter-cluster identity matrices were generated and plotted as heatmaps using R v.4.4.2 [56] and Seaborn [50] in Python,

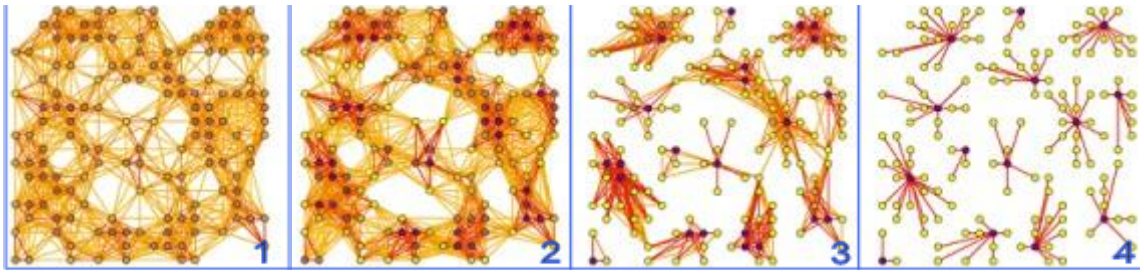

Figure 6: Overview of the MCL algorithm. Image adapted from the MCL website[47].

respectively.

Conserved Domain Search

Sequences from each cluster were analyzed against the conserved domain database (CDD v3.21, accessed 22.11.24 and 4.12.24)[57]. Since the online tool limits input to 1000 sequences per FASTA file, the splitfasta function from GenomeTools v.1.6.2 [58] was used to maintain proper fasta format for the input files.

Synteny Analysis

Feature table files downloaded from NCBI were parsed to extract sequences of interest, including three upstream and three downstream coding regions for synteny comparison. This information was used to determine whether a genome contained the necessary proteins to form a bc_1 complex, b_6f complex or neither. For genomes lacking a full set of subunits within this gene window, further analysis using InterPro and BLAST was conducted.

Interproscan v.5.74-105.0-11.0.4 [59] was used on all proteins sequences of our initial set, to gather as much information as possible, instead of just focusing on the hit sequences and their neighbors

Classification of proteins

Since KEGG does not distinguish between the short and long cytochrome *b* versions, additional steps have been taken to classify these proteins correctly. Protein sequences were queried against a reference database of canonical proteins using DIAMOND BLAST. The canonical protein dataset was downloaded from UniProt on 02.05/05.05 and 15.07.2025 [60–64]. To distinguish between different versions of the cytochrome *b* subunit, a length-based classification was implemented. Canonical b_6 sequences typically range from 200 to 250 amino acids. However, some sequences were present as fusion proteins and were manually split as described before. During this splitting process, 15 amino acids were retained at the overlap regions to avoid losing any information. Therefore, a threshold of 260 amino acids was chosen to ensure that all biologically relevant variants, including these extended split sequences, were captured within the analysis.

Improved Taxonomic Distribution

After completing the classification steps described above, the results were integrated into the taxonomic distribution heatmap to identify genomes capable of forming a complete bc_1 or b_6f complex. Complex formation was assessed by using three criteria. First, KEGG annotations were examined to verify the presence of all required subunits. Second, the synteny was evaluated to determine whether all subunits' genes occurred together within the defined window. Third, cases were included in which all necessary subunits were present in the genome but scattered outside the defined synteny window. As a result, a new column was added to the heatmap to highlight genomes able to assemble a full complex according to any of these criteria.

Multiple Sequence Alignments

Retained sequences were combined into FASTA files corresponding to the three subunits and aligned with Clustal Omega v.1.2.4 [65]. Alignments were visualized with a custom script, trimmed using TRIMAL v.1.5 [66] with a conservation threshold of 60%, and re-plotted to check the quality of trimming and plotting.

Phylogenetic Reconstructions

Phylogenetic reconstruction is a visualized representation of relationships between either organisms, genes or proteins. Often, they are also referred to as trees, based on their looks. They are widely used to better understand the key changes in evolution and for inferring the common origin of genes or species. Maximum likelihood phylogenetic reconstructions were generated using IQTREE v.2.3.6 [67] with automated model selection[68]. The Minimal Ancestor Deviation (MAD) method was applied for rooting the phylogenies[69]. Final trees were visualized and annotated using Figtree v.1.4.4 [70], running locally.

Results and Discussion

HMMER Results

The dataset used consisted of 15.518 genomic assemblies, of which 14.885 were bacterial and 633 were archaeal. Following the separation of the genomes into their respective domains, the option hmmsearch of HMMER was run on the proteomes of the Bacteria, followed by the proteomes of Archaea. In total, HMMER identified 41.9761 sequences with potential functional orthologous relationships to the bc_1 complex. Subunits could be present more than once in a genome.

The list of potential functional orthologs was filtered based on the set of initial threshold values. A total number of 27.165 sequences passed the initial thresholds and the distribution for each model can be seen in Table 2. No archaeal sequences passed the initial thresholds.

Table 2: List of the used KEGG models and the number of assigned sequences for each model.

KEGG models	Total Number of Sequences that passed the initial threshold
K00410	329
K00411	5366
K00412	5254
K00413	5052
K03886	1205
K03887	1007
K03888	970
K03889	2505
K03890	2405
K03891	3072

Validation of Acquired Sequences Against Experimentally Confirmed bc_1 Complexes via BRENDA

The results were compared to experimentally confirmed cases for species possessing a functional complex III. The EC entry page for the bc_1 complex included 98 species also present in the internal database used by our group. Using the initial cut-offs, bc_1 subunits could only be retrieved in 32 species, with a total number of retrieved sequences of 190. This indicated that the initial threshold covered only 33.3% of the experimentally characterized species diversity and needed to be adjusted.

Table 3: Comparison based on BRENDA. Number of sequences found by HMMER and the number of sequences that passed the threshold per model.

KEGG models	Number of total sequences found by HMMER	Number of sequences that passed the threshold
K00410	150	3

K00411	178	26
K00412	168	21
K00413	209	20
K03886	297	40
K03887	106	40
K03888	437	40
K03889	564	0
K03890	314	0
K03891	331	0

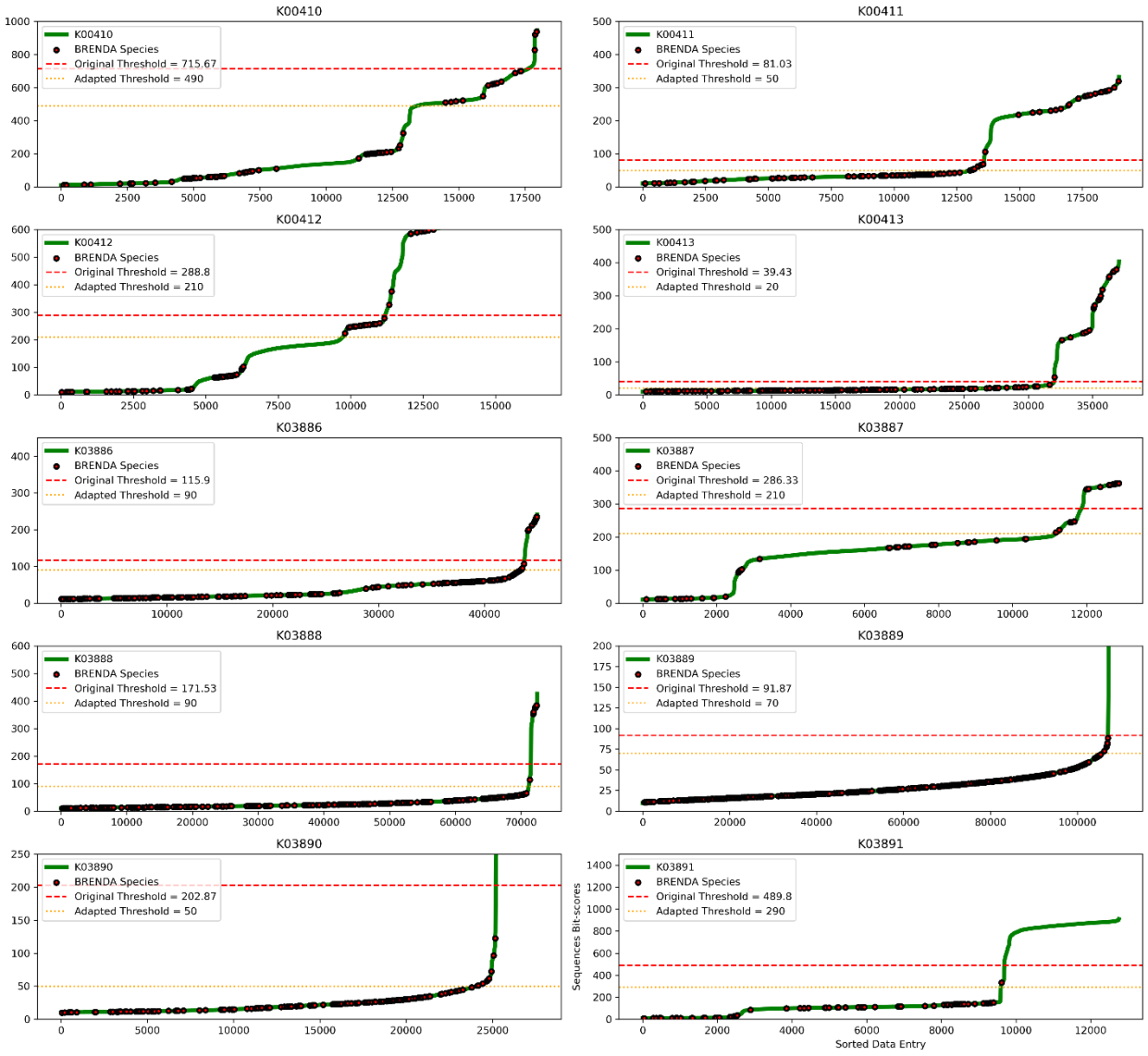

Figure 7: Comparison between Brenda entries scores in relation to initial and adjusted threshold values. The red line represents the original threshold value; the orange line represents the new threshold value. The green line represents each hit that was found by HMMER. The black dots represent BRENDA data.

As seen in the values of Table 3, the initial threshold was too high in order to capture most of the orthologous sequences. Consequently, the threshold had to be reduced significantly, as further illustrated in the plots in Figure 7. Based on this analysis, bit-scores were adjusted to the final values shown in Table 1.

The threshold values were reduced and then compared again with the BRENDA Line plot in Figure 7. With the adjusted thresholds, the overall hit recovery rate was increased for all models, as can be seen in Table 4.

Table 4: Improvement of the identification of Brenda' bc_1 *characterized sequences with the new cut-offs.*

Model	Total Sequences found	Sequences that passed the original Threshold	Sequences that passed the adapted Threshold	Improvement (%)
K00410	215	3 (1.4%)	19 (8.8%)	+7.4%
K00411	210	28 (13.3%)	48 (22.9%)	+9.6%
K00412	220	21 (9.5%)	118 (53.6%)	+44.1%
K00413	231	20 (8.7%)	41 (17.7%)	+9.0%
K03886	340	69 (20.3%)	86 (25.3%)	+5.0%
K03887	143	85 (59.4%)	97 (67.8%)	+8.4%
K03888	480	56 (11.7%)	57 (11.9%)	+0.2%
K03889	644	0 (0.0%)	6 (0.9%)	+0.9%
K03890	256	0 (0.0%)	11 (4.3%)	+4.3%
K03891	152	0 (0.0%)	1 (0.7%)	+0.7%

Lowering the threshold also allowed to detect functional orthologous sequences belonging to Archaea. Moreover, with the new thresholds, it was possible to retrieve hits for three models for which no hit could be assigned with the initial thresholds. Those were cases in which hits were only slightly below the initial threshold values (Figure 7).

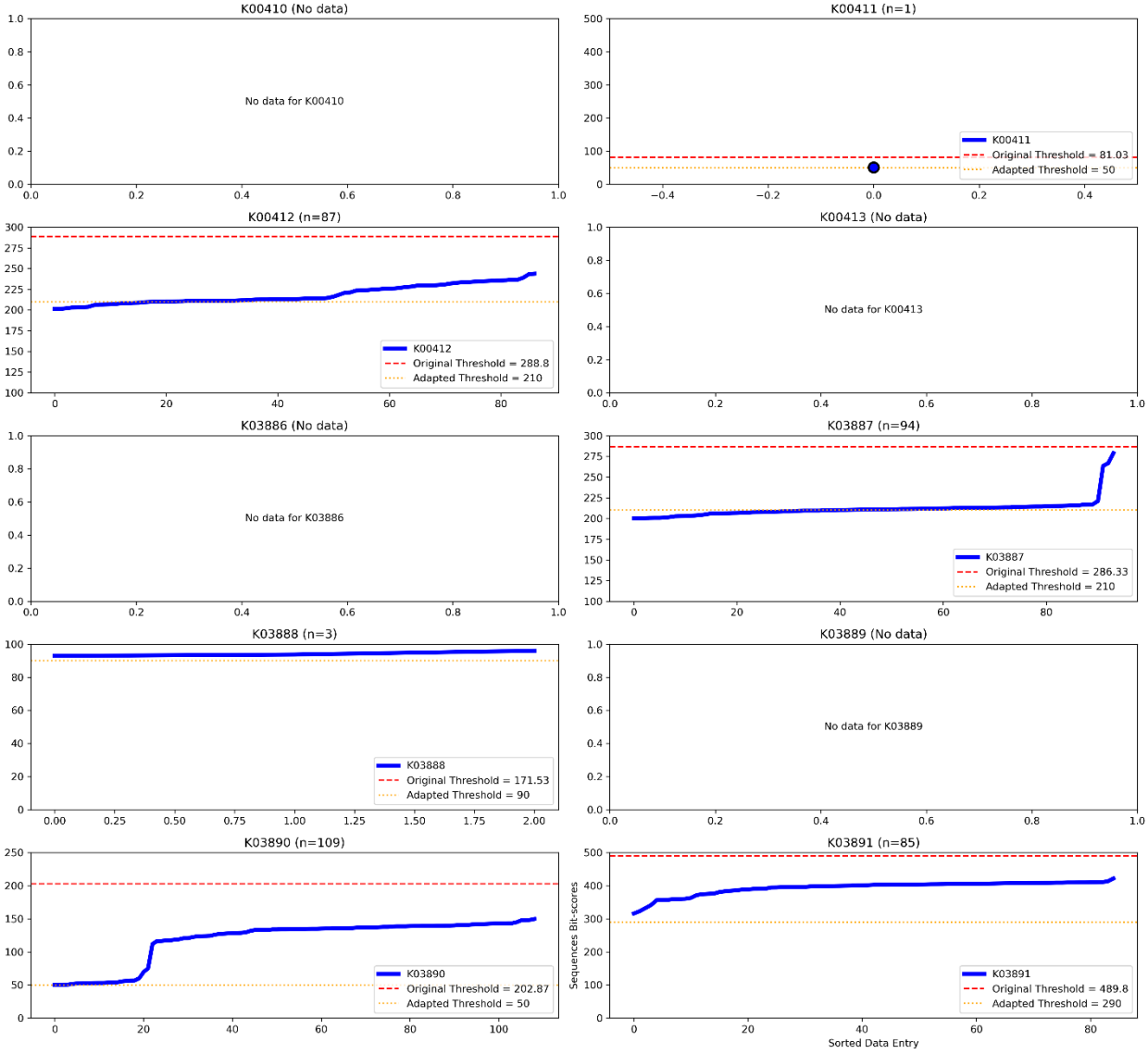

Figure 8: Comparison of retrieved archaeal hits before and after threshold adjustment. The blue line represents archaeal hit scores.

The KEGG models do not appear to have been designed with archaeal analyses in mind, as no matches passed the initial significance threshold. Only after lowering the threshold values significantly, hits were reported. As a comparative approach, HMMs derived from alternative databases, such as Clusters of Orthologous Groups (COG)[71] or eggNOG[72] could have been used.

Table 5: Threshold adjustment effect in archaeal sequences.

Model	Entries	Score Range	Mean Score	Original Pass	Adapted Pass
K00410	0	-	-	-	-
K00411	1	51.3	51.3	0 (0%)	1 (100%)
K00412	87	201.3 - 243.8	218.4	0 (0%)	69 (79.3%)
K00413	0	-	-	-	-
K03886	0	-	-	-	-
K03887	94	200.3 - 278.9	211.8	0 (0%)	57 (60.6%)
K03888	3	92.9 - 96.0	94.2	0 (0%)	3 (100%)
K03889	0	-	-	-	-
K03890	109	50.2 - 149.7	118.5	0 (0%)	109 (100%)
K03891	85	316.0 - 421.8	393.9	0 (0%)	85 (100%)

Using the adjusted thresholds, a total of 52792 bacterial and 640 archaeal hits for Archaea were retrieved. An in-house PFAM annotation table was used to annotate these hits, and the resulting descriptions were compared to the earlier KEGG-based classifications. Sequences with contradictory annotations were removed and not considered for further analysis. Following these steps, 28315 bacterial sequences were kept.

The number of cytochrome *c* per sequence was also analyzed as described in Methods. Most cytochrome *c* subunit sequences analyzed for this project had only one heme *c* motif, while some had two. In addition, using the PFAM and Kegg annotations, 728 sequences were identified as a fusion of different subunits. These were divided into their unfused components and added to the respective clusters (see below). In total, 29033 bacterial and 379 archaeal sequences were kept.

Clustering

After completing the filtering step, local and global alignments were made as described in Methods. The initial set of results contained over 74 million unique relationship pairs for comparison. After applying a filter to keep the sequence pairs with an E-value of less

than or equal 1^{-10} and an identity score above 25%, approximately 65 million pairs remained. To optimize computational and time efficiency, only the ones passing the thresholds, were selected for global alignment. Results were filtered to keep only pairs that had a global identity score above 25%. The resulting global relationship pairs were used as input for the Markov Cluster Algorithm (MCL), which grouped them based on their identity scores.

MCL produced 71 clusters with a total number of sequences of 29,412. For each individual cluster, inter- and intracluster heatmaps were generated. Intercluster relationships were plotted for each subunit to evaluate if the clusters could be merged for subsequent analysis. The distribution of KEGG assignments per cluster is presented in Figure 9. Green tones correspond to the cytochrome *b* subunit, blue tones for ISP/Rieske, red tones for cytochrome c_1 and the fusion cytochrome b/c_1 is represented in purple.

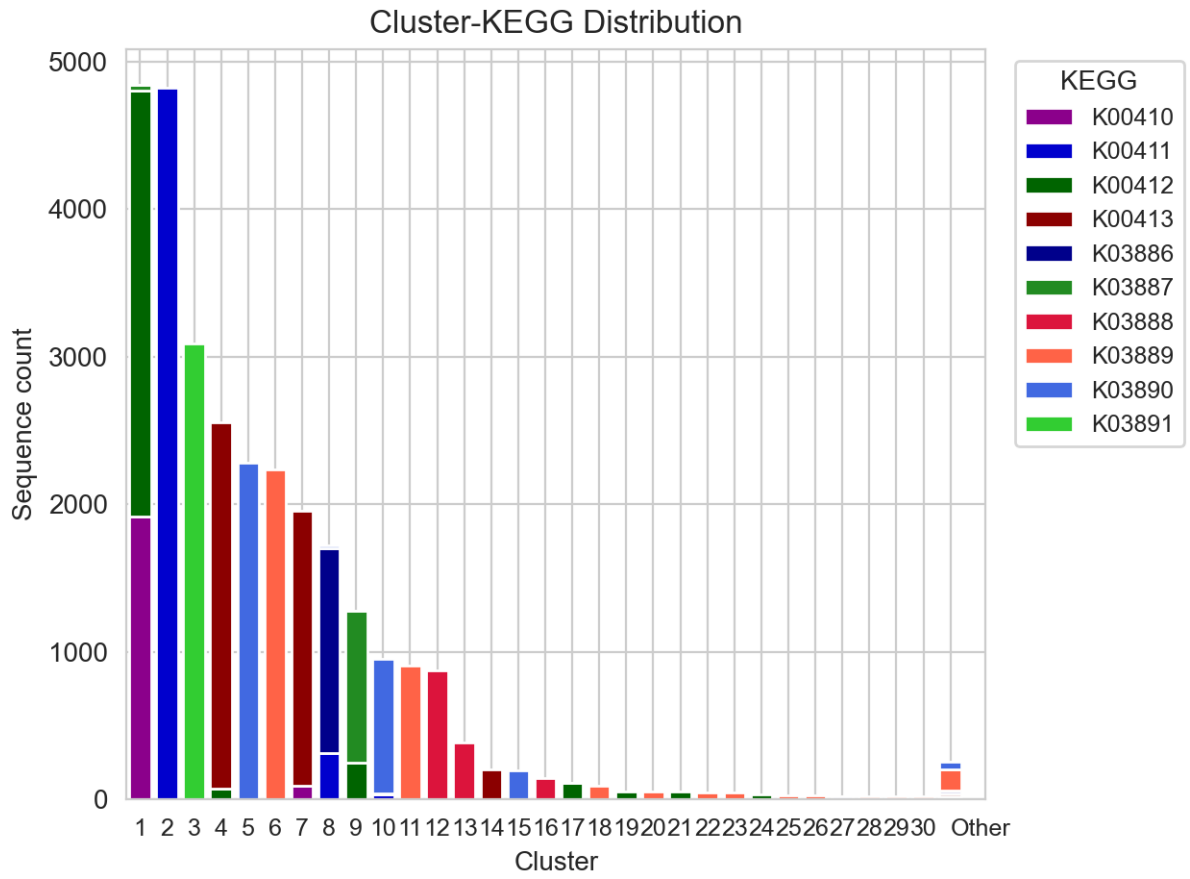

Figure 9: KEGG KO classification per cluster.

The relationships between clusters were plotted and are shown in Figure 10A–C. No intercluster relationships were observed between different subunits. All entry pairs with a similarity score above 25% are displayed. Only the first 30 clusters were included in the analysis, as the number of sequences per cluster declined rapidly and clusters with fewer than 20 sequences were omitted. Given the high computational demands of phylogenetic reconstruction, this restriction was applied to reduce analysis time and computational load. Since the heatmaps indicated that some sequences and clusters might have been incorrectly annotated, all sequences were subsequently checked using the NCBI Conserved Domain Search (CDS). For example, as can be seen in Figure 10C, clusters 12 and 13 show a clear relationship to each other and further inspection revealed that these clusters contained sequences belonging to complex IV. This indicates that KEGG models are not able to correctly distingluish between *cbb3* suunits and cyctochrome c_1 from bc_1 complex. For the cytochrome *b* and ISP subunits, no additional refinement of clusters was required. Each set of clustered sequences subunit group.

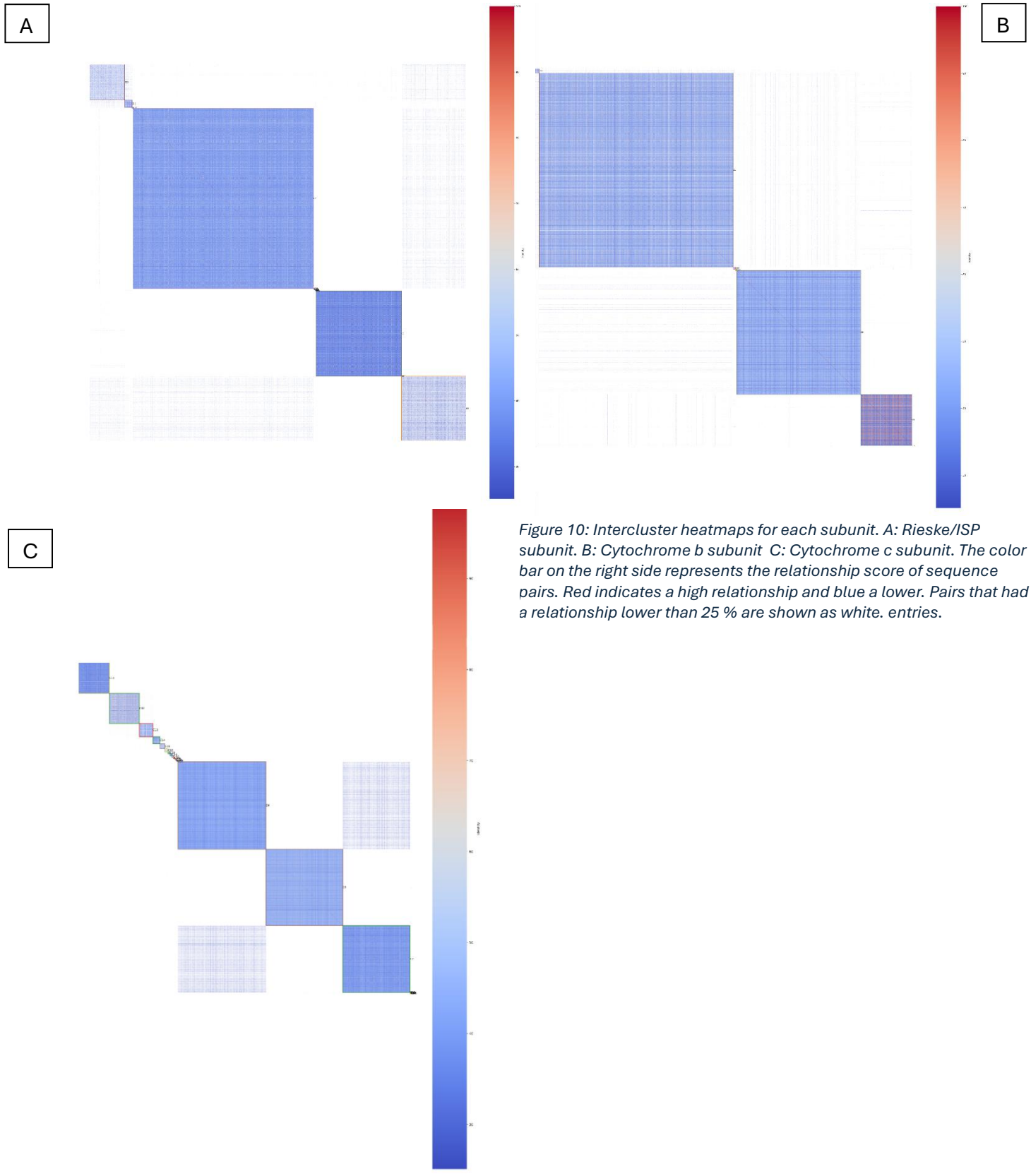

Figure 10: Intercluster heatmaps for each subunit. A: Rieske/ISP subunit. B: Cytochrome b subunit C: Cytochrome c subunit. The color bar on the right side represents the relationship score of sequence pairs. Red indicates a high relationship and blue a lower. Pairs that had a relationship lower than 25 % are shown as white. entries.

Multiple sequence alignments

Each set of clustered sequences was merged within its respective subunit group, and separate MSAs were then generated for cytochrome *b* and for ISP.

Multiple MSAs were generated using different combinations of cytochrome c_1 subunit clusters, as the initial alignment contained numerous misalignments (Figure 11B) and could not be used for downstream analysis. To improve the alignment quality, alternative cluster groupings were tested and evaluated. The most reliable MSA was obtained by combining clusters 4, 6, and 7 as shown in Figure 11A. In total, 10002 ISP/Rieske sequences, 9409 for cytochrome *b* sequences and 6752 cytochrome c_1 sequences were used for the following steps.

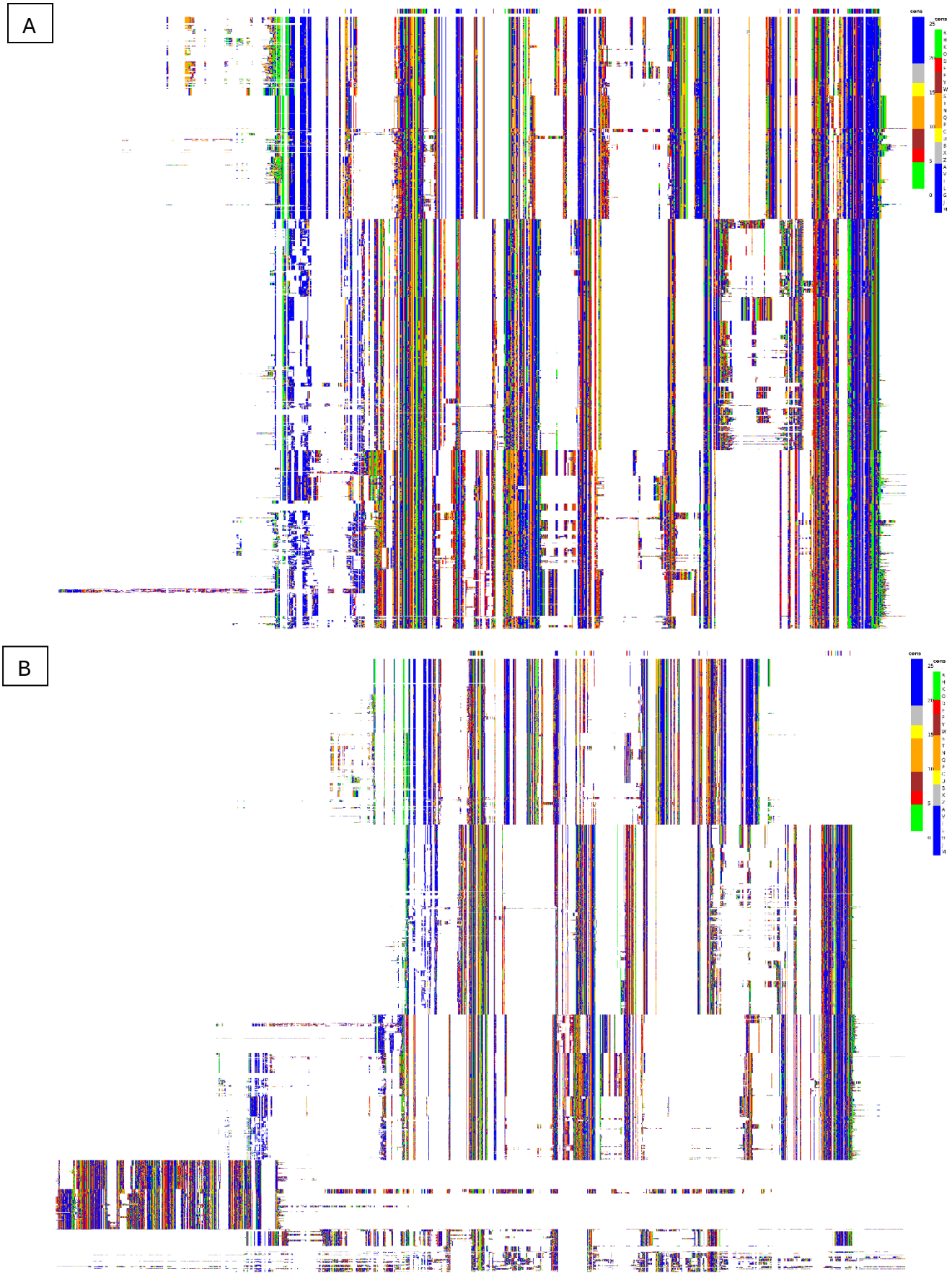

Figure 11: A, MSA of clusters 4,6 and 7. When compared to the other alignment the quality has improved, All present sequences are well aligned. B, MSA for all Cyt c containing clusters. The resulting alignment was of poor quality and was not used for further analysis.

The alignments for the subunit cytochrome *b* showed a section that appeared much

Taxonomic Distribution

A total of 120 different classes of Bacteria and Archaea were represented and checked based on the presence of the necessary subunits required to form a full complex. A general overview of the formation capabilities of the Rieske/cytochrome *b* complexes can be found in Table 6, while Figure 12 summarizes the taxonomic distribution of classes that were able or unable to form a complete complex.

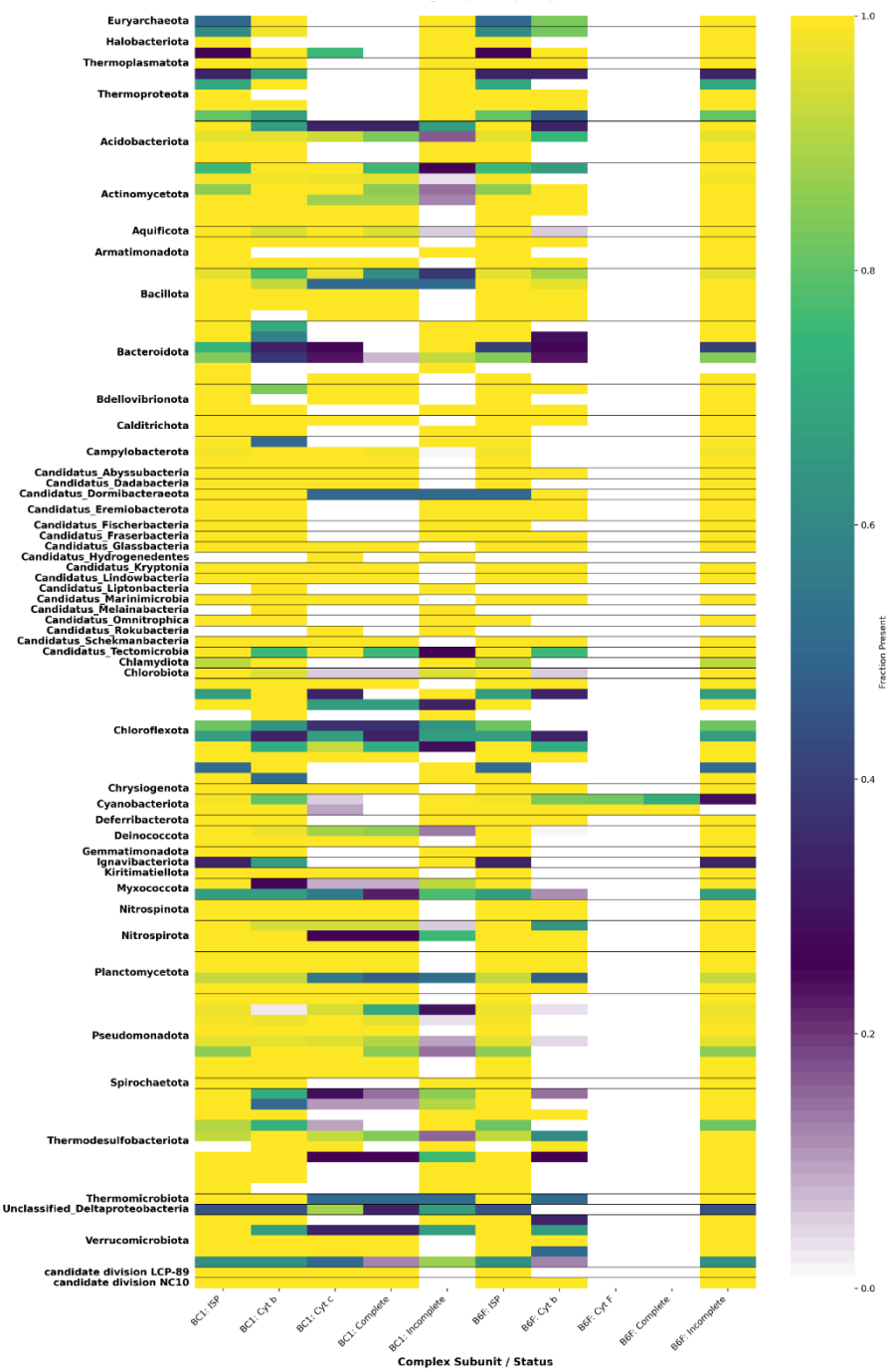

Figure 12: Taxonomic Distribution on phyla level. Normalized Prescence Absence Matrix: This figure shows classes of bacteria and archaea with the presence or absence of each subunit indicated at different annotation levels. White cells indicated the absence. The first 4 columns represent the presence of a subunit for the bc_1 complex and if a complete complex could be formed. The following 4 columns show the same but for the b_6f complex. At the top we have Archaeal entries followed by Bacterial.

Table 6: Breakdown of the number of genomes that had all necessary subunits to form a full complex.

	ISP	*Cytochrome b*	*Cytochrome* c_1
Total	10002	9409	6752
Complete bc_1	7955 (79.5%)	7747 (82.3%)	6028 (89.3%)
Incomplete bc_1	1727 (17.3%)	1429 (15.2%)	724 (10.7%)
Complete b_6f	276 (2.8%)	178 (1.9%)	0
Incomplete b_6f	44 (0.4%)	55 (0.6%)	0

From the 29.412 sequences that were analyzed, a total of 22.184 were assigned to be part of a complete complex, either based on synteny or whole-genome content.

For the Rieske/ISP subunit, out of 10,002 sequences, 7,955 were able to form a full bc_1 complex and 276 were assigned to the b_6f complex. The most frequently missing subunit in these cases was cytochrome b, either in the short or long form.

The cytochrome *b* dataset contained 9,409 sequences, of which 7,747 were assigned to be able to form a full bc_1 complex and 178 to a b_6f complex. Here, the most commonly missing subunit was cytochrome c_1.

The cytochrome c_1 subunit showed one major difference, as no sequences were associated with the b_6f complex, since this complex contains cytochrome *f* instead of cytochrome c_1. Out of 6,752 sequences, 6,028 were assigned to be able to form a full bc_1 complex. Overall, the most commonly subunit absent across all datasets was cytochrome *b*.

No archaeal genome was found to be capable of forming a complete complex, as neither cytochrome c_1 nor cytochrome *f* could be identified with the selected threshold values. The b_6f complex was largely absent from bacterial phyla, apart from the *Cyanobacteriota*. In contrast, bc_1 complexes were widely distributed within bacteria classes.

Maximum Likelihood Phylogenetic Reconstructions

The phylogenetic reconstructions corresponding to subunits Rieske/ISP, *b*, and c_1 are shown in Figures 13 to 15, respectively. Overall, the phylogenies present a similar topology with a split between *Alphaproteobacteria* and *Gamma-*, and *Betaproteobacteria* sequences and with a clade dominated by *Actinomycetota sequences on the other end of the phylogeny*. In between those clades, the same neighboring classes (e.g. Bacilli and Cyanophycea) are found. The *cytochrome* c_1 phylogenetic reconstruction, although based on a smaller number of sequences, still follows the same overall pattern of distribution in relation to *Pseudomonadota* and *Actinomycetota taxa*.

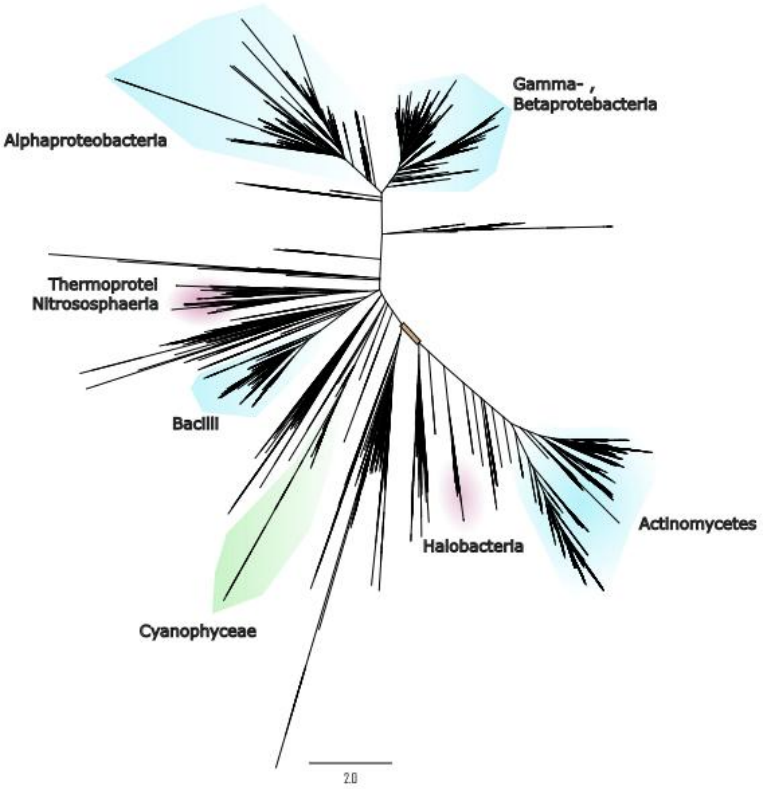

Figure 13: Phylogenetic reconstruction of the Rieske/ISP subunit, with a minimal ancestor deviation root (marked by the yellow square).

Rieske/ISP sequences were identified in both prokaryotic domains, (46 different phyla and 103 classes, including Candidatus taxa). Archaeal sequences belonging to the classes Halobacteria, Nitrospheria and Thermoprotei are highlighted in red, while bacterial lineages are marked by blue with cyanophycean highlighted in a lighter green. As can be observed, archaeal sequences do not form a monophyletic clade. Based on their scattered presence and phylogenetic position, it is more likely that they were acquired via HGT events, rather than being present in a common ancestor, with subsequent loss across many groups. The assignemt of the Rieske/ISP subunit to the b_6f

complex was based on the presence of cytochrome *f* and/or sequence position in the phylogeny since the models used in the analysis did not distinguish between the subunits used in both complexes.

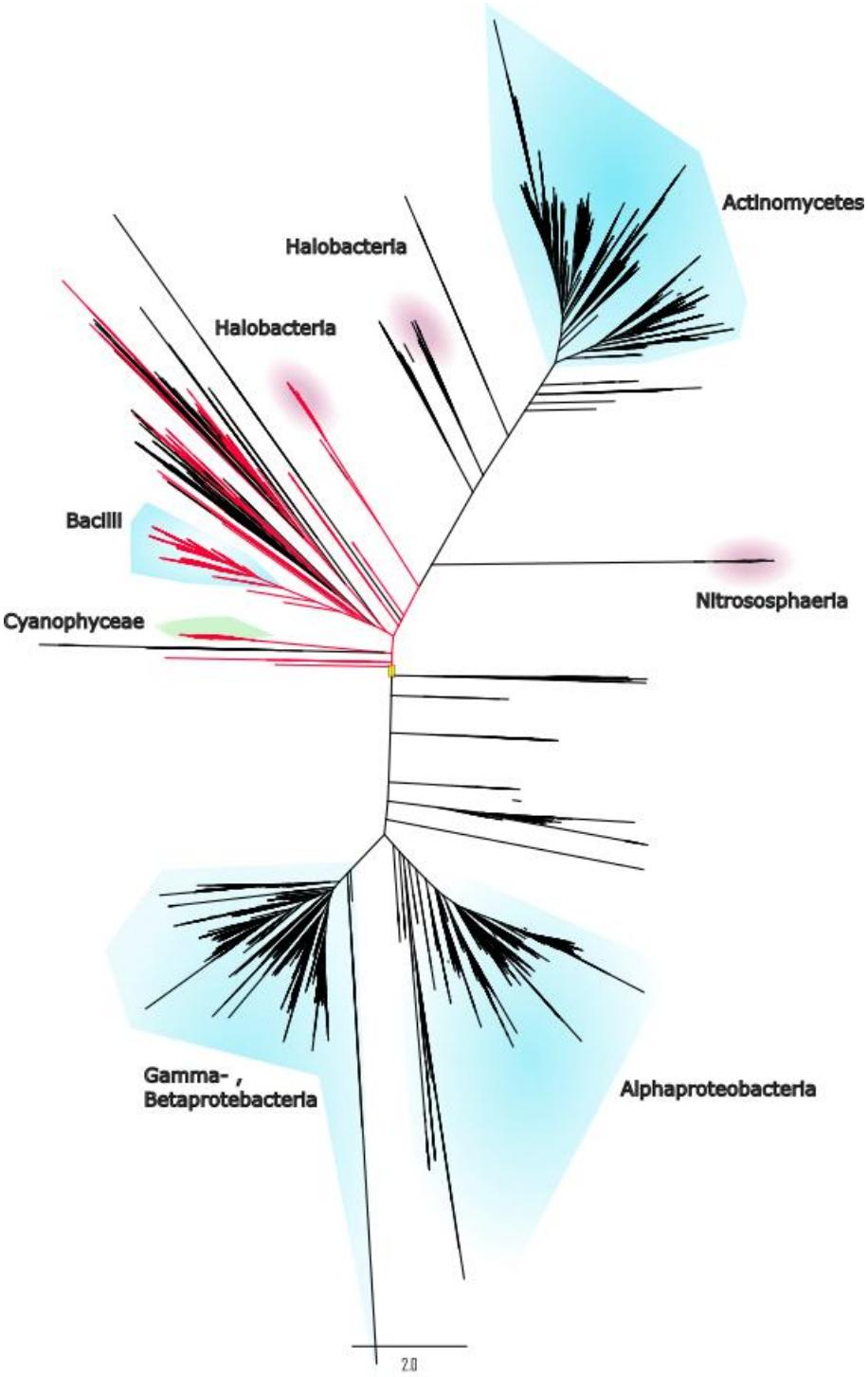

Figure 14: Phylogenetic reconstruction of the cytochrome b subunit, with a midpoint root (marked by the yellow square). The branches that are marked in red are the short version cytochrome b_6.

Figure 14 shows the reconstruction of the cytochrome *b* subunit where the colour scheme used in Figure 13 is used, with red lines indicating shorter cytochrome *b* versions, based on their sequence length. The sequences were the most diverse based on taxonomic results. 53 phyla and 113 classes, again including candidates, were reported. Archaeal

groups are present marked by soft red, but they do not form distinct separate clades but rather are embedded within the bacterial lineages. In this case, midpoint rooting was chosen since MAD rooting position the root within two archaeal known to have acquired their subunits via HGT events. Similar to the previous phylogeny, here, a predominantly bacterial signal is also observed. Moreover, the scattered position of the short version of cytochrome *b* across the phylogeny shows that not only they are non-monophyletic, but seem to have appeared multiple times via fission events and/or passed on via HGT events.

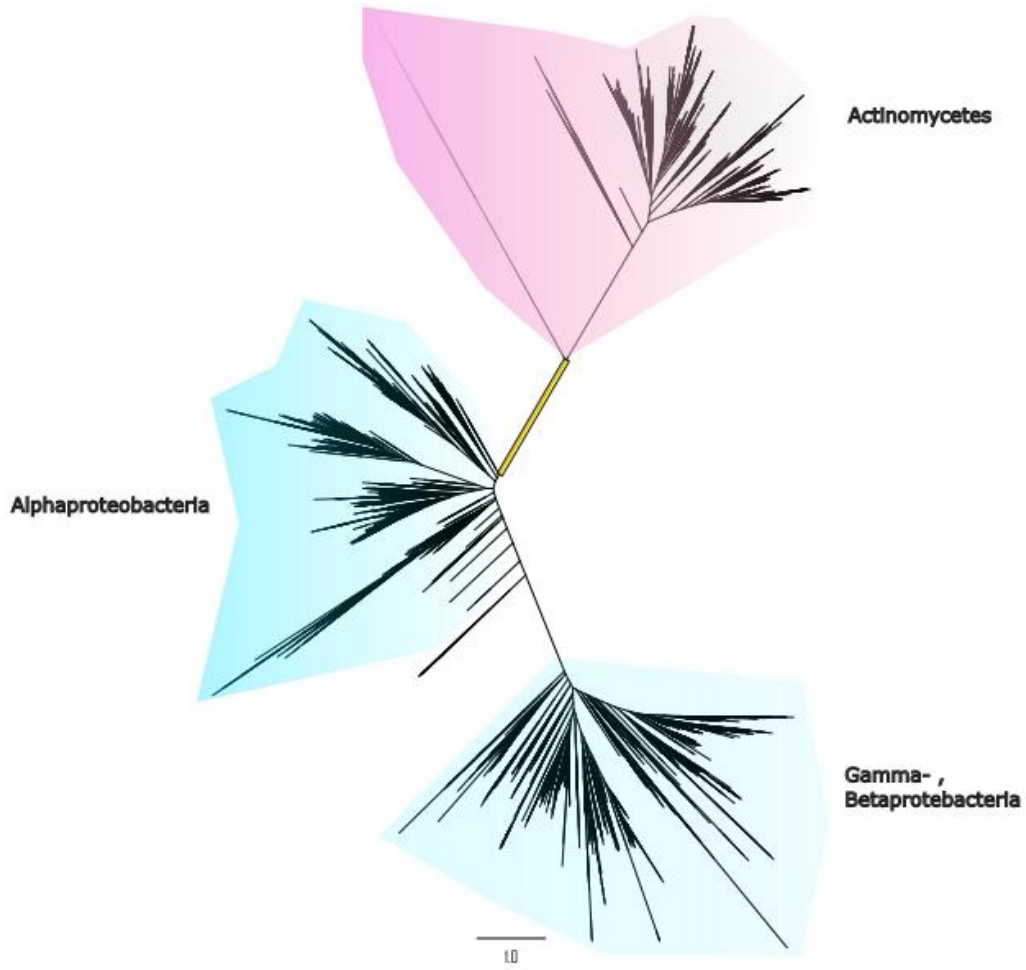

Figure 15 : Phylogenetic reconstruction of the cytochrome c_1 subunit, with a minimal ancestor deviation root (marked by the yellow square). Monoheme sequence are marked in blue and diheme sequences are marked in violet.

The cytochrome c_1 subunit is the one with the least number of sequences and with the smallest taxonomic diversity. In any case, also here, the topology of the major bacterial phyla is similar to that observed for cytochrome *b* and the Rieske/ISP subunit.

Revised Scenario

A revised evolutionary scenario can be proposed based on the results obtained in this project. Since no archaeal genomes contained all the subunits required to form a complete complex, it is likely that the Last Universal Common Ancestor (LUCA) did not possess either a bc_1 or a b_6f complex. The phylogenetic reconstructions indicate instead, that the bc_1 complex originated within Bacteria, where the ancestral complex probably consisted of a long cytochrome *b* subunit together with a Rieske/ISP subunit and a cytochrome *c*–type subunit.

Regarding cytochrome c_1, earlier work proposed that the ancestral state was a di-heme version of the protein, which later lost one heme-binding domain to give rise to the mono-heme cytochrome c_1[24]. However, this cannot be confirmed based on our phylogenetic reconstruction, although, the phylogeny presented in Figure 15 clearly separates mono-heme and di-heme cytochrome c_1 proteins into two distinct groups. To our knowledge

there is no known advantage of having a diheme versus a mono heme cytochrome c_1 in the complex.

Within the progenitor of Cyanobacteria, the ancestral bc_1 complex appears to have undergone substantial structural changes, giving rise to the b_6f complex, characterized by the shorter cytochrome b_6 and the presence of cytochrome *f* instead of cytochrome c_1. Our results do not support the ancestral presence of b_6f *over* bc_1 complex.

In Archaea, only partial sets of subunits were identified, typically including the Rieske/ISP and cytochrome b/b_6 components. These are likely the result of horizontal gene transfer (HGT) from bacterial donors. In some archaeal genomes, single subunits were present, but a complete complex could not be formed due to the absence of cytochrome *c* or cytochrome *f*. Whether Archaea possesses alternative subunits fulfilling similar functions, or whether these proteins serve a different purpose, remains to be determined experimentally. Moreover, examination of neighboring genes revealed no consistent patterns or shared genomic contexts, supporting the idea of independent acquisition events rather than inheritance from a common ancestor.

Implications for the Origin

Our results do not support the theory of a complete bc_1 complex in LUCA, based on the fact that no archaeal genome was capable of forming a full complex and that the distribution of archaeal genomes within the phylogenetic reconstruction is not monophyletic. What we would expect from a common ancestry is that we have a clear monophyletic group for *bacteria* and *archaea* but as they are dispersed within the bacterial lineages, we would reject the idea that it was already present in LUCA. Our interpretations are more consistent with the theory of Dibrova et al.[20,22] but here we differ that they suggest that the ancestral state resembles more the b_6f version of the complex.

Our findings contradict this as we have shown that the shorter version b_6 is also not monophyletic, arguing against its ancestral role.

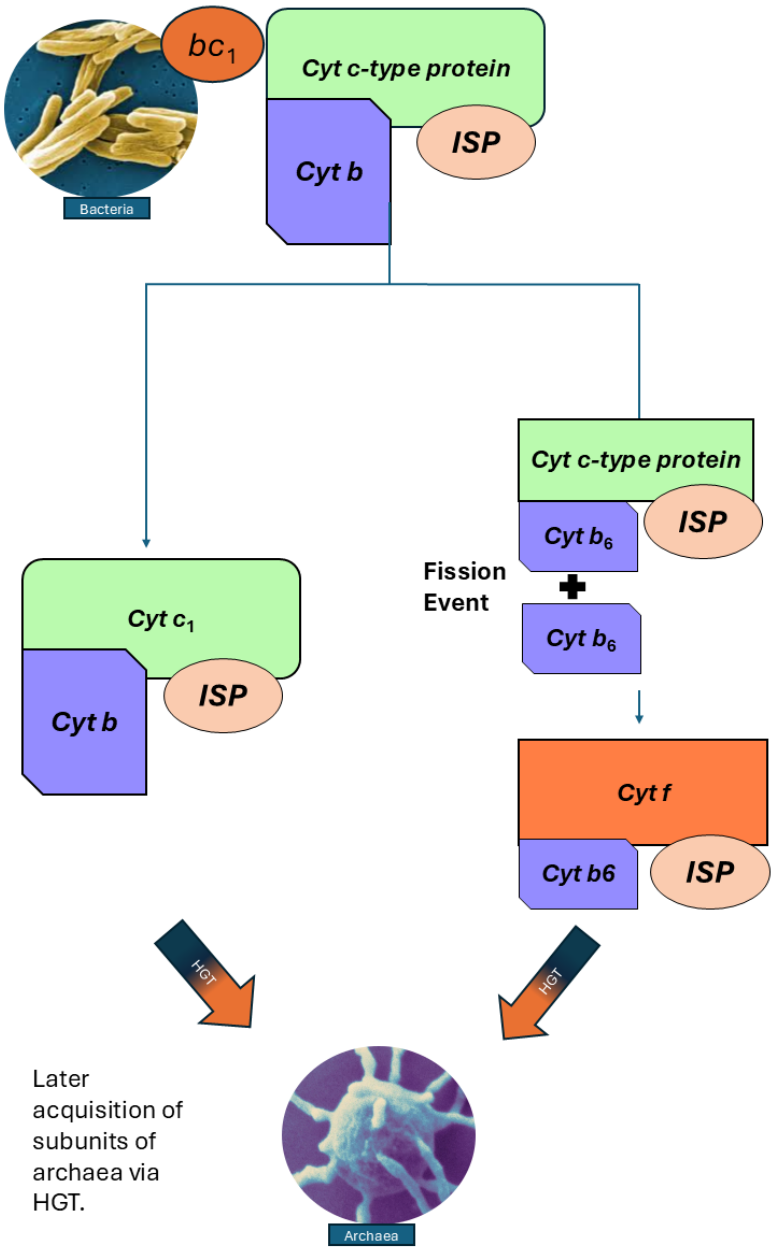

Figure 16: Revised scenario for the evolution of the bc1 complex. LUCA did not possess a bc_1 complex, and the ancestral state would contain the long version of the cytochrome b, which later would have undergone a fission event. This shorter version would have been used by the ancestor of the b_6f complex. HGT events are responsible for the presence of bc1 subunits in Archaea. Subunits that share a color are homologous protein complexes. Archaea are represented in the figure by Lokiarchaeum ossiferum [29], Bacteria are represented by Mycobacterium tuberculosis [30].

Conclusion

A large-scale comparative genomic analysis was performed to find out whether the bc1 complex was already present in LUCA or if it emerged later within bacterial lineages. A total number of 15.518 bacterial and archaeal genomes were analyzed with then KEGG based profile HMM, resulting in 53.432 sequences that were either associated to bc_1 or b_6f complex. Out of these, 22.184 sequences originated from a genome that contained all three catalytic subunits to form a full complex.

The most notable findings are in relation to the Archaea. Despite previous research purifying cytochrome *b* and *c* from Archaea[73] we were unsuccessful in identifying cytochrome c_1 in any archaeal genome. Further experimental validation is needed to clarify the function of the functional role of these Rieske/ISP and cytochrome *b* subunits identified in the archaeal genomes.

Future work could also incorporate additional sources of profile Hidden Markov Models, such as COG or eggNOG, to further improve functional annotation. Although multiple annotation resources were applied in this study; using KEGG models for initial identification, followed by PFAM domain annotation and a final validation step with the NCBI Conserved Domain Database; these analyses were performed sequentially rather than integrated into a unified consensus framework. By using a consensus annotation approach, it could potentially reduce model-specific biases especially for the archaeal sequences.

References

1. Ahmad, M., Wolberg, A. & Kahwaji, C. I. Biochemistry, Electron Transport Chain. in *StatPearls* (StatPearls Publishing, Treasure Island (FL), 2024).
2. Anoar, S., Woodling, N. S. & Niccoli, T. Mitochondria Dysfunction in Frontotemporal Dementia/Amyotrophic Lateral Sclerosis: Lessons From Drosophila Models. *Front. Neurosci.* **15**, (2021).
3. Scientific Image and Illustration Software | BioRender. https://www.biorender.com/.
4. Crofts, A. R. *et al.* The Q-Cycle Mechanism of the bc1 Complex: A Biologist's Perspective on Atomistic Studies. *J. Phys. Chem. B* **121**, 3701–3717 (2017).
5. Sarewicz, M. *et al.* Catalytic Reactions and Energy Conservation in the Cytochrome bc1 and b6f Complexes of Energy-Transducing Membranes. *Chem. Rev.* **121**, 2020–2108 (2021).
6. Xia, D. *et al.* Structural Analysis of Cytochrome bc1 Complexes: Implications to the Mechanism of Function. *Biochim Biophys Acta* **1827**, 1278–1294 (2013).
7. Conte, L. & Zara, V. The Rieske Iron-Sulfur Protein: Import and Assembly into the Cytochrome bc 1 Complex of Yeast Mitochondria. *Bioinorg Chem Appl* **2011**, 363941 (2011).
8. Hoffmeier, K. *Electron Transport Chain Complex III (Coenzyme Q : Cytochrome c — Oxidoreductase, Cytochrome Bc1 Complex)*. (2006).
9. Mitchell, P. The protonmotive Q cycle: A general formulation. *FEBS Letters* **59**, 137–139 (1975).

10. Crofts, A. R. *et al*. The Q-cycle reviewed: How well does a monomeric mechanism of the bc1 complex account for the function of a dimeric complex? *Biochimica et Biophysica Acta (BBA) - Bioenergetics* **1777**, 1001–1019 (2008).
11. CROFTS, A. R., MEINHARDT, S. W., JONES, K. R. & SNOZZI, M. THE ROLE OF THE QUINONE POOL IN THE CYCLIC ELECTRON-TRANSFER CHAIN OF RHODOPSEUDOMONAS SPHAEROIDES. *Biochim Biophys Acta* **723**, 202–218 (1983).
12. Crofts, A. R. The Cytochrome bc1 Complex: Function in the Context of Structure. *Annual Review of Physiology* **66**, 689–733 (2004).
13. Widger, W. R., Cramer, W. A., Herrmann, R. G. & Trebst, A. Sequence homology and structural similarity between cytochrome b of mitochondrial complex III and the chloroplast b6-f complex: position of the cytochrome b hemes in the membrane. *Proc Natl Acad Sci U S A* **81**, 674–678 (1984).
14. Refojo, P. N., Teixeira, M. & Pereira, M. M. The Alternative complex III: Properties and possible mechanisms for electron transfer and energy conservation. *Biochimica et Biophysica Acta (BBA) - Bioenergetics* **1817**, 1852–1859 (2012).
15. Refojo, P. N., Sousa, F. L., Teixeira, M. & Pereira, M. M. The alternative complex III: A different architecture using known building modules. *Biochimica et Biophysica Acta (BBA) - Bioenergetics* **1797**, 1869–1876 (2010).
16. Refojo, P. N., Teixeira, M. & Pereira, M. M. The alternative complex III of Rhodothermus marinus and its structural and functional association with caa3 oxygen reductase. *Biochim Biophys Acta* **1797**, 1477–1482 (2010).

17. Pereira, M. M., Refojo, P. N., Hreggvidsson, G. O., Hjorleifsdottir, S. & Teixeira, M. The alternative complex III from Rhodothermus marinus - a prototype of a new family of quinol:electron acceptor oxidoreductases. *FEBS Lett* **581**, 4831–4835 (2007).
18. Wu, W. *et al.* Crystal structure of the alternative complex III from the phototrophic bacterium Chloroflexus aurantiacus. *Structure* **33**, 29-38.e2 (2025).
19. Sun, C. *et al.* Structure of the alternative complex III in a supercomplex with cytochrome oxidase. *Nature* **557**, 123–126 (2018).
20. Dibrova, D. V., Cherepanov, D. A., Galperin, M. Y., Skulachev, V. P. & Mulkidjanian, A. Y. Evolution of cytochrome *bc* complexes: From membrane-anchored dehydrogenases of ancient bacteria to triggers of apoptosis in vertebrates. *Biochimica et Biophysica Acta (BBA) - Bioenergetics* **1827**, 1407–1427 (2013).
21. Wald, G. The origins of life. *Proceedings of the National Academy of Sciences* **52**, 595–611 (1964).
22. Dibrova, D. V., Shalaeva, D. N., Galperin, M. Y. & Mulkidjanian, A. Y. Emergence of cytochrome bc complexes in the context of photosynthesis. *Physiologia Plantarum* **161**, 150–170 (2017).
23. Kao, W.-C. & Hunte, C. The molecular evolution of the Qo motif. *Genome Biol Evol* **6**, 1894–1910 (2014).
24. Baymann, F., Lebrun, E. & Nitschke, W. Mitochondrial cytochrome c1 is a collapsed di-heme cytochrome. *Proc Natl Acad Sci U S A* **101**, 17737–17740 (2004).
25. Bertini, I., Cavallaro, G. & Rosato, A. Evolution of mitochondrial-type cytochrome *c* domains and of the protein machinery for their assembly. *Journal of Inorganic Biochemistry* **101**, 1798–1811 (2007).

26. Bowman, S. E. J. & Bren, K. L. The Chemistry and Biochemistry of Heme c: Functional Bases for Covalent Attachment. *Nat Prod Rep* **25**, 1118–1130 (2008).
27. Baymann, F., Schoepp-Cothenet, B., Lebrun, E., van Lis, R. & Nitschke, W. Phylogeny of Rieske/cytb Complexes with a Special Focus on the Haloarchaeal Enzymes. *Genome Biol Evol* **4**, 832–841 (2012).
28. Ducluzeau, A. L., Chenu, E., Capowiez, L. & Baymann, F. The Rieske/cytochrome *b* complex of Heliobacteria. *Biochimica et Biophysica Acta (BBA) - Bioenergetics* **1777**, 1140–1146 (2008).
29. Sokol, J. Primitive Asgard Cells Show Life on the Brink of Complexity. *Quanta Magazine* https://www.quantamagazine.org/primitive-asgard-cells-show-life-on-the-brink-of-complexity-20230411/ (2023).
30. Ray, B. Mycobacterium tuberculosis. *Encyclopædia Britannica* https://www.britannica.com/science/bacteria#/media/1/48203/218081 (2025).
31. Van Rossum, G. & Drake, F. L. *Python 3 Reference Manual*. (CreateSpace, Scotts Valley, CA, 2009).
32. ChatGPT. https://openai.com/chatgpt.
33. Perplexity. https://www.perplexity.ai/.
34. GitHub & OpenAI. GitHub Copilot. (2025).
35. team, T. pandas development. pandas-dev/pandas: Pandas. Zenodo https://doi.org/10.5281/zenodo.10697587 (2024).
36. Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585**, 357–362 (2020).
37. Sayers, E. W. *et al.* Database resources of the national center for biotechnology information. *Nucleic Acids Res* **50**, D20–D26 (2022).

38. Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering* **9**, 90–95 (2007).
39. Kapli, P., Yang, Z. & Telford, M. J. Phylogenetic tree building in the genomic age. *Nat Rev Genet* **21**, 428–444 (2020).
40. Linard, B. *et al.* Ten Years of Collaborative Progress in the Quest for Orthologs. *Molecular Biology and Evolution* **38**, 3033–3045 (2021).
41. Eddy, S. R. Profile hidden Markov models. *Bioinformatics* **14**, 755–763 (1998).
42. Aramaki, T. *et al.* KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* **36**, 2251–2252 (2020).
43. Eddy, S. R. HMMER User's Guide.
44. KEGG BRITE: KEGG Orthology (KO). https://www.genome.jp/brite/ko00001.
45. Yoon, B.-J. Hidden Markov Models and their Applications in Biological Sequence Analysis. *Curr Genomics* **10**, 402–415 (2009).
46. hmmscan vs. hmmsearch speed: the numerology ·. http://cryptogenomicon.org/hmmscan-vs-hmmsearch-speed-the-numerology.html.
47. Chang, A. *et al.* BRENDA, the ELIXIR core data resource in 2021: new developments and updates. *Nucleic Acids Research* **49**, D498–D508 (2021).
48. Information on EC 7.1.1.8 - quinol-cytochrome-c reductase - BRENDA Enzyme Database. https://www.brenda-enzymes.org/enzyme.php?ecno=7.1.1.8#SYNONYM.
49. BRENDA download - BRENDA Enzyme Database. https://www.brenda-enzymes.org/download.php.
50. Waskom, M. L. seaborn: statistical data visualization. *Journal of Open Source Software* **6**, 3021 (2021).

51. Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Research* **49**, D412–D419 (2021).
52. Sutherland, M. C. *et al.* In vitro reconstitution reveals major differences between human and bacterial cytochrome c synthases. *eLife* **10**, e64891.
53. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods* **18**, 366–368 (2021).
54. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* **48**, 443–453 (1970).
55. Van Dongen, S. Graph Clustering Via a Discrete Uncoupling Process. *SIAM J. Matrix Anal. Appl.* **30**, 121–141 (2008).
56. R Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, Vienna, Austria, 2024).
57. Wang, J. *et al.* The conserved domain database in 2023. *Nucleic Acids Res* **51**, D384–D388 (2023).
58. Gremme, G., Steinbiss, S. & Kurtz, S. GenomeTools: A Comprehensive Software Library for Efficient Processing of Structured Genome Annotations. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **10**, 645–656 (2013).
59. Blum, M. *et al.* The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res* **49**, D344–D354 (2021).
60. UniProt P0CY49. *UniProt* https://www.uniprot.org/uniprotkb/P0CY49/entry.
61. UniProt P83791. *UniProt* https://www.uniprot.org/uniprotkb/P83791/entry.
62. UniProt P83792. *UniProt* https://www.uniprot.org/uniprotkb/P83792/entry.
63. UniProt P26287. *UniProt* https://www.uniprot.org/uniprotkb/P26287/entry.

64. UniProt P23577. *UniProt* https://www.uniprot.org/uniprotkb/P23577/entry.

65. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology* **7**, 539 (2011).

66. trimAl: trimAl Main Page. https://vicfero.github.io/trimal/index.html#introduction_sec.

67. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution* **37**, 1530–1534 (2020).

68. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods* **14**, 587–589 (2017).

69. Tria, F. D. K., Landan, G. & Dagan, T. Phylogenetic rooting using minimal ancestor deviation. *Nat Ecol Evol* **1**, 1–7 (2017).

70. FigTree. http://tree.bio.ed.ac.uk/software/figtree/.

71. COGs - Clusters of Orthologous Groups. https://www.ncbi.nlm.nih.gov/research/cog-project/.

72. Huerta-Cepas, J. *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research* **47**, D309–D314 (2019).

73. Kletzin, A. *et al.* Cytochromes c in Archaea: distribution, maturation, cell architecture, and the special case of Ignicoccus hospitalis. *Front Microbiol* **6**, 439 (2015).

Supplements

Table 7: Total amount of pairs of DIAMOND sequences alignments that were performed.

Local Alignment (DIAMOND)	Count
Total Count	153,657,260
Total Unique Pairs	77,414,063
Unique Pairs After Filtering	67,373,047

Table 8: Total amount of pairs of NEEDLE sequence alignments that were performed.

Global Alignment (Needleman-Wunsch)	Count
Total Count	132,612,440
Total Unique Pairs	67,373,047
Unique Pairs After Filtering	40,542,745

Table 9: List of clusters that were used with their total amount of sequences and which subunit they represent.

Subunit	Clusters	Total number of sequences
ISP	2,5,8,10,15,27	10002
B	1,3,9,17,21,24	9409
C	4,6,7	6752

Table 10: List of the clusters with the number of corresponding sequences

Cluster	**Sequences**	**Cluster**	**Sequences**	**Cluster**	**Sequences**
Cluster 1	4844	Cluster 2	4820	Cluster 3	3091
Cluster 4	2559	Cluster 5	2280	Cluster 6	2236
Cluster 7	1958	Cluster 8	1725	Cluster 9	1277
Cluster 10	956	Cluster 11	906	Cluster 12	877
Cluster 13	390	Cluster 14	205	Cluster 15	197
Cluster 16	146	Cluster 17	112	Cluster 18	91
Cluster 19	57	Cluster 20	56	Cluster 21	54
Cluster 22	51	Cluster 23	49	Cluster 24	33
Cluster 25	30	Cluster 26	28	Cluster 27	24
Cluster 28	23	Cluster 29	22	Cluster 30	21
Cluster 31	19	Cluster 32	19	Cluster 33	18
Cluster 34	16	Cluster 35	13	Cluster 36	13
Cluster 37	10	Cluster 38	10	Cluster 39	9
Cluster 40	8	Cluster 41	8	Cluster 42	8
Cluster 43	7	Cluster 44	6	Cluster 45	6
Cluster 46	6	Cluster 47	6	Cluster 48	6
Cluster 49	6	Cluster 50	5	Cluster 51	5
Cluster 52	5	Cluster 53	4	Cluster 54	4
Cluster 55	4	Cluster 56	4	Cluster 57	3
Cluster 58	3	Cluster 59	3	Cluster 60	3
Cluster 61	2	Cluster 62	2	Cluster 63	2
Cluster 64	2	Cluster 65	2	Cluster 66	2
Cluster 67	2	Cluster 68	2	Cluster 69	2
Cluster 70	2	Cluster 71	2		