Unveiling the Terra Cognita of Sequence Spaces using Cartesian Projection of Asymmetric Distances

Alban Ramette[1,2]

Affiliations
[1] Institute for Infectious Diseases, University of Bern, Bern, Switzerland
[2] Multidisciplinary Center for Infectious Diseases, University of Bern, Bern, Switzerland

*Correspondence: Alban Ramette. Institute for Infectious Diseases, University of Bern, Friedbühlstrasse 25, Bern 3001, Switzerland. Tel +41 31 632 9540, Email alban.ramette@unibe.ch

Keywords: sequence analysis – multiple sequence alignment – genetic distances

Abstract

In the vast and ever-expanding realm of biological sequences, effectively visualizing relationships among numerous, potentially very long sequences is crucial. We introduce Cartesian Projections of Asymmetric Distances (CAPASYDIS), a novel approach designed to map and visualize the 'terra cognita'—the already explored regions—of a given sequence space. CAPASYDIS uniquely calculates asymmetric sequence distances, which account for the directional nature of sequence information, and then associates each distinct sequence to a unique location in a low-dimensional projection space, called a "seqverse". The approach is by design computationally scalable, and allows a standardized exploration of the seqverses of massive datasets, comprising hundreds to millions of potentially very long sequences. The approach is illustrated by analyzing the seqverses of rRNA sequences from the three domains of life. We show that the seqverses of Bacteria, Archaea and Eukaryota are not only spatially distinct, but also differ in shapes, sizes and patterns of variation around their centroids. *asymdist* values retain most of the taxonomic signal present in the sequences, when analyzed at different taxonomic levels from the domain to the single base difference levels. Overall, CAPASYDIS represents a novel mapping strategy that helps define the shape and boundaries of seqverses, allowing their exploration in a reproducible and scalable manner.

Introduction

Genetic diversity analysis and phylogenetic reconstruction are cornerstones of biological research. They provide insights into evolutionary relationships, ecological dynamics, and the intricate workings of life itself. The advent of next-generation sequencing (NGS) has unveiled the vast extent and complexity of genetic diversity, revealing that even few grams of soil (Howe *et al.*, 2014), milliliters of sea water (Singer *et al.*, 2019), or square centimeters of human skin (Byrd, Belkaid and Segre, 2018), can harbor a staggering array of lifeforms. This unprecedented availability of sequence data presents opportunities, but also challenges to analyze and visualize the large amount of information (Yin *et al.*, 2017). While traditional methods like distance matrix calculations and phylogenetic tree reconstruction offer valuable perspectives to represent the extent of the diversity in sequenced data, they can struggle to efficiently process and visualize the high-dimensional nature of sequence space (Cavalli-Sforza and Edwards, 1967; Roch, 2010), especially with increasingly larger datasets.

A significant challenge is, first, the curse of dimensionality, which makes it computationally infeasible to compare all possible sequence combinations. For a DNA sequence of length n, there are 4^n possible sequence variations. To create a definitive distance matrix that maps the relationships between all these variations, one would have to calculate a matrix of size 4^{2n}. As sequence length grows, this number expands exponentially, quickly exceeding the computational limits of even the most powerful supercomputers. Consequently, a complete, universal map of all possible evolutionary relationships of even a 1,000-base sequence is impossible to create, as this would correspond approximately to 10^{1204} possible sequences. In comparison, the estimated total number of stars in an inflationary universe (observed and unobserved) would be in the range of 10^{100} (Totani, 2020). This challenge forces researchers to rely on methods that only compare the sequences actually observed in their datasets, rather than representing the entire potential sequence space. Second, because it is impossible to compute a comprehensive distance matrix for all possible sequences, constructing evolutionary trees (dendrograms or phylograms) becomes a challenging task. Even for a modest number of observed sequences, the number of possible tree topologies grows at a rate that is superexponential. It means that number increases far faster than an exponential function: For example, the number of possible unrooted binary trees for just 10 sequences is over two millions, and for 50 sequences, it is a number with 74 digits (Cavalli-Sforza and Edwards, 1967; Felsenstein, 1978). This consideration makes it impractical to search for a single, "perfect" tree. As a result, bioinformatic methods for tree construction must rely on heuristics, which are algorithms that use assumptions to find a reasonably good, though not necessarily optimal, solution. Approaches, such as Neighbor-Joining or Maximum Likelihood, iteratively build a tree that best fits the input data, but they do not guarantee that the absolute best tree has been found. This reliance on heuristics introduces a level of uncertainty and subjectivity into the analysis. Third, beyond the computational difficulty of building a meaningful tree, a key limitation lies in how dendrograms and phylograms are visualized. These diagrams represent tree topology (the branching pattern) and branch

lengths (often proportional to evolutionary distance), but they do not convey any information about the absolute spatial positions of the branches or leaves in the tree. This means that a tree can be visually transformed by simply rotating a branch at any node, and the underlying evolutionary relationships remain exactly the same. This non-uniqueness can create a misleading visual interpretation, making it difficult to compare different tree plots, especially when tree size grows larger. Thus, while phylograms are powerful for showing hierarchical relationships, their visual arrangement does not contain any unique or definitive information.

To address these limitations, we introduce CAPASYDIS (Cartesian Projections of Asymmetric Distances), a framework designed to map and visualize the "terra cognita" of a given sequence space. CAPASYDIS embraces the inherent asymmetry in reading direction of a pair of aligned sequences, and acknowledges that, even if the same number of changes between two sequences exist, if those changes occur at different sequence positions, they should have *different impacts on the resulting distances* between those two sequences. Starting from a multiple sequence alignment, the novel concept associates a distinct distance value to each unique query sequence relative to a given reference sequence. Then using those unique vectors, a reduced dimensional space can be built by plotting the sequences using those unique coordinates into a lower-dimensional space. By employing Cartesian projections, sequences are mapped onto a visualizable space with predefined, fixed axes, enabling reproductible exploration and analyses. In this study, the CAPASYDIS methodology is presented, including the introduction of the novel asymmetric distance metric, *asymdist*, that differentially accounts for differing positions along the compared sequences. Possible applications of CAPASYDIS are illustrated on a large dataset (rRNA gene sequences) where the known diversity of the three domains of life is investigated at different taxonomic levels.

Methods

Data set. The file "SILVA_138.2_SSURef_NR99_tax_silva_full_align_trunc.fasta.gz" was downloaded from the SILVA rRNA database project (https://www.arb-silva.de; (Quast *et al.*, 2013; Glöckner *et al.*, 2017)) on 2024-07-11. The dataset is based on the full SSU Ref 138.2 dataset and include a total of 510,508 rRNA sequences from the three domains of life. The CAPASYDIS processed data used for the NR99 analysis is available at https://doi.org/10.5281/zenodo.17055026.

Implementation. The initial proof-of-concept was implemented with the R language (version 4.4.1; https://www.r-project.org/) to develop the algorithm using small datasets. To allow the pre-processing of large Multiple Sequence Alignments (MSA), further computational scripts were developed in the *go* language (aka "Golang"; https://go.dev/; version 1.24.5) based on the *biogo* library (version v1.0.1; (Kortschak *et al.*, 2017)). Further analyses of the csv files produced by CAPASYDIS *go* scripts were processed with the R library *dplyr* (version 1.1.4; (Roussel *et al.*, 2025)), and visualized using the R libraries

ggplot2 (version 3.5.2; (Wickham, 2016)), *plotly* (version 4.10.4; (Sievert, 2020)), *grid* (version 4.4.1; base R), *geometry* (version 0.5.2; (Roussel *et al.*, 2025)), *FNN* (version 1.1.4.1; https://cran.r-project.org/web/packages/FNN/), *patchwork* (version 1.3.0; https://cran.r-project.org/web/packages/patchwork/index.html), *ggalign* (version 1.0.2; (Peng and Wang, 2025)), and *forcats* (version 1.0.0; https://cran.r-project.org/web/packages/forcats/index.html). In addition to R, interactive data visualization was also implemented with the JavaScript *D3.js* (version 7.9.0; https://d3js.org/) and *Three.js* (version 0.179.1; https://threejs.org/) libraries, which enable the rendering of large objects directly in the web browser of the user. Gemini 2.5 Flash and Pro were used throughout the project to help develop and correct code snippets in the R, Golang and JavaScript languages.

All CAPASYDIS analyses on NR99 datasets were performed using an AlmaLinux 8.10 compute server (AMD EPYC-Rome processors 2.26 GHz; 491 GB of RAM). Further data analyses in R or with the JavaScript libraries were done on a standard Windows office laptop (Intel Core Ultra 5 processor 135U with 1.60 GHz, 16 GB RAM, 8 GB Intel Graphics GPU).

Code availability. R, Golang and JavaScript scripts are available at https://github.com/RametteLab/CAPASYDIS. The repository also includes the workflow and scripts used to analyze the NR99 dataset presented in the study.

Results

CAPASYDIS concepts. The goal of CAPASYDIS is to associate a query sequence in a MSA dataset to a unique Euclidean coordinate. Given a sequence alignment of length *n* bases between a candidate sequence S and a reference sequence R, a *score x*i for the base at position i of sequence S is first calculated as:

$$xi = i + delta \quad (1)$$

where $delta$ is null, except if the base at position i is different between S and R. In the latter case, the *delta* value is either a small constant value (e.g. 0.01) as compared to the magnitude of the indices, as in the simple example (**Fig. 1A**), or a unique value associated with a specific base pair combination (see section 2 below; **Table 1**). Second, the individual scores are square-root transformed and then summed up across all the sequence positions:

$$SumSqrtScore(S|R) = f(x1, x2, \ldots, xn) = \sum_{i}^{n} \sqrt{x_i} \quad (2)$$

where xi represents the *score* at position i of the sequence S as defined in Equation 1. For example, a one-base difference at position 2 between S1 and R leads to the following computations for S1: SumSqrtScore (S1|R)= $\sqrt{1}$ + $\sqrt{2.01}$ + $\sqrt{3}$ + … + $\sqrt{10}$ = 22.472. For R, we obtain: SumSqrtScore (R|R)= $\sqrt{1}$ + $\sqrt{2}$ + $\sqrt{3}$ + … + $\sqrt{10}$ = 22.468 (**Fig. 1B**).

The function $SumSqrtScore$ (Equation 2) produces a numeric sequence that diverges indefinitely as *n* increases through all positive integers. There is no simple, concise formula

to directly calculate the nth term of the sequence without performing the summation of square roots (no closed-form expression). Importantly, the gaps between successive *n* terms increase, which means that terms appearing later will create a larger effect on the overall distance for a given sequence. This is because the difference between consecutive terms [f(n + 1) - f(n)] is equal to $\sqrt{n+1}$, and as *n* increases, so does the square root, meaning the gaps between terms get wider and wider. Third, the output of Equation 2 is then re-scaled to a convenient interval, e.g. [0,1]:

$$asymdist(S|R) = \frac{(SumSqrtScore(S|R) - absMin)}{(absMax - absMin)} \quad (3)$$

where $SumSqrtScore(S|R)$ is calculated in Equation 2, $absMin$ is the smallest value obtained if the S and R are identical sequences, and $absMax$ is the largest value that is obtained if the sequence S differs maximally at all positions with R. Equation 3 provides a quantity that we name $asymdist$ to reflect the asymmetric nature of the results. Indeed, the calculation takes into account the positions at which base modifications occurred along the alignment, and provides a different result as compared to the usual count of base differences (**Fig. 1B**). Noticeably, changes that occur early in the sequence positions (i.e. towards the 5′ end, such as S1) of the reference sequence have more weight than changes occurring later (i.e. towards the 3' end, such as S3), and contribute larger distances (**Fig. 1C**).

The asymmetric nature of the proposed metric can be illustrated graphically using a set of 100 aligned sequences, which differ by one base only as compared to the reference sequence. These sequences were ordered by the position of the modification, from left to right (**Fig. 2A,B)**. Early changes at the beginning of the sequences have much larger impact on the *asymdist* values than changes occurring downstream, e.g. after base 50. By design, a distinct *asymdist* value is associated with each unique sequence provided that the sequence differs from the reference in a unique manner. In addition, when compared to classical phylogenetic distances, discrimination is found among all sequences that differ from the reference sequence by only one single variation (**Fig. 2C)**: On this set of 100 distinct sequences, the JC69 model produces only one constant value for all sequences, which is not surprising given that this model assumes that all changes of a base by another one have the same probability (Jukes and Cantor, 1969). This probability is the same for all sites along the DNA sequence, so it does not reflect where the change occurred in the sequences. A similar result is obtained if one relaxes the assumption of equal base frequencies, by using e.g. the F81 model, which is a generalized Jukes–Cantor model (Felsenstein, 1981). Alternatively, one may also try probabilistic models for nucleotide substitution such as the BH87 model (Barry and Hartigan, 1987), which accounts for the observed proportions of changes among the four bases. This latter distance is not symmetric (**Fig. 2D)**, which results yet in only two distance classes (0.00125, 0.00083) for all 100 distinct sequences. A similar result is obtained when using the TN93 model (Tamura and Nei, 1993), which assumes distinct rates for both kinds of transition (A ↔ G versus C ↔ T), and transversions: Here, the base frequencies are not assumed to be equal and are estimated from the data. The TN93 model produces three classes of values corresponding

to distances of 0.01008, 0.01021, and 0.01021 (**Fig. 2E)**. Overall, the *asymdist* calculation approach allows the set of 100 query sequences to be associated with definitive and unique coordinates using only one dimension.

Using distinct *delta* values. With a constant $delta$ value, different types of changes at a given position would not be distinguishable from one another, and would produce *asymdist* coordinate values that are thus confounded. Therefore, a matrix containing distinct *delta* values for all possible mismatches, including those involving gaps in the reference sequence, could thus be proposed (**Table 1)**, where a unique numerical weight is assigned to each specific change, in order to influence the final score of the sequence.

Table 1. A possible coding of case-specific *delta* values.

S|R change	*delta* value
Transition	
A ↔ G	0.01
C ↔ T	0.02
Transversion	
A ↔ C	0.06
G ↔ T	0.07
A ↔ T	0.08
G ↔ C	0.09
Unknown	
- ↔ A	0.11
- ↔ T	0.12
- ↔ G	0.13
- ↔ C	0.14

When using case-specific *delta* values, the uniqueness of the representation of each single sequence is maintained, but also that a different, more granular pattern is obtained for the set of 100 sequences as compared to using a unique *delta* value for all changes (**Fig. 2B**). Although this coding of case-specific *delta* values is arbitrary, we could justify this choice by stating that delta values should be relative small to the smallest index values (*delta* << 1), and by assuming that transitions (A ↔ G, C ↔ T) involve bases of similar shape, and are thus less likely to result in amino acid substitutions (due to "wobble"), and are therefore more likely to persist as "silent substitutions", and could thus be more likely to occur. On the contrary, transversions (A ↔ C, G ↔ T, A ↔ T, G ↔ C) involve exchange of one-ring and two-ring structures, and as such, transversions are generated at lower frequency than transition mutations. This is one way of coding the *delta* values and probably other options, with may reflect biological knowledge, could also be considered, as long as the magnitude of those values is small as compared to the smallest base position index value. This choice enables the base modifications to be associated with a

value whose magnitude is smaller than that of the case when no difference to the reference sequence is observed.

A possible workflow using the CAPASYDIS. In essence, the concept of CAPASYDIS is to produce one coordinate axis per chosen reference sequence on which query sequences are positioned in a unique manner given the uniqueness of their *asymdist* coordinates (see section 2). It is thus essential that the starting dataset consists of a set of high-quality sequences, for which the aligned positions are meaningful. As such, a series of preparatory steps may be needed before applying the CAPASYDIS approach to a MSA in order to build a 2D or 3D visualization of the sequence space, which we defined as "seqverse" (**Fig. 3**). We start with a MSA of rows of sequences by columns of aligned bases in which A, U, T, G, C, - (dash, or alignment gap) and "." (missing base) are considered valid information. In step **a**) the MSA is vertically truncated to keep only the positions of the MSA that are not missing or not a constant gap for a given *k* proportion of bases (step a_1). For instance, for $k = 0.9$, the MSA is truncated at specific starting and ending positions so that 90% of the sequences starting in the new MSA contain at least *k* bases that are not missing bases ("."). The same approach is applied to the last position of the alignment. Positions consisting of "-" or "." only, as well as duplicated sequences, are removed from the MSA in a second stage (a_2). The headers of the remaining FASTA sequences are appended with the count of identical sequences found in the MSA. In step **b**) CAPASYDIS axes are computed for all the sequences in the MSA based on the indices of the chosen reference sequences, and the 2D (or 3D) coordinates are merged into a single csv file. In step **c**) specific color labels can optionally be added to sequence features (e.g. color "red" for a given species or taxonomic level such as the domain "Archaea"). In step **d**) the coordinate file of the seqverse can be displayed in one to three dimensions using dynamic visualizations, which allow the interactive analysis of very large number of points and their patterns. The user can thus identify points interactively, delineate grouping structures (hulls, contours), and hover to see details, zoom to explore fine-grained topics, or retrieve the information from specific sequences. Optionally, identified sequences or patterns can be retrieved from the original MSA (step **e**). Those steps are illustrated with an example in the next section.

The rRNA seqverses of the three domains of life. The CAPASYDIS workflow was applied to ribosomal RNA sequences obtained from the SILVA release 138 SSU Ref NR99, which includes 510,495 RNA sequences with 20,389 (3.99%) sequences classified as Archaea, 58,940 (11.54%) as Eukaryota, and 431,166 (84.46%) as Bacteria. This large MSA consists of 50,000 aligned nucleotide positions, accommodating varying structures and regions within the 16S and 18S rRNA gene sequences (total information content of 25,524,750,000 bases). The script *truncAte* was first applied with a threshold *k* of 0.90 that select positions in the truncated MSA to be kept if 90% of the bases at those positions are not missing information or alignment gaps. The procedure also removes sequences with unknown ("N") bases or wobble bases. The analysis was completed in 1h16 min when using 60 cores and produced a new MSA of dimensions 387,633 sequences by 29,932 aligned base positions, so keeping effectively around 76% of the initial sequences, and 60% of the initial

aligned base positions. The resulting MSA was further deduplicated and degapped with *deduplicAteSeq*, resulting in 331,663 unique sequences. This step took about 10 min using 30 cores. This deduplicated MSA now encompassed 6,031 sequences for Archaea (1.8%), 291,195 sequences for Bacteria (87.8%) and 34,437 sequences for Eukaryota (10.4%), resulting in a total dataset size of 9,927,336,916 bases, with average sequence lengths of 1,278 bases (minimum 842, maximum 2,787) for Archaea, 1,313 (1,036, 2,642) bases for Bacteria, and 1,605 (1,087, 3,232) bases for Eukaryota.

Escherichia coli O157:H7 (GenBank Accession number: AB035920; identified as AB035920.964.2505 in the MSA) was chosen as first reference for the CAPASYDIS axes, as its sequence was the most abundant among the *Escherichia-Shigella* sequences with a total of 379 sequences in the cleaned MSA. The function *build_axes*() provides several parameters such as "*-stat*" to retrieve sequence identity of the most distant sequences, the overall mean, standard deviation and maximum *asymdist* values, based on the input of the index of the chosen reference sequence. When applied to our example using case-specific *delta* values (**Table 1**), the most distant sequence to the 16S rRNA gene of *Escherichia coli* O157:H7 was the 18S rRNA gene of the marine eukaryotic cellular organism *Heterostegina depressa* (GenBank accession number AJ879131.1; NR99 accession AJ879131.1.3425). The latter is classified in the Phylum Foraminifera, and produced the largest *asymdist* value of 0.0983304. The second CAPASYDIS axis for the NR99 MSA was then calculated using as reference the *H. depressa* sequence in the MSA. The *build_axes*() function identified the 16S rRNA gene sequence of the archaeon *Pyrobaculum* sp. (GenBank accession number AB302407.1; NR99 accession AB302407.1.2962) at an *asymdist* value of 0.1298366, as the most distantly related to the Eukaryotic reference. Finally, the *asymdist* values with references to *E. coli* (Bacteria; axis "x"), *H. depressa* (Eukaryota; axis "y"), and *Pyrobaculum* (Archaea; axis "z") were merged into a comma-separated-value (csv) file, with a simple structure of the type "x, y, z, label" where (x, y, z) represents the *asymdist* coordinates in different dimensions, and the "label" describes the taxonomic information string of the sequences. *asymdist* calculation via *build_axes*() for one reference sequence in the MSA took about 1-2 min using 60 cores in parallel. The uniqueness of the coordinates of each sequence was obtained with a decimal point precision of e-10. The choice of the precision level (tunable parameter in the *build_axes* script) depends on the intrinsic properties of the MSA, such as dimensions and complexity. Smaller and simpler MSA would require less precision to obtain unique point coordinates for each sequence in the seqverse. Here, we determine empirically the needed precision by defining the number of decimals needed to differentiate a sequence with only one change at the most proximal index position and considering a change associated with the smallest possible *delta* value.

The taxonomic information provided with the NR99 was further parsed to remove inconsistencies and missing information using the script *SILVA_go*, which is a Golang script that removes erroneous terms (e.g. habitat or sample descriptors) in the taxonomic fields. After removing taxonomic inconsistencies, the data set labels contained 14 phyla and 28 classes for Archaea, 92 phyla and 185 classes for Bacteria, while Eukaryota consisted of 10 phyla and 21 classes. Taxonomic levels in the coordinate file were further associated with

specific color labels using either the R programming language or Golang (colorCSV). The resulting labelled coordinate file was finally rendered with the R library *plotly* (**Fig. 4**) or via a suite of HTML interactive pages using the JavaScript 3D (Three.js) or 2D (D3.js) libraries, which are suitable for rendering very high-dimensional data. Several exemplary JavaScript scripts are provided in the package repository.

The rRNA gene seqverses of the three domains of life fall into distinct, non-overlapping regions (**Fig. 4A**), and spanned different sizes and shapes (**Fig. 4B-J**). The 3D projecton of the seqverses indicated that the Eukaryota seqverse appears rather flat, that of the Archaea compact and conic, while the bacterial seqverse seems to display an extensive "wing" shape (**Movies 1, 2**). Two-dimensional plots confirmed the distinct landscapes (**Fig. 4B-G**) and density profiles (**Fig. 4H-J**) of the 16S rRNA gene seqverses of the three domains of life. The seqverses revealed distinct multi-modal density distributions in the 2D coordinate space, as seen for the archaeal seqverse (**Fig. 4H**) or bacterial seqverse (**Fig. 4I**), around their respective domain centroids. The coordinates of the centroids for each domain were calculated as the median coordinate on each dimension using all sequences whose information at the phylum level was available (n= 331,105 cases). As expected the most distant sequences were those of *Pyrobaculum* (Archaea) and Phylum Foraminifera (Eukaryota), which contributed to the large spread of the archaeal and eukaryotic seqverses away from their centroids, respectively. The centroid (coordinates 0.017664, 0.094707, 0.046474) of the Archaea seqverse had the sequence of Thermoplasmata (identified as KP091008.1.1423 in the MSA) as closest sequence (distance of 0.0002192) (**Movie 3**; **Fig. 4E, H**). For bacteria, the closest sequence to the centroid (coordinates x=0.010409, y=0.097673, z=0.051626) was the sequence of the uncultured Actinobacteria of the genus *Blastococcus* (identified as KC554984.1.1516) at a distance of 3.734e-05 (**Movie 4**; **Fig. 4F, I**). For Eukaryota, the closest sequence to the centroid (coordinates 0.028267, 0.087945, 0.063611) belonged to the Animalia *Bilateria, Cnidaria, Ctenophora* (BCP; KY077283.1.1814) at a distance of 4.294182e-05 (**Movie 5**; **Fig. 4G, J**). The average distance of all the sequences to their domain centroids was 0.00209 ± standard deviation 0.001867 for Archaea, 0.00395 ±0.002892 for Bacteria, and 0.00310± 0.004643 for Eukaryota, while the distances between domain centroids were on average 0.01808 ± 0.007626 (**Fig. 5A-B**). Therefore, the between-distance was on average 9.7, 4.6, and 5.80 times larger than the respective within-domain distances, underlying the sequence cohesiveness around their domain centroids.

The distribution of distances to centroids for the domain Bacteria displayed a bimodal distribution, which was not evidenced for the other domains (**Fig. 5D**). To investigate this pattern, bacterial sequences were then divided into two groups, based on whether their distances to the centroid were smaller or larger than the observed distance cutoff of 0.0033, which was found as the mid distance between the two modes of the distribution (**Fig. S1**). Group 1 and group 2 thus included 134,978 (46.4%) and 155,744 (53.6%) bacterial sequences, respectively. To understand whether these groups were associated with particular taxonomic preferences, detailed analysis at the Phylum level for each group was undertaken: A partitioning at the phylum level was observed for 22 out of

51 phyla (43.1%). For those phyla, >90% of their sequences were classified into one specific group, with group 1 with 21 cases and group 2 with 1 case corresponding to Pseudomonadota. The latter case corresponded to 6,545 (8.0%) and 75,748 (92.0%) sequences belonging to each group, respectively, suggesting a clear trend favoring group 2, i.e. the more diverging group of sequences to their domain centroid (**Table S1A**). Similarly, 41/105 (39.0%) of the bacterial Classes were associated with >90% of their sequences belonging to a specific group (group 1 with 37 cases and group 2 with 5 cases; **Table S1B**). Those patterns suggest that few bacterial Phyla and Classes do have a tendency to systematically diverge from their domain centroid more than others. This observation is also supported by the 2D analyses, which indicated that the point density distribution of the bacterial domain was not uniformly distributed, as some denser areas were evidenced, much like an archipelago of small "islands" around a larger island where the domain centroid was located (**Fig. 4F, I**). In the 2D plots (**Fig. 4B-D**), the area of the Archaea seqverse was also 5.6 and 18.1 times smaller than that of the Bacteria and Eukaryota, respectively (area for Archaea: 0.00022, Bacteria: 0.00124, Eukaryota: 0.00401 asymdist^2), suggesting a smaller diversity space for Archaea. Differences in domain shape were mostly driven by the presence of few "atypical" sequences, whose effects on the distributions were mostly seen for the Archaea (**Fig. 4E**) and Eukaryota (**Fig. 4G**) domains.

Atypical sequences in each domain were then arbitrarily defined as sequences whose distances to their domain centroid were greater than the respective 99% percentile (q99) of all the 3D distance values for this domain. The q99 values of 7.8 $\times 10^{-03}$, 11.4 e^{-03}, and 21.8 e^{-03} were obtained for the domains Archaea, Bacteria, and Eukaryota, respectively, which mirror their observed spreads and shapes (**Fig. 5D**), and total 3D variance of 7.40 e-06, 2.36 e-05, and of 2.84e-05, respectively. At these q99 values, 61, 2,912 and 344 atypically distant sequences were observed for each domain, respectively. For the archaeal domain with 61 outliers, the phylum Thermoproteota was represented with classes Thermoproteia (n=39 sequences; 63.9%), and Bathyarchaeia (3 sequences; 4.9%), the phylum Aenigmarchaeota with the class "Deep Sea Euryarchaeotic Group" (DSEG) (8 sequences; 13.1%), and the phylum Thermoplasmatota with class Thermoplasmata (2 sequences; 3.3%). For the domain Bacteria, with 2,912 sequences with distances above q99, the top three phyla were phylum Bacillota with classes Clostridia (787 sequences; 27%), Bacilli (269 sequences; 9.3%), Negativicutes (104 sequences; 3.6%), Desulfotomaculia (101 sequences; 3.5%), and Desulfitobacteriia (65 sequences; 2.2%), phylum Pseudomonadota with classes Alphaproteobacteria (602 sequences; 20.7%), Gammaproteobacteria (46 sequences; 1.6%), and phylum Patescibacteria with classes Saccharimonadia (299 sequences; 10.3%), Parcubacteria (59 sequences; 2%) and Microgenomatia (57 sequences; 2%). For the domain Eukaryota with 345 outlier sequences, the identified phyla were Discoba with class Discicristata (143 sequences; 41.6%), phylum Amorphea with classes Obazoa (99 sequences; 28.8%) and Amoebozoa (18 sequences; 5.2%) and phylum Archaeplastida with classes Chloroplastida (45 sequences; 12.8%) and Rhodophyceae (26 sequences; 7.6%), and phylum SAR with class Rhizaria (13 sequences; 3.8%).

Given that *E. coli* was chosen as the reference for the rRNA gene seqverses, a pertinent question was whether the positions of the sequences pertaining to *E. coli* and related species in the obtained seqverse were associated with the taxonomic labels of these sequences, and if so, at which levels of taxonomic resolution. We thus considered the bacterial Family Enterobacteriaceae to which *E. coli* belongs, and kept the top 15 genera whose sequence abundances in the dataset consisted of at least 10 sequences. The clustering of the sequences by genus indicated that along the first two dimensions of the CAPASYDIS projection axes a continuum of diversity was observed (**Fig. 6**). Individual projection of each single genus on each axis (**Fig. S2**) clearly revealed that most Enterobacteriaceae genera, including *Escherichia*, *Enterobacter*, *Klebsiella*, *Kluyvera*, *Salmonella*, *Shigella,* of the Enterobacteriaceae displayed very peaky distributions on each projection axis, suggesting well-defined seqverses. On the third axis, however, more overlap between the coordinates of different genera was observed. The lack of resolution observed for *E. coli* and relatives on the third axis could be expected, given that this axis was built using the most distant sequences to *E. coli*, which was the Eukaryotic reference (*H. depressa*), and then taking the most distant sequence to the latter, which was found with the Archaeal sequence *Pyrobaculum*. Thus, the third axis is expected to be distantly related to the *E. coli* sequences in general.

Upon closer examination of the *E. coli* sequences, which constituted 948 of the 4,343 total Enterobacteriaceae sequences (**Fig. 7A**), a main "island" of diversity was found with two main peaks (**Fig. 7B**), which also matched the observed split in the dendrogram that could be obtained using classical phylogenetic representation of 200 sequences differing by only one difference from each other (**Fig. 7C**). Both the classical rectangular dendrogram and the unrooted tree representations (**Fig. 7D**) confirmed the existence of two main clusters. We compared the number of base differences (Hamming distance) between *E. coli* sequences, and identified 655 pairs involved a single nucleotide difference, with a median of 15 and a mean of 22 nucleotide differences, and a maximum of 385 nucleotide differences. Using standard phylogenetic analyses, many sequences could not be distinguished from each other because they had the exact same base pair difference, even though these differences occurred at different positions along the aligned sequences. If we take the presence of only one base difference between any two sequences again, the base difference "TC" occurred, for instance, in 134 undistinguishable sequence pairs (20.5%), but also "CT" in 60 undistinguishable pairs (9.2%). With *asymdist* computations, all those single differences at unique positions are translated into distinct coordinates. Of course, evolutionary models can differentiate a single difference between CT and AG, but not among the different cases of CT or AG occurring at different positions along the sequences.

Next, we investigated whether the *distances* between the sequence points in the (1D, 2D, 3D) CAPASYDIS seqverses were related to the number of nucleotide differences observed for the same sequences. We found that the number of differences was significantly and positively correlated with the distances between sequences based on their 1D, 2D, and 3D CAPASYDIS coordinates with Pearson correlation coefficients of 0.643,

0.640, and 0.640, respectively (all $P<0.001$; **Fig. S3**). So even at this fine resolution level, the number of nucleotide differences is correlated with the distance between the sequences in the seqverses be they of 1, 2 or 3 dimensions. Finally, the relationships between ML distances obtained with TIM2+F+I model in IQ-TREE (Minh *et al.*, 2020) were compared to the distances obtained between the 1D, 2D and 3D seqverses of the 948 *E. coli* sequences with 29,932 nucleotide sites. In the classical phylogenetics approach, a pair that involves "-" (e.g. "A-") was given 0.0000010 vs. ("AG") 0.0012879 (1,287 times larger). Of note, the distances estimated by maximum likelihood-based models and *asymdist* are not directly comparable, as the former treat gaps ("-") as unknown characters, essentially ignoring them during likelihood calculations and ignoring them for the site-likelihood calculation of a given tree. This is in contrast to the calculations used here, where a specific weight is given also to cases where a gap is involved at a position (**Table 1**). Overall, the calculated ML distances fell into two groups, with a range of values from 0-0.003, and there were only 27 out of 655 (4.1%) distances which were numerically unique, illustrating the lack of discrimination already evidenced at the distance matrix level among closely related sequences when using the classical calculation approach. These 655 distances between the corresponding sequences in the seqverses consisted of 441 (67.3%), 516 (78.8%), and 585 (89.3%) unique distances (at the e-10 precision level) when considering the 1D, 2D or 3D seqverses, respectively. This suggests that, not only the positions of the sequences are unique in the seqverse, but also that small differences (one nucleotide) between sequences in the seqverse are well identifiable. The traditional cluster analysis of the 200 *E. coli* sequences that only had one nucleotide difference to another (**Fig. 7C-D**), further confirmed the difficulty to represent the sequence information in a definitive, comparable way using a traditional dendrogram or phylogram, as the representation is equivalent even if the leaves or branches are rotated in the dendrogram.

Discussion

CAPASYDIS is a computational framework that aims at mapping sequences in a low-dimensional space in a way that is non-arbitrary, consistent and unique, by directly leveraging the information content in the sequences themselves. The approach is built around the novel concept of asymmetric distances between sequences, which produce a unique score (coordinate) for each distinct sequence, and thus allow a distinct point representation of each single sequence in a reduced projection space. The choice of the function involved in the calculation of *asymdist* was found empirically by trials and errors, but other mathematical functions could be envisioned, given that they map each unique event uniquely into a resulting coordinate space, and ideally provide a distance metric based on biological properties. The projection space dimensionality is defined by the number of chosen reference sequences. Ideally, the projection space is of dimension three or lower, so that users can display and interact easily with the rendered seqverses. The mapping of each sequence is guaranteed to be unique if the same reference sequence(s) and *delta* values are used for a given query sequence. This standardized approach allows the description of the properties of seqverses (shape, boundaries, spread), and the

identification of areas of high or low sequence densities, as well as outliers. The seqverses of massive datasets can be calculated and easily explored, as demonstrated here with the analyses of a large set of very long sequences of rRNA sequences from the three domains of life.

Beyond the possibility of the CAPASYDIS approach to embed sequence information into a reduced space, an essential question was whether the obtained projections were associated with biologically meaningful patterns. This question is pertinent because given their asymmetric nature, the observed *asymdist* distances could potentially distort the taxonomic relationships between sequences. Similarly, a change that occurs at the same position in the beginning of two unrelated sequences could lead to an unwanted vicinity in the map. Here is where the choice of the magnitude of the *delta* values is essential, as a way to retain the biological relationships encoded in the sequences: On the one hand, each positional variation to the reference sequence should have a noticeable, unique impact on the resulting score of the query sequence. On the other hand, this impact should be very small as compared to cases of identical positions to retain the signal from the information present in the sequences. To ensure this is the case, *delta* values were chosen here to be ten to thousand times smaller than the magnitude of the corresponding index values.

To address the hypothesis that the signal in the sequences is also preserved by CAPASYDIS, we used the example of the rRNA seqverses, and compared the obtained positions in the seqverses using the taxonomic label information associated with the FASTA sequences, but also small-scale variations among sequences differing by one single change to each other. We showed that the taxonomic information was very well preserved in the resulting CAPASYDIS mapping, and that the hypothesis-free, three-dimensional seqverses generated for the three domains of life were spatially distinct, but that they also differed in shapes, sizes and patterns of variation around their centroids. Using as input a MSA and making simple assumptions about how base changes should be encoded, we identified Archaea as the most distant sequence to that of the bacterium *E. coli*, and that the most distant to Archaea was then an Eukaryotic organism. Organisms of the domain Archaea are often considered the most distant relatives of bacteria, including *E. coli*, and that Extremophiles, such as thermophiles, acidophiles, or halophiles, are likely to be distantly related to *E. coli*. Further, the detailed analyses at the phylum, order or family levels for bacteria also revealed that closely related genera may display a "continuum" in their seqverses, as illustrated by the case of Enterobacteriaceae. For specific bacterial taxa, we also identified diversity patterns associated with more systematic divergence from the rest of the sequences of the domain Bacteria, suggesting the existence of either diversification patterns specific to those taxa or some difficult placement of rRNA sequences for these organisms in comparison with the rest of the rRNA gene sequences in the dataset. Not only were the coordinates of the sequences in the seqverses taxonomically meaningful, but also the distances between points in the seqverses preserved meaningful information related to sequence differences. While investigating the

underlying causes of the identified diversification patterns lies beyond the scope of this study, it would be a very interesting avenue for future research.

The strengths of CAPASYDIS is that it addresses the fundamental problem of providing absolute and not relative positioning of sequences from large datasets in a reduced dimensional space. The CAPASYDIS approach avoids the calculation of all-versus-all sequence distances, by choosing *a priori* main axes of references, so that the patterns of diversity can be visualized in a reduced space. Based on simple requirements consisting of choosing specific reference sequences (in and outgroups of interest for instance) and *delta* values, the same coordinate system could be obtained across studies or experiments, enabling the description of already known seqverses, their expansion direction and rate, or the discovery of yet uncharted seqverse areas. Some regions in the seqverses are also expected to remain unpopulated. The intricate folding of molecules like ribosomal DNA (rDNA) places functional constraints on which mutations are possible. This is because the molecule's secondary and tertiary structures—its precise 2D and 3D shape—are essential for its function. Not all possible mutations can be accommodated, as many would disrupt the molecule's critical folding, rendering it nonfunctional (Maeder, Conn and Draper, 2006). As the Tree of Life continues to be a central organizing principle in biology, our understanding of it should grow in a scalable manner, i.e. in a manner that can cope with the ever-increasing amount of data generated by sequencing. Unlike traditional methods that focus on summarizing variability mathematically using indices or graphically using ordination or phylogenetic trees, CAPASYDIS provides a direct and definitive visualization of the sequence space itself. The reverse function, i.e. going from a random point on the seqverse back to a sequence of bases, is less trivial, because points are not randomly located in the seqverse and the identification of a point would need to be made at the same precision as the map, e.g. e-10. In the case of a simple and constant *delta* value, one could find the most likely sequence by narrowing down the search of the most candidate sequence based on available neighboring sequences. For variable *delta* values, the back transformation from a point back to the sequence could be mathematically intractable, except maybe for short and simple sequence alignments. To deal with computational requirements for very large datasets, CAPASYDIS is implemented in Golang, a simple yet, high-performance, cross-platform scripting language. Several examples of detailed exploration and statistical summarizations are provided in the associated code base.

Despite the novelty of the concept, some limitations can be foreseen: First, the calculations essentially depend on the quality and extent of the MSA and of the underlying sequences being compared. Not only for CAPASYDIS but also for usual phylogenetic and bioinformatic approaches are the sequences expected to be of high quality, chimera checked, and optimally taxonomically classified. Noisy sequences are expected to distort the appearance of the seqverses, and adding inaccurate metadata on the seqverses would further obscure any meaningful biological patterns. When introducing a new sequence in the MSA to be plotted in an existing seqverse, if the alignment of the sequence does not modify the MSA length, then adding new points to the seqverse can be done trivially without the need to recompute the whole map again. Indeed, just the distance between

the references and new sequence needs to be computed. Yet, any change in the underlying MSA positions would result in a new set of coordinates for the whole map. Thus, the magnitude of the expected MSA changes in comparison to the alignment length and sequence composition would need to be accounted for to best estimate the expected changes in the whole map. For well-studied molecules like 16S and 18S rRNA genes, which are about 1.2–2.2 kilobases in length, utilizing a large reference MSA of around 50,000 possible positions is a useful strategy, as it includes enough positions to accommodate the alignment of highly variable rRNA regions. Thus, this helps create a stable framework that can accommodate new sequences without significantly altering the number of positions. Further, query sequences differing in length must be shortened to the smallest alignable region. Thus, this limitation precludes the direct combination of sequences originating from short and long read sequencing technologies for instance. This makes sense because the same degree of information should be available for all sequences and samples being compared on a given map.

Second, *asymdist* values are not intended to represent phylogenetic or evolutionary distances. Although the correct alignment of the sequences already implies *de facto* functional conservation, and the choice of the *delta* values can be made to reflect some "biological" knowledge about the mutational patterns, the resulting *asymdist* values do not distinguish between different types of evolutionary relationships (e.g., recent vs. ancient divergence), and, as such, some taxonomic groups may not be necessarily represented as expected by traditional phylogenetic or genomic reconstruction methods. The detection of horizontal gene transfers have led some to propose a “web of life” instead of a tree – with evidence of hybridization between species, even across different genera, which might blurred some of the traditionally established taxonomic boundaries (Goris *et al.*, 2007). Others have detected clear species boundaries among prokaryotic genomes (Jain *et al.*, 2018). Therefore, the definition of the properties (shape, density, spread, area) of sequences from a typical "species" in a seqverse is a discussion that remains to be addressed in the future. The CAPASYDIS concept does not aim at producing a new way to define species or operational taxonomic units, because the pairwise distances between all sequences have not been computed by design. In the low-dimensional projection space some distances between the sequences can thus be misleading as they may not necessarily reflect the expected taxonomic or phylogenetic distances. In addition, when calculating *asymdist* values for a given sequence direction, some regions inside a long sequence may be more informative or may evolve faster or slower than others, but the resulting effect on the seqverse is not yet known.

Third, the choice of the reference sequences directly affects the mapping results. As this choice is arbitrary, it can be seen as a limiting factor. To facilitate systematic comparison across studies, individual research groups could yet adopt a standardized reference system, akin to the convention of orienting geographic maps with North at the top. This convention would enable consistent interpretation and a scalable integration of findings about the already explored realms of biological diversity. Further, the easy specification (as the reference sequences are part of the same MSA) and fast calculations

enable as much flexibility as needed, especially for explorative studies. The "optimal" references, i.e. those that best render the diversity space, would depend on the sequences of interest and may also be study specific: For instance, if one is interested in a global map, then the most distantly related references could be used, as done here for the NR99 dataset. One can then use a global map and zoom-in using the specific coordinates to still explore a subspace of the seqverse while maintaining data coordinate comparability. On the other hand, if patterns of micro-evolution are the objectives of the study, then the reference sequences could be restricted to a much narrower sequence variability space, allowing a more targeted exploration of variational landscapes for sequences expected to vary only little. Finally, one may want to further rotate the seqverses to best view the data (this is indeed what is done interactively in the interactive, three-dimensional visualizations; see movies), using e.g. rigid axis rotation (PCA), and this may prove useful to explore the data structure and visualize the largest direction of variation for instance. Yet, if the rotation angles are based on the observed variance in the data, this process may lead to some non-standard presentation of the seqverses, which would defeat the use of a standardized method across studies.

Finally, further applications not demonstrated here might include, but are not limited to, a) mapping of metadata (associated functions or phenotype, sample descriptors, etc.) as additional layers onto the displayed points or seqverses. In this study, we focused on the taxonomic signal associated with the sequences, but study specific layers of information can easily be combined and analyzed using spatial statistics for instance. b) As long-read sequencing becomes more readily available, classification of the generated, potentially very long, sequences could be envisioned by using CAPASYDIS' ability to rapidly score a sequence given just a reference sequence and *delta* values, and to identify in which seqverse region it will be found. This may have implications for diagnostics or pathogen discovery for instance by reducing the detection time of novel or targeted features, while discarding unwanted sequences bases on their distances to a given reference. c) One could also make predictions about where an artificial sequence would theoretically be located in a given seqverse. This may help investigate the novelty of candidate sequences and define seqverse regions of interest. As shown with the NR99 case, the most diverging sequences to any sequence can readily be calculated, further enabling predictions and about the extent of divergence and spatial positioning in a given seqverse. d) Applications beyond DNA sequences could also be foreseen, as the concept could be used with any ordered characters for which a sorted classification is beneficial (e.g. RNA sequences, alignment of amino acids). Here again, *delta* values may serve as an adjustable parameter, enabling precise tailoring of the analysis to the specific nuances of different sequence types and variations as compared to their references.

In conclusion, CAPASYDIS offers a novel approach to delimit the boundaries of the *terra cognita* in molecular diversity surveys and will be useful for the fast classification and identification of new regions in the sequence universe. Future research should address how the approach can be further leveraged by integrating additional layers of biological

information. Such an approach would unlock a wealth of biological insights, allowing for a richer exploration within and between seqverses.

Funding

This work was supported by intramural funding of the Institute for Infectious Diseases, University of Bern, Switzerland.

Acknowledgements

The idea of an entire galaxy existing within a marble in the 1997 film "Men in Black" served as a key inspiration for this work on dimensionality, mapping, and interactive scaling.

References

Barry, D. and Hartigan, J.A. (1987) "Asynchronous distance between homologous DNA sequences," *Biometrics*, 43(2), pp. 261–276.

Byrd, A.L., Belkaid, Y. and Segre, J.A. (2018) "The human skin microbiome," *Nature Reviews Microbiology*, 16(3), pp. 143–155. Available at: https://doi.org/10.1038/nrmicro.2017.157.

Cavalli-Sforza, L.L. and Edwards, A.W.F. (1967) "Phylogenetic Analysis: Models and Estimation Procedures," *Evolution*, 21(3), pp. 550–570. Available at: https://doi.org/10.2307/2406616.

Felsenstein, J. (1978) "The Number of Evolutionary Trees," *Systematic Biology*, 27(1), pp. 27–33. Available at: https://doi.org/10.2307/2412810.

Felsenstein, J. (1981) "Evolutionary trees from DNA sequences: A maximum likelihood approach," *Journal of Molecular Evolution*, 17(6), pp. 368–376. Available at: https://doi.org/10.1007/BF01734359.

Glöckner, F.O. *et al.* (2017) "25 years of serving the community with ribosomal RNA gene reference databases and tools," *Journal of Biotechnology*, 261, pp. 169–176. Available at: https://doi.org/10.1016/j.jbiotec.2017.06.1198.

Goris, J. *et al.* (2007) "DNA–DNA hybridization values and their relationship to whole-genome sequence similarities," *International Journal of Systematic and Evolutionary Microbiology*, 57(1), pp. 81–91. Available at: https://doi.org/10.1099/ijs.0.64483-0.

Howe, A.C. *et al.* (2014) "Tackling soil diversity with the assembly of large, complex metagenomes," *Proceedings of the National Academy of Sciences*, 111(13), pp. 4904–4909. Available at: https://doi.org/10.1073/pnas.1402564111.

Jain, C. *et al.* (2018) "High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries," *Nature Communications*, 9(1), p. 5114. Available at: https://doi.org/10.1038/s41467-018-07641-9.

Jukes, T.H. and Cantor, C.R. (1969) "Evolution of Protein Molecules," in *Mammalian Protein Metabolism*. Elsevier, pp. 21–132. Available at: https://doi.org/10.1016/B978-1-4832-3211-9.50009-7.

Kortschak, R.D. *et al.* (2017) "bíogo: a simple high-performance bioinformatics toolkit for the Go language," *Journal of Open Source Software*, 2(10), p. 167. Available at: https://doi.org/10.21105/joss.00167.

Maeder, C., Conn, G.L. and Draper, D.E. (2006) "Optimization of a ribosomal structural domain by natural selection," *Biochemistry*, 45(21), pp. 6635–6643. Available at: https://doi.org/10.1021/bi052544p.

Minh, B.Q. *et al.* (2020) "IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era," *Molecular Biology and Evolution*, 37(5), pp. 1530–1534. Available at: https://doi.org/10.1093/molbev/msaa015.

Peng, Y. and Wang, S. (2025) "ggalign: A 'ggplot2' Extension for Consistent Axis Alignment. R package version 1.0.2.9000, https://github.com/Yunuuuu/ggalign."

Quast, C. *et al.* (2013) "The SILVA ribosomal RNA gene database project: improved data processing and web-based tools," *Nucleic Acids Research*, 41(D1), pp. D590–D596. Available at: https://doi.org/10.1093/nar/gks1219.

Roch, S. (2010) "Toward Extracting All Phylogenetic Information from Matrices of Evolutionary Distances," *Science*, 327(5971), pp. 1376–1379. Available at: https://doi.org/10.1126/science.1182300.

Roussel, J.-R. *et al.* (2025) "geometry: Mesh Generation and Surface Tessellation." (doi: 10.32614/CRAN.package.geometry). Available at: https://cran.r-project.org/web/packages/geometry/index.html.

Sievert, C. (2020) *Interactive Web-Based Data Visualization with R, plotly, and shiny*. Chapman and Hall/CRC.

Singer, G. a. C. *et al.* (2019) "Comprehensive biodiversity analysis via ultra-deep patterned flow cell technology: a case study of eDNA metabarcoding seawater," *Scientific Reports*, 9(1), p. 5991. Available at: https://doi.org/10.1038/s41598-019-42455-9.

Tamura, K. and Nei, M. (1993) "Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees.," *Molecular Biology and Evolution*, 10(3), pp. 512–526. Available at: https://doi.org/10.1093/oxfordjournals.molbev.a040023.

Totani, T. (2020) "Emergence of life in an inflationary universe," *Scientific Reports*, 10, p. 1671. Available at: https://doi.org/10.1038/s41598-020-58060-0.

Wickham, H. (2016) "ggplot2: Elegant Graphics for Data Analysis. https://ggplot2.tidyverse.org," in.

Yin, Z. *et al.* (2017) "Computing Platforms for Big Biological Data Analytics: Perspectives and Challenges," *Computational and Structural Biotechnology Journal*, 15, pp. 403–411. Available at: https://doi.org/10.1016/j.csbj.2017.07.004.

List of Movies

Movie 1. "Three domains_plotly"; https://doi.org/10.5281/zenodo.17049073
Movie 2. "Three domains_polygon_plotly"; https://doi.org/10.5281/zenodo.17049153
Movie 3. "Archaea seqspace", "Archaea_polygon_centroid_plotly"; https://doi.org/10.5281/zenodo.17049195
Movie 4. "Bacteria seqspace", "Bacteria_polygon_centroid_plotly"; https://doi.org/10.5281/zenodo.17049222
Movie 5. "Eukaryota_seqspace", "Eukaryota_polygon_centroid_plotly"; https://doi.org/10.5281/zenodo.17049247

A

	Reference		S1				S2				S3				Maximum diverging sequence			
Position	Bases	SqrtScore	Bases	Delta	Score	SqrtScore	Bases	Delta	Score	SqrtScore	Bases	Delta	Score	SqrtScore	Bases	Delta	Score	SqrtScore
1	A	1.000	A	0	1	1.000	A	0	1	1.000	A	0	1	1.000	**N**	**0.01**	1.01	1.005
2	C	1.414	**A**	**0.01**	2.01	1.418	C	0	2	1.414	C	0	2	1.414	**N**	**0.01**	2.01	1.418
3	G	1.732	G	0	3	1.732	G	0	3	1.732	G	0	3	1.732	**N**	**0.01**	3.01	1.735
4	T	2.000	T	0	4	2.000	T	0	4	2.000	T	0	4	2.000	**N**	**0.01**	4.01	2.002
5	A	2.236	A	0	5	2.236	**C**	**0.01**	5.01	2.238	A	0	5	2.236	**N**	**0.01**	5.01	2.238
6	T	2.449	T	0	6	2.449	T	0	6	2.449	T	0	6	2.449	**N**	**0.01**	6.01	2.452
7	G	2.646	G	0	7	2.646	G	0	7	2.646	G	0	7	2.646	**N**	**0.01**	7.01	2.648
8	C	2.828	C	0	8	2.828	C	0	8	2.828	C	0	8	2.828	**N**	**0.01**	8.01	2.830
9	A	3.000	A	0	9	3.000	A	0	9	3.000	**T**	**0.01**	9.01	3.002	**N**	**0.01**	9.01	3.002
10	T	3.162	T	0	10	3.162	T	0	10	3.162	T	0	10	3.162	**N**	**0.01**	10.01	3.164
SumSqrtScore		22.468				22.472				22.471				22.470				22.493

B

	Reference	S1	S2	S3	Max
Number of difference	0	1	1	1	10
SumSqrtScore	22.468	22.472	22.471	22.470	22.493

C

	Reference	S1	S2	S3	Max
asymdist	0	0.1407937	0.0891124	0.0664	1

Figure 1. A conceptual example. A) Example of alignment dataset detailing the calculation of **B**) *SumSqrtScore* (Eq. E2), and **C**) *asymdist* (Eq. E3) consisting of re-scaled *SumSqrtScore* values to [0,1] range for the simple case of constant $delta$ (0.01 for a base difference; 0 otherwise). Sequences S1, S2, and S3, differ only by a single mutation as compared to the reference sequence, at the beginning, middle and end of their sequences, respectively. The maximum diverging sequence contains "N" (i.e. unknown, yet different base) at all positions of the aligned sequences.

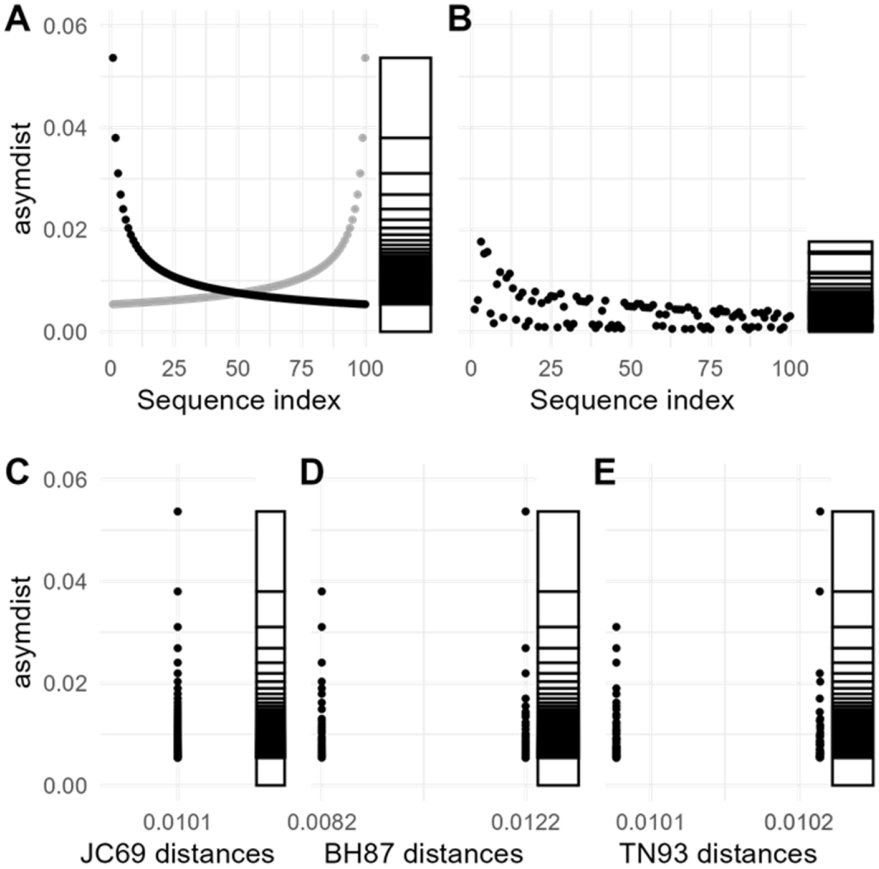

Figure 2. Illustration of *asymdist* properties on a set of 100 aligned sequences. A 100-base long nucleotide sequence was modified one base at the time, from left to right, to generate a set of 100 distinct sequences. **A**) These sequences were sorted by the position (index) of the nucleotide change in the sequences (x axis) from left to right, and *asymdist* with constant delta (0.01) was calculated for each sequence using the original sequence as reference. The y axis depicts the asymdist for each of those sequences (black dots), with a marginal histogram to confirm the uniqueness of each *asymdist* value. Grey dots correspond to asymdist calculated for the reverse order of the aligned sequences (reading from right to left). **B**) *asymdist* values were calculated using case-specific *delta* values, as proposed in **Table 1** for the same set of sequences. Comparison of *asymdist* values (*delta* = 0.01) with distances estimated by different genetic distance models: **C**) Jukes–Cantor 1969 (JC69) or Felsenstein 1981 (F81), **D**) Barry and Hartigan 1987 (BH87), and **E**) Tamura and Nei 1993 (TN93) models for the same set of 100 sequences.

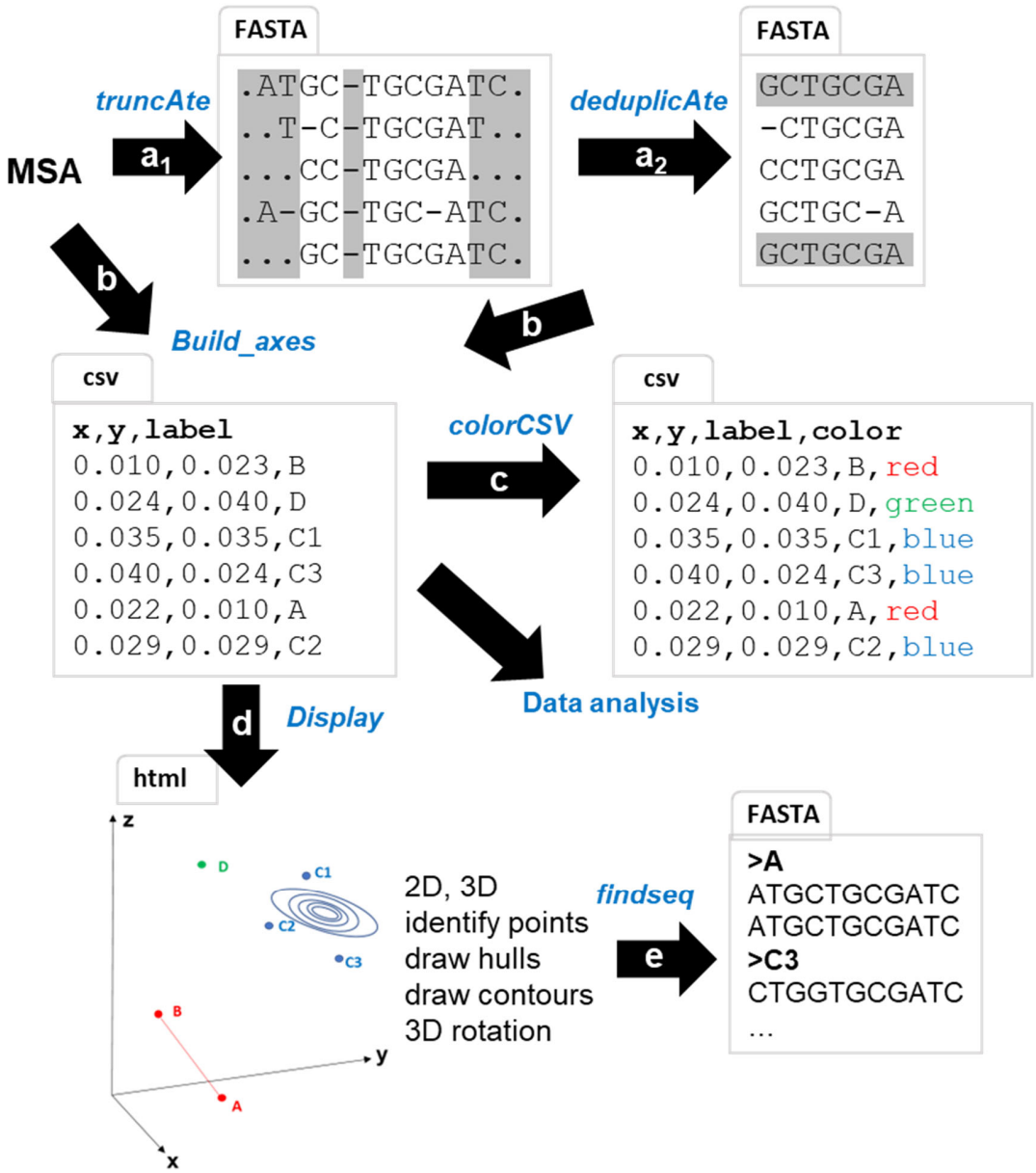

Figure 3. **Workflow for CAPASYDIS**. Starting from a multi-sequence alignment (MSA), the following steps are performed: **a₁**) The MSA is truncated to keep only the positions of the MSA that are not missing or unknown bases for a given k proportion of bases. In grey, the information that is removed from the MSA. **a₂**) Redundant sequences in the MSA are removed. **b**) the *asymdist* distances are computed relative to the chosen reference and are merged into one CSV file. Bad entries in the taxonomy fields are removed. **c**) Specific colors can be assigned to sequence features (e.g. taxonomic levels) to facilitate their visualization. The file can also be used directly for further data analyses. **d**) The coordinate file can be displayed interactively to allow exploration of the seqverse. **e**) Sequences can be retrieved from the original MSA for downstream processing. The resulting file formats (FASTA, csv or html) are also indicated for each step.

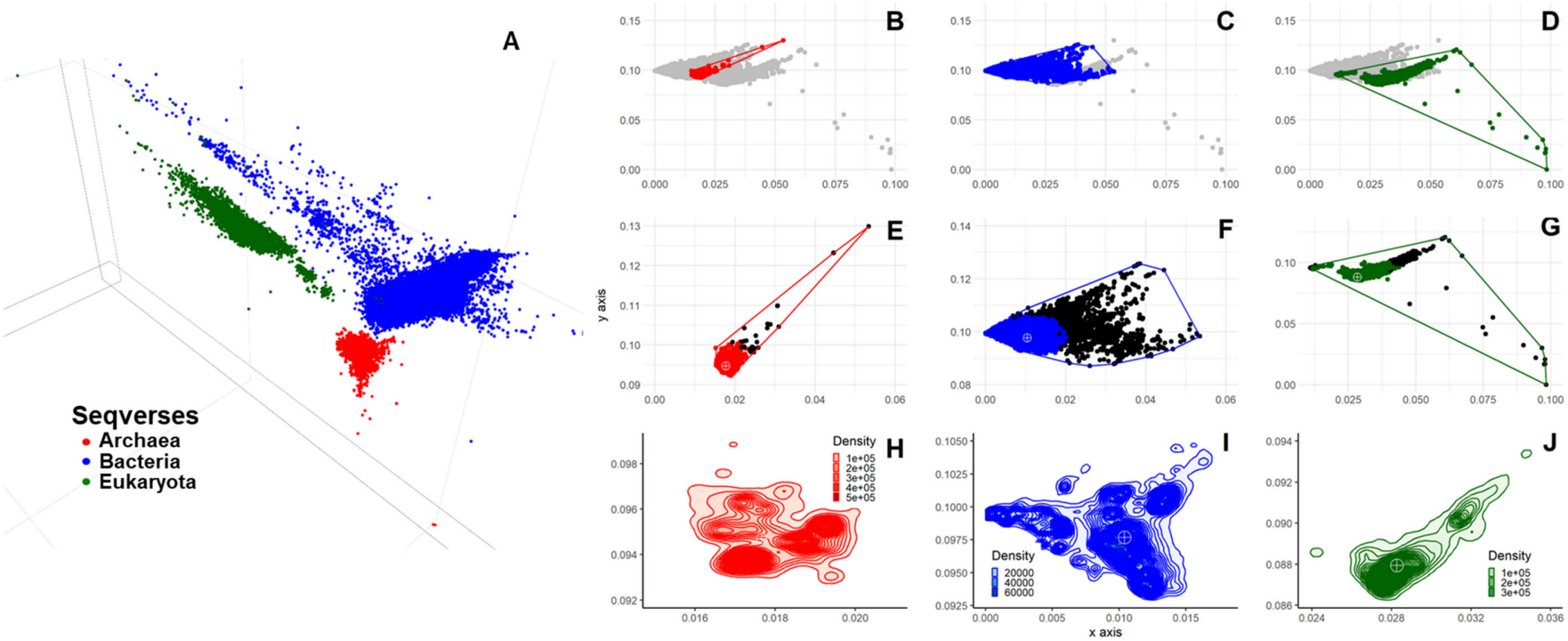

Figure 4. Visualization of the rRNA gene seqverse (n = 331,663 unique sequences) of the three domains of life. The NR99 MSA was analyzed with the CAPASYDIS workflow, which produced *asymdist* values with references to *E. coli* (Bacteria; axis "x"), *H. depressa* (Eukaryota; axis "y"), and *Pyrobaculum* (Archaea; axis "z"). Each sequence point was colored by its associated domain with Archaea in red Bacteria in blue, Eukaryota in green. **A**) Three-dimensional visualization of the 16S rRNA gene seqverse). Two dimensional convex hulls of the **B**) Archaea seqverse, **C**) Bacteria seqverse, and **D**) Eukaryota seqverse, with sequences from other domains colored in grey. Panels **E)**, **F)**, **G)** represent magnified views of each respective seqverse with sequences belonging to the 1% most distant sequences to the domain centroid indicated in black points. Panels **H), I), J)** represent 2D kernel density estimations (with 50 bins) which are displayed as contour lines, for Archaea, Bacteria, and Eukaryota, respectively. The domain centroids are indicated as white markers in panels **E** to **J**.

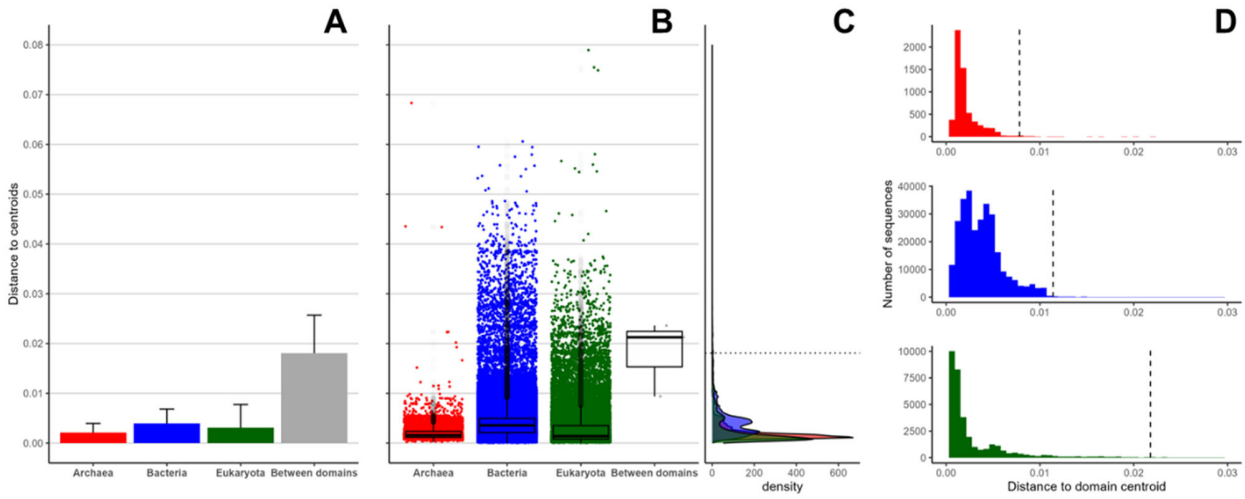

Figure 5. **Between- and within-domain distances to centroids**. Using the CAPASYDIS 3D distances of the NR99 dataset, **A**) the centroids of each domain seqverse were calculated and the within-domain distances to the centroids are depicted as barplot (means and standard deviation), as well as the average distance between the domain centroids (grey bar). Given the nature of the dataset, the numbers of sequences for each domain are unbalanced (with 6,031, 291,195, 34,437 points for Archaea, Bacteria, and Eukaryota, respectively), and with only three points corresponding to the distances between the domain centroids. **B**) A boxplot visualization of the same data with jittered points and their **C**) marginal density plots. The dotted horizontal line represents the mean distance between centroids. Seven points are not shown in panel B) because they are located between 0.08-0.13. **D**) Distribution of distances to the centroid for each domain with a dashed line representing the 99th percentile for each distribution.

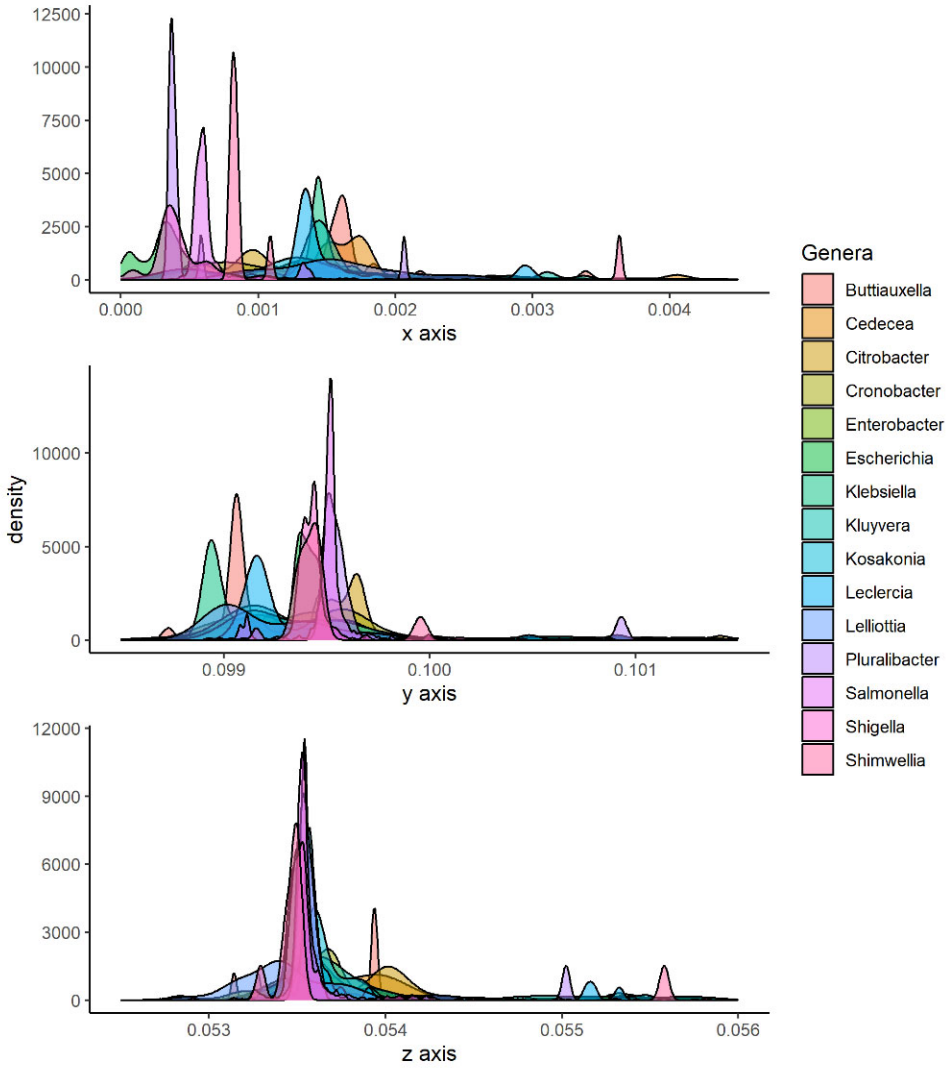

Figure 6. Density curves of sequence positions of the top 15 most abundant genera of the family Enterobacteriaceae onto individual CAPASYDIS axes.

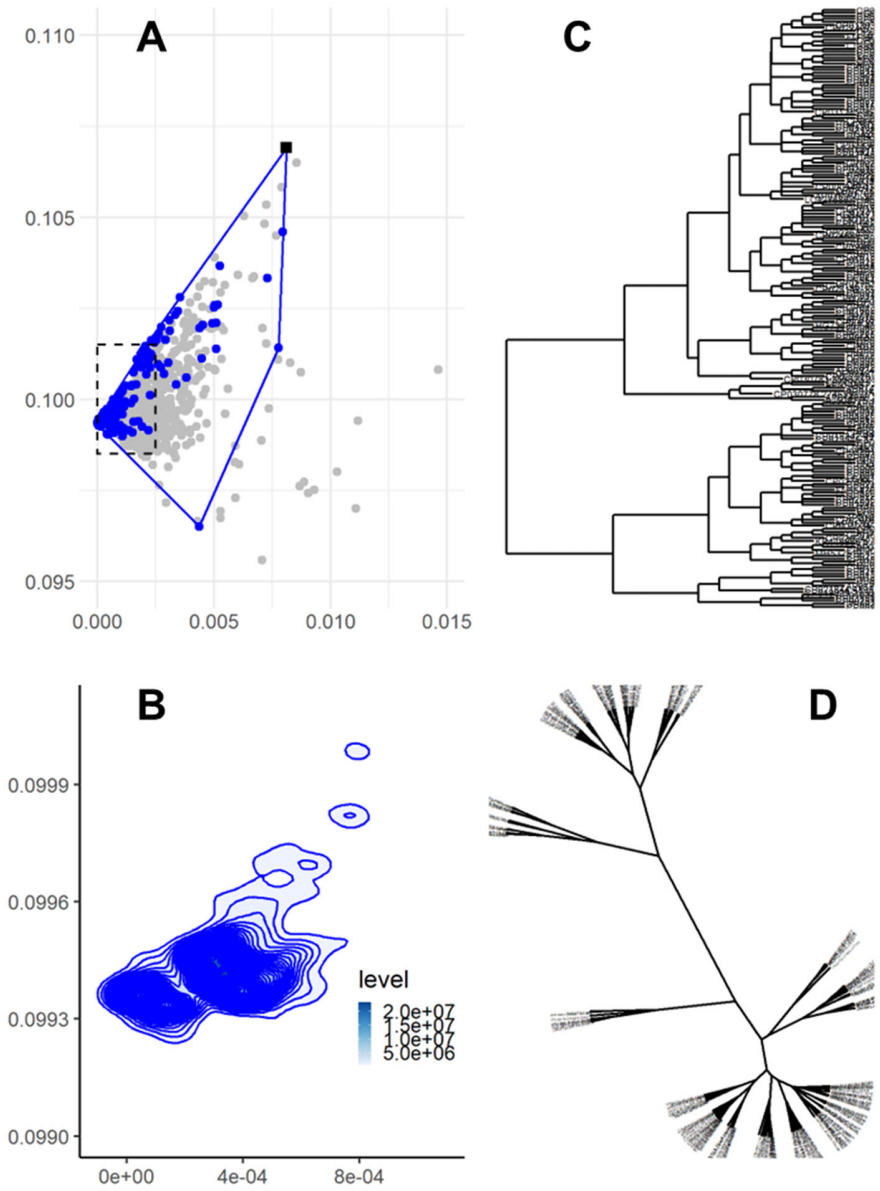

Figure 7. **Visualization of *E. coli* seqverse within the Enterobacteriaceae. A**) Within the Enterobacteriaceae (n = 4,408 sequences) colored in grey, the seqverse of *E. coli* sequences (n= 948) is indicated as blue points, as well as its corresponding convex hull. The black square represents the only sequence that was found at more than 99th percentile from the centroid (AOQL01000076.469.2061). The dashed rectangle indicates the coordinates of the region shown in panel B. **B**) Contour plot (bins= 50) of the *E. coli* seqverse with density lines. **C**) Rectangular phylogram of the 200 *E. coli* sequences that only differed by one nucleotide change from each other. The ML tree was inferred from the alignment of 29,932 nucleotide sites with the TIM2+F+I model of substitution implemented in IQ-TREE (version 2.1.2). **D**) An unrooted version of the rectangular phylogram shown in panel C.