Clustering Large-scale Genetic Data for Hypercubic Inference of Evolutionary Pathways of Antimicrobial Resistance

[a,b,*], [a,b,c], [a,b]

[a] *UIB, Mathematics*

[b] *CAMRIA*

[c] *Kwara State University, Malete, Nigeria*

Abstract

Antimicrobial resistance (AMR) is rapidly emerging as a major threat to the global public health system. The pathogen of interest in the study is Klebsiella pneumoniae known to be associated with AMR pathogenic infections and commonly found in hospitalized patients.In recent years, genomic technology has been widely used to provide a far more detailed picture of the evolution and spread of AMR, resulting in large-scale genomic data- a major challenge. Using large-scale genomic data to learn and predict AMR evolutionary pathways is not only a scientific necessity but also a matter of significant public interest, as it directly impacts global public health. The large-scale genomic data can be perceived as a sequential stochastic acquisition of binary traits, involving the presence or absence of genes. Therefore, to effectively address the increasing threat of AMR, it is imperative to understand how bacteria acquire AMR genes from large-scale genomic data through evolutionary dynamics. Previous studies explored the Hypercubic inference method based on a hidden Markov chain model to dynamically infer the acquisition of binary traits through the Bayesian method (HyperTraPS) and adapted Baum–Welch (expectation–maximization) algorithm (HyperHMM). However, the previous studies lack the capability of directly handling large-scale genomic data. To overcome this barrier, we develop a flexible approach that combines the power of the clustering approach and HyperHMM inference method called cluster-HyperHMM. The study applies the proposed methods to the synthetic and real-life data on Klebsiella Pneumonia to compare the resistance patterns and evolutionary progressive pathways in the six continents. The results of evolutionary pathways revealed differences in the genome feature acquisition in various continents.

Keywords: Cluster analysis, Hypercubic inference, Large-scale data, Evolutionary pathways

[*] Corresponding author

Email addresses: (), `kazeem.dauda@kwasu.edu.ng (Alternative Affiliation: Kwara State University, Malete)` (), ()

1. Introduction

Antimicrobial resistance (AMR) is a global challenge to the modern healthcare system and it has become a major concern to computational biology, medicine, and other scientific domains [1, 2, 3]. One of the opportunistic bacterial species associated with antimicrobial resistance (AMR) is Klebsiella pneumoniae, particularly due to a significant increase in the production of extended-spectrum β-lactamases (ESBLs)[4, 5, 6, 7]. This increase has the potential to result in serious pathogenic infectious diseases affecting human health and is a major source of hospital infections[8].

Antimicrobial resistance (AMR) genes can be acquired stochastically through various mechanisms, including horizontal gene transfer (HGT), mutation, efflux pumps, and others [9]. The term 'mechanisms' refers to the transfer of genetic material, including AMR genes, from one bacterial cell to another, often across different species or strains, and it contributes significantly to the spread of antibiotic resistance among bacterial populations [10].

Despite several interventions by the healthcare system and actions (”A European One Health Action Plan against AMR”, [11]) to prevent and contain the emergence, evolution, and spread of AMR bacteria on a local and global scale. A severe pathogenic infection caused by bacteria continues to emerge and spread across the continents[12]. The persistence and spread of the bacteria species have been linked to the alteration of the microbial genomic sequence, providing insights into the mechanisms through which bacteria develop resistance genes, and resulted in large-scale genomic data[13, 14]. The large-scale genomic data can be viewed as a stochastic acquisition of binary traits consisting of the presence or absence of genetic features[15, 16]. Understanding these processes at the genomic level is crucial for developing strategies to address the evolutionary dynamics of AMR gene acquisition - a global health concern.

In previous studies, various approaches have been employed to reconstruct historical events and forecast the acquisition of binary traits in the fields of cancer progression and evolutionary biology. In cancer progression studies, several deterministic dependence models have been developed to explore patterns of dependencies in the irreversible acquisition of binary traits using cross-sectional data[17]. These models include Oncogenetic trees (OT)[18, 19], OncoBN (DBN)[20], Conjunctive Bayesian Networks (CBN)[21, 22], and Hidden Extended Suppes-Bayes Causal Networks (H-ESBCN, PMCE)[23], respectively. The cancer progression models rely on the assumption that each subject provides a single data point and, as such, is independently observed. These represent cross-sectional data, which is a standard format for the cancer progression models. However, these models do not incorporate essential evolutionary assumptions (e.g. species may share a common ancestor, heritable traits, etc.) and interpretations[17].

In parallel, some methods have been developed for cross-sectional data, which are also capable of handling phylogenetic and longitudinal dependencies (i.e. they incorporate evolutionary assumptions) in cancer science data. These methods are MHN[24],HyperTraPS[25, 15], and HyperHMM[26], respectively. These models have also been categorized as stochastic dependencies, as events can dynamically alter the probability of acquiring other events, either increasing or decreasing[17].

Challenges remain in applying these methods to infer the evolutionary dynamics and stochastic acquisition of binary traits (presence/absence of genes) from large-scale genomic data. These approaches leverage the use of hypercubic inference with a Hidden Markov Model (HMM)[27] to dynamically infer the acquisition of binary traits through the Bayesian method (HyperTraPS[25, 15]) and adapted Baum–Welch (expectation–maximization) algorithm (HyperHMM[26]).

So far, HyperTraPS has primarily been used to address specific evolutionary questions related to the dynamics of multiple coupled traits on small-scale data (i.e. number of traits are less than equal to 15). But, as large-scale scientific and biomedical data become more readily available, questions about the structure of dynamic pathways are expanding and gaining relevance in evolutionary biology and precision medicine. Therefore, the availability of large-scale data presents new opportunities and challenges that may require some modifications to the existing HyperTraPS methodology.

In 2020, Greenbury et al. [15] formalized the use of HyperTraPS in learning the structure of pathways in multiple coupled traits. This formalized HyperTraPS represents a novel and efficient parametric inference and model selection platform. It offers simultaneous capabilities for Bayesian inference of pathways and the identification of model structures that accurately describe the dynamics and interactions present within a given set of observations.

Furthermore, the scope of the HyperTraPS illustration has been extended to encompass various influence structures, such as cross-sectional, longitudinal, or phylogenetically interconnected data structures depending on the specific scientific context [15]. Additionally, the concept of pairwise interactions has seen recent advancements and rapid progress, as exemplified by the work of Gotovos et al.[28].In the realm of understanding cancer progression, diverse methodologies and models have been developed, leveraging alternative techniques as their foundation. To comprehensively understand these models, we encourage the reader to explore the research authored by [29, 30].

Until now, HyperTraPS[25, 15] uses Bayesian inference to dynamically infer pathways and optimally capture the dynamics and interactions present within a provided set of observations. However, the Bayesian methods have often required significant time to achieve converged results, which sometimes imposes practical limitations on their utilization. Therefore, Moen and Johnston[26] introduced and developed HyperHMM, an alternative method to infer dynamic pathways on directed hypercubes, all without imposing any limitations on state space or trait interactions and less time-consuming.

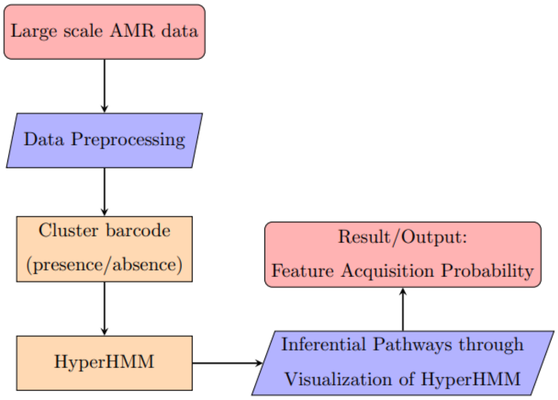
Thus far, HyperHMM[26] embarked on the utilization of the Hypercubic inference technique, employing a hidden Markov chain model for dynamically inferring binary trait acquisition. This process was enhanced using an adapted Baum–Welch (expectation–maximization) algorithm. Despite these advancements, the HyperHMM faced challenges when dealing with large-scale data, particularly those encompassing a larger number of traits. This intricate complexity has the potential to influence the probability of traits transitioning from one state to another. Hence, this study aims to develop a modified version of HyperHMM called Cluster-HyperHMM to overcome the challenges associated with handling large-scale genomic data and addressing the evolutionary dynamics of AMR genes across different continents.

The development of the Cluster-HyperHMM approach emerged as a strategic means to reduce the dimensionality of the original datasets, thus generating underlying clusters containing the presence/absence of genetic features. This refinement aims to augment the adaptability, practicality, and comprehensibility of the HyperHMM methodology, which was initially introduced by Moen and Johnston[26]. The resultant algorithm, named Cluster-HyperHMM, not only demonstrates reliability but also efficiency in terms of computational time − a focal point of the HyperHMM framework.

In this study, we showcase the effectiveness of this technique (Cluster-HyperHMM) through two distinct scenarios: first, utilizing synthetic datasets, and second, employing large-scale real-life data from K. pneumoniae, focusing on the stochastic acquisition of AMR genes. We conduct a comparative analysis of the Cluster-HyperHMM technique with other methods from the disease progression and evolutionary literature, emphasizing its unique position at the intersection of these fields.

2. Methodology

The research methodology of this study relies on the use of data preprocessing, cluster analysis, and HyperHMM ([26]), respectively. The methodological pipeline of our study is illustrated through the flowchart and algorithm presented in Figure 1. The large-scale AMR genetic data will be accessed through publicly available sources. Subsequently, the data will be cleaned, curated, and transcribed into cluster binary traits, a usable format required by this study. The process of constructing cluster barcodes (cross-sectional), phylogenetic estimation procedure, and cluster protocol are well explained in the subsequent section.


Large scale AMR data
Data Preprocessing
Cluster barcode (presence/absence)
HyperHMM
Inferential Pathways through Visualization of HyperHMM
Result/Output: Feature Acquisition Probability

(a) Flowchart Pipeline of Cluster-HyperHMM

```
Input  : Large-scale genomic dataset
Output: Visualization of HyperHMM output
Data preprocessing;
Construct cluster barcodes (presence/absence);
Cross-Sectional Analysis:;
Construct barcodes for cross-sectional snapshots;
Run barcodes through HyperHMM;
Estimated Phylogeny Analysis:;
Construct barcodes for estimated phylogenetic relationships;
Run barcodes through HyperHMM;
```

(b) Algorithm Pipeline of Cluster-HyperHMM

Figure 1: **Flowchart-Algorithm of Cluster-HyperHMM General Protocol**. The Cluster-HyperHMM protocol finds the clusters within the dataset and constructs cluster barcodes (presence/absence) capturing their evolutionary dynamics under different contexts in (a) cross-sectional views and (b) estimated phylogenies. Subsequently, these barcodes are processed through HyperHMM to extract meaningful patterns and associations, and the resulting output is visualized to provide insights into the underlying evolutionary processes.

2.1. Clustering protocol

The clustering approach is commonly used to identify hidden biological information (e.g. pattern, structure, and so on) in genetic features by grouping similar entities (such as genes, samples, etc.) based on shared characteristics[31, 31, 32]. Over the past years, a range of cluster analysis techniques and algorithms have been developed and explored in the literature to solve data clustering problems([33, 34, 35, 36]). In this study, the application of k-means clustering ([37, 38]) is employed, incorporating Manhattan distance and the gap-statistic method for determining the optimal number of clusters ([39]). To visually represent the clustering outcomes, we employ Principal Component Analysis (PCA) ([40]) and then create a plot illustrating the structure of PC1 against PC2. Additional details regarding the clustering algorithms, PCA protocol, and the selection of the optimal cluster number can be found in the Supplementary Information.

2.2. Construction of Cluster Barcodes (Presence/Absence)

The large-scale genomic data matrix M consists of the presence/absence of genes (g_1, g_2, ..., g_p) across different isolates. The construction of cluster barcodes (Presence/Absence) begins by first determining the optimal number of clusters using the gap statistic ([39]). Next, the dimensionality of this large-scale data is reduced using clustering techniques based on the optimal number of clusters to represent the presence or absence of certain genetic features within each cluster. A cluster is assigned '1' if at least one gene in that cluster is present (1). Otherwise, '0' is allocated to that cluster. Here, we denote presence with '1' and absence with '0'. For further understanding of how the $N \times L$ matrix was generated and the cluster barcode, please refer to synthetic example 1 in the supplementary material

The constructed cluster barcodes (presence/absence) can be viewed within different evolutionary contexts, such as cross-sectional snapshots or estimated phylogenies. The barcode created from matrix M is considered an independent observation (cross-sectional) because the data were generated individually for each isolate. Thus, we have an $N \times L$ matrix that represents isolates and traits (L), called cluster barcodes (Presence/Absence), and this is tagged **cross-sectionally** observed data. The details of the phylogenetic estimation procedure are given in the next section2.3.

2.3. Phylogenetic Estimation Procedure

The Cluster-HyperHMM algorithm can effectively estimate a phylogeny from the $N \times L$ matrix, which represents isolates and cluster barcodes, providing insights into the evolutionary history and relationships. The estimation process involves constructing a dendrogram through clustering methods from cross-sectional data and reconstructing a tree from the dendrogram to extract ancestor-descendant pairs. Ancestor and descendant pairs are extracted when all descendant nodes, such as A (101) and B (110), in the binary tree plot exhibit a '1' at the same position, we assign '1' to the corresponding position for the ancestor C (100); otherwise, we append '0'. This process iterates through all descendants and ancestors until reaching the root node.

2.4. HyperHMM: Efficient Pathways Inference on Transition Hypercubes using Baum–Welch Algorithm

HyperHMM([26]) is an evolutionary dynamics of feature acquisition techniques that rely heavily on the hypercubic transition graph based on hidden Markov models. The HyperHMM method uses the power of expectation–maximization (EM) called Baum–Welch Algorithm[41, 42] for the efficient estimation of parameters (graph edges/states). The algorithm of HyperHMM is well defined and described in the paper[26] by parameterizing $L2^{L-1}$ hypercubic edges individually, where L represents the number of features. In the Cluster-HyperHMM graph model, each possible state of the graph represents a binary barcode (0 *or* 1) of the number of clusters denoted by L. At the i^{th} position, a value of 0 signifies the absence of the i^{th} gene at a particular cluster and isolate, while a value of 1 indicates the presence of the i^{th} gene at a particular cluster and isolate. After organizing the data into an $N \times L$ matrix consisting of isolates and cluster barcodes (presence/absence), we used this matrix as input for HyperHMM ([26]). The major assumption of Cluster-HyperHMM is that the presence or absence of genes at a particular cluster with the corresponding isolate is seen as a stochastic and irreversible process, that is once a gene is present it can not be absent again at a particular cluster and vice versa. Thus, a hypercubic transition graph is then constructed based on the clusters, with each node (state) of the graph corresponding to the binary barcode of each cluster and a weighted edge from node $\mathscr{A}$ to node $\mathscr{B}$, if $\mathscr{B}$ differs from $\mathscr{A}$ with at least one absent/present of gene in the barcode. If the optimal number of clusters is 5, then each state of the transition graph will contain five binary barcodes as shown in the synthetic data in figure 2. For a clearer comprehension of our procedure, we illustrate our idea using a system (e.g. patient with a progressive disease) navigating across the hypercube.

This movement is from the binary string of all 0s to the binary string of all 1s, possibly but not always attaining the state of 1s, and traverses the edges from a given state in a probabilistic manner according to their weights.

The purpose of hypercube inference is to estimate the sets of edge weights that give a higher likelihood of observing the dynamics of a given system. For this purpose, this study embraces the use of Cluster-HyperHMM which utilizes the power of cluster and hidden Markov model based on the likelihood of emitting the processes on the given transition graph at various nodes (states). Additionally, Cluster-HyperHMM also allows the removal of noisy emissions and accounts for incomplete emissions in the transition graph.

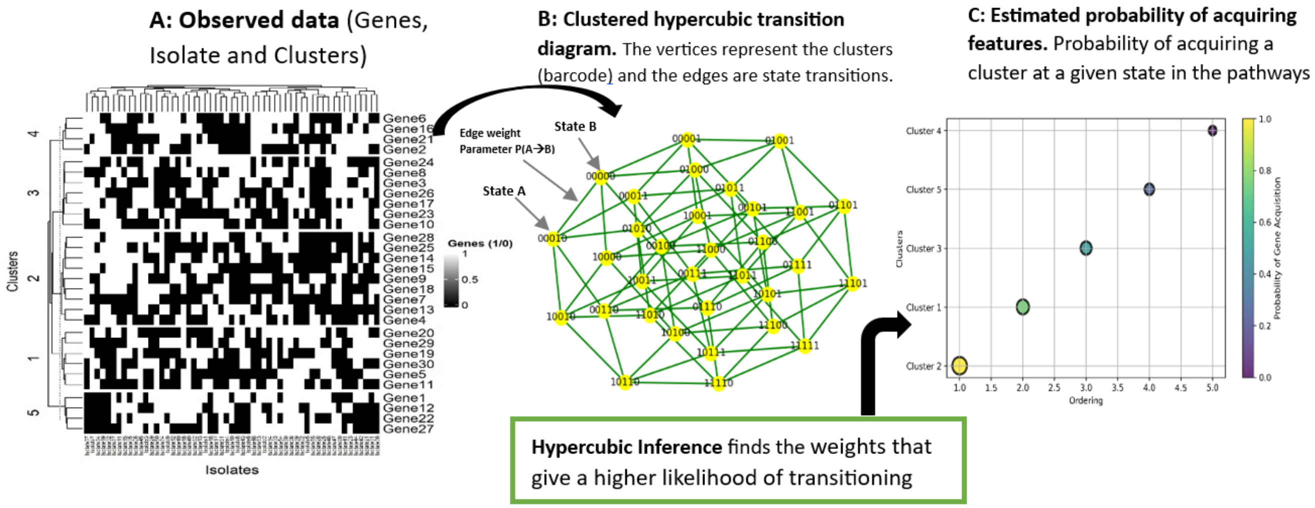


Figure 2: **Overview of Cluster-HyperHMM inference.** (A) The original observed data consist of present (1) and absent (0) of AMR genes "barcode", which is high-dimensional, and we then perform dimension reduction using cluster analysis to allow effective implementation of hypercubic inference. The five clusters are used to represent the original genes which may be independently or phylogenetically observed. (B) The cluster pairs in (A) are fed into the hypercubic inference to estimate the transition probability of moving from one cluster state to another. The latter is further used to find the most compatible parameters with a set of emitted observations. (C) The final output after learning the parameterization can be interpreted in several ways and not restricted to the following: 1. as a likelihood map showing which feature is most likely to be acquired at and what stage 2. clearly defined routes over the hypercube area 3. links between orderings of features, and other things.

2.5. Cross-sectional and Phylogenetically Coupled Data

In general, we don't have access to the phylogenetic relationship between our isolates. Therefore, assuming independence among isolates, when in reality, they are not truly independent, may be challenging. The concept underlying cross-sectional and phylogenetically coupled data is to independently examine the present status of genetic entities and dependently explore the historical context of the evolutionary dynamics of inferential pathways. In this study, cross-sectional data is generated from a population with different isolates and

AMR genes. The original AMR genes are then clustered into different group clusters and subsequently generate a cross-sectional matrix consisting of the presence/absence of clusters. Meanwhile, the phylogenetically coupled was generated in this study using the idea of Moen and Johnston ([26, 15]).

3. Results

*3.1. **Identifying Inferential Pathways from Synthetic Data***

Here, we consider the use of two simulated synthetic data preconstructed to enhance the applicability of Cluster-HyperHMM. The primary focus of this synthetic data is the generation of matrix M that represents the presence/absence of clusters across the isolates. In the first scenario, we consider synthetic data with 7 isolates and 5 genes, and we predetermine the number of clusters k to be 4. In this case, the dimension of M is 7×4 resulting in a simple system with $L = 4$ traits and with observations 1010, 1101, 1110, 0010, 0110, 0111, 1100. In this context, the obtained observations are independent and can be regarded as cross-sectionally observed data. The findings illustrated in Figure S1 in the supplementary material (A) depict the likelihood of acquiring features for the cross-sectionally observed synthetic data, as determined by the Cluster-HyperHMM method.

In the second scenario of the synthetic data, we consider evolutionary relationships among genes that are quantified through phylogenetic tree reconstruction. This study considered the use of a dendrogram with Manhattan distance as our method of tree reconstruction and consequently used the principle described in [26] to find ancestor and descendant linkage. The results of the tree reconstructed based on earlier synthetic data are as follows: 0000-1100, 0000-0010, 0010-0111, 0010-1010, 0010-0110, 1010-1110, and 1100-1101. The results presented in Figure S1 (B) reveal the probability of feature acquisition in the phylogenetically observed synthetic data, as determined through the Cluster-HyperHMM method.

Detailed in the supplementary material (A.1: Example 2), the second case of our synthetic data enriches comprehension of Cluster-HyperHMM. This presentation involves manual computation and data generation, encompassing both cross-sectional and phylogenetically coupled scenarios. It serves to deepen insight into the method's applications and nuances, fostering a more comprehensive understanding of its capabilities.

*3.2. **Real-life Data: Klebsiella pneumoniae AMR genes***

Following the validation of the Cluster-HyperHMM technique on synthetic datasets, next, we turned our attention to the investigation of a real-life medical dataset. The epidemiological genome study of bacterial isolates consists of 1574 Klebsiella pneumonia isolates with 120 genetic features were considered from publically available data (Institut Pasteur[43]). All isolates were collected from the BIGSdb-Kp database[43] together with other information such as the continent, country, hosts, and information relating to resistance info, respectively. The process for acquiring the data is intricately tied to all loci and antimicrobial resistance. We subsequently filtered the data by selecting entries with nonzero values in the resistance information column and specifically focusing on Klebsiella pneumoniae as a particular bacterium of interest.

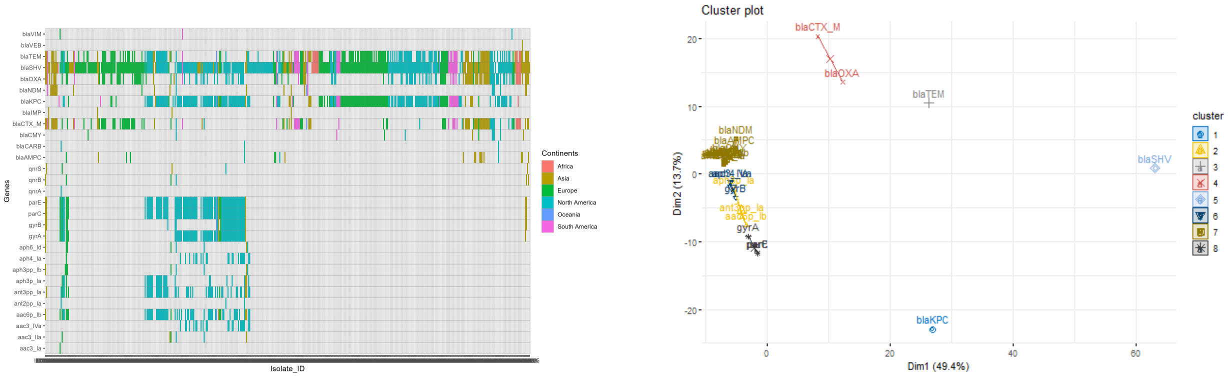
Following this, we assigned predictor genes a value of 0 for empty cells (indicating the absence of genes) and 1 for non-empty cells (indicating the presence of genes). Subsequently, non-informative genes were eliminated from the dataset. Dimensionality reduction was achieved by removing empty cells (see figure S5(a)) from both rows and columns, retaining only informative data points, resulting in 468 isolates with 29 genomic features. This process is termed the preprocessing method and is commonly used in advanced statistical analysis and machine learning to correct potential outliers, missingness, and incomplete data. After the preprocessing, the data, thus, consists of six continents with an unequal number of isolates representing the samples of individual hosts taken from each continent. The distribution of sequenced isolates across continents is as follows: Africa (13), 2.8%; Asia(78), 16.7%; Europe (157), 33.5%; North America (189), 40.4%; Oceania (3), 0.6%; and South America (28),6%. This distribution is visually represented in Figure 3(a).

Afterward, the study systematically applies the developed Cluster-HyperHMM to this real-life data on Klebsiella Pneumonia to compare the resistance patterns and evolutionary pro- aggressive pathways in the six continents. We start by implementing the clustering approach to identify hidden biological information (clustering structure) in the AMR genetic features. The optimal K identified through the Gab-statistic is 8 (i.e. $k = 8$, number of clustering) as shown in figure S5(b), and the corresponding identified latent pattern of the similar AMR genes was observed and presented in table S5(c).

Furthermore, this study deems it appropriate to assess the quality of the identified clusters using the average silhouette method (ASM) [33]. The purpose of the ASM method in this study is to evaluate how effectively genes align within their respective clusters, with a high average silhouette width serving as an indicator of effective clustering quality. The results from Figure S5(d) reveal that Cluster 8 stands out as a good cluster, exhibiting the highest average silhouette width compared to every other cluster. This suggests a common association of genes parC, parE, and gyrA with antimicrobial resistance (AMR) in bacteria.

In the regional comparative sample, a higher proportion of clusters were found in Asia, Europe, and North America, accounting for the higher proportion of genomes from these regions, while 50% of clusters were found in South America and with a higher proportion of genomes. The remaining continents Africa and Oceania possess a relatively low number of clusters, however, they both consist of a higher proportion of genomes since cluster 7 with the highest number of genomes present in both continents (see figure 3(a)).

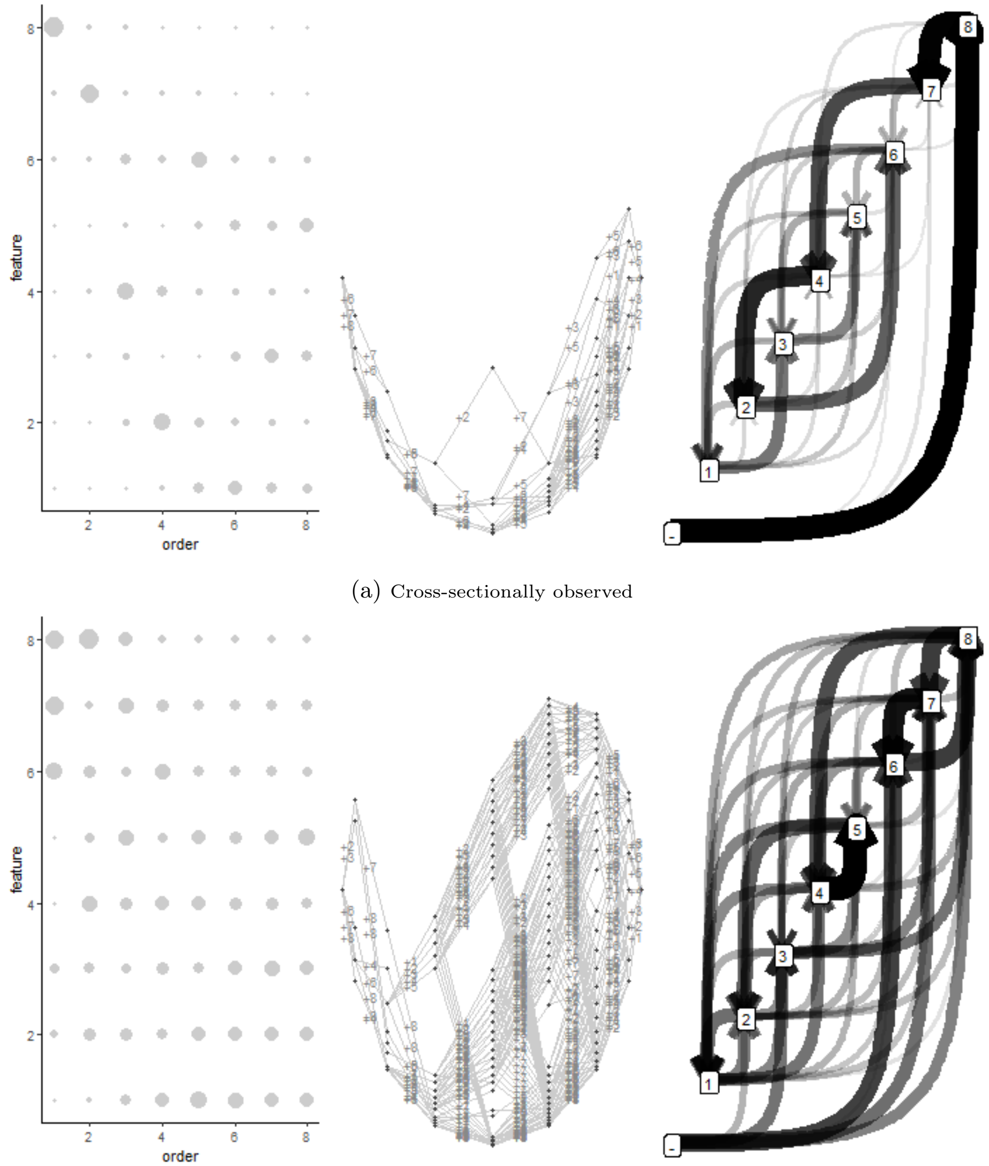
Following the successful demonstration of cluster analysis on this real-life data, the subsequent step involves advancing to the final stage of our algorithm, wherein we apply the HyperHMM procedure, transforming it into cluster-HyperHMM. We start by by examining the assumption of independence (cross-sectional) among the isolates, and the results are visualized in Figure 4(a). Figure 4(a) displayed the orderings from the HyperHMM approach and the probability of transitioning from state to state. The bubble shows the probability of getting AMR genes at a particular time point, after 100 bootstraps and the inference of evolutionary pathways on the hypercubic transition network, respectively. The initial feature acquired is cluster 8, succeeded by clusters 7, 4, 2, 6, 1, 3, and 5. In addition to the assumption of independence, the HyperHMM procedure can be employed to model phylogenetically observed data, providing a framework to elucidate the evolutionary relationships among these isolates. The findings from figure 4(b) disclose the progression of cluster acquisition, with the initial feature being cluster 7, followed by clusters 8, 4, 6, 1, 3, 2, and 5 in sequence.



(a) Prevalence (present/absent) of AMR genes across the continents

(b) PCA virtualization of clustering results

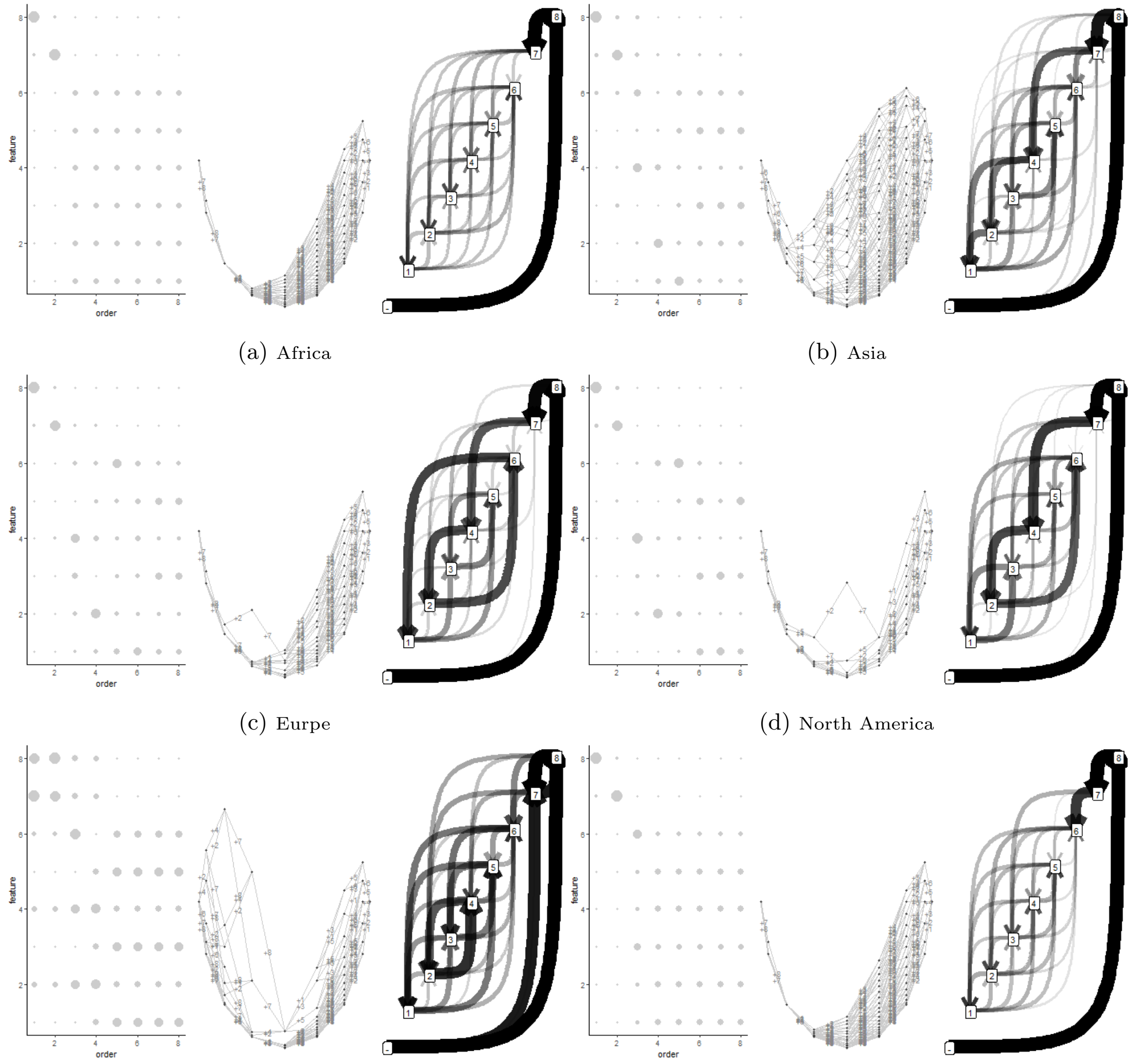
Figure 3: Exploring the Distribution of Antibiotic Resistance Genes (Present/Absent) Across Continents and Visualizing Clustering Results through PCA



(a) Cross-sectionally observed

(b) Phylogenetically observed

Figure 4: Cluster-HyperHMM Inferential Pathways for Cross-sectional and Phylogenetically coupled AMR gene



(a) Africa

(b) Asia

(c) Eurpe

(d) North America

(e) Oceania

(f) South America

Figure 5: Cluster-HyperHMM inferential pathways for regional Differences

4. Discussion of Results

5. Conclusion

Acknowledgements

The authors acknowledge CAMRIA research group

References

[1] A. Santos-Lopez, C. W. Marshall, M. R. Scribner, D. J. Snyder, V. S. Cooper, Evolutionary pathways to antibiotic resistance are dependent upon environmental structure and bacterial lifestyle, eLife 8 (2019) e47612. doi:10.7554/eLife.47612.
URL https://doi.org/10.7554/eLife.47612

[2] M. Ferri, E. Ranucci, P. Romagnoli, V. Giaccone, Antimicrobial resistance: A global emerging threat to public health systems, Critical Reviews in Food Science and Nutrition 57 (13) (2017) 2857–2876, pMID: 26464037. arXiv:https://doi.org/10.1080/10408398.2015.1077192, doi:10.1080/10408398.2015.1077192.
URL https://doi.org/10.1080/10408398.2015.1077192

[3] A. Cassini, L. D. Högberg, D. Plachouras, A. Quattrocchi, A. Hoxha, G. S. Simonsen, M. Colomb-Cotinat, M. E. Kretzschmar, B. Devleesschauwer, M. Cecchini, D. A. Ouakrim, T. C. Oliveira, M. J. Struelens, C. Suetens, D. L. Monnet, R. Strauss, K. Mertens, T. Struyf, B. Catry, K. Latour, I. N. Ivanov, E. G. Dobreva, A. Tambic Andraševic, S. Soprek, A. Budimir, N. Paphitou, H. Žemlicková, S. Schytte Olsen, U. Wolff Sönksen, P. Märtin, M. Ivanova, O. Lyytikäinen, J. Jalava, B. Coignard, T. Eckmanns, M. Abu Sin, S. Haller, G. L. Daikos, A. Gikas, S. Tsiodras, F. Kontopidou, Ákos Tóth, Ágnes Hajdu, Ólafur Guólaugsson, K. G. Kristinsson, S. Murchan, K. Burns, P. Pezzotti, C. Gagliotti, U. Dumpis, A. Liuimiene, M. Perrin, M. A. Borg, S. C. de Greeff, J. C. Monen, M. B. Koek, P. Elstrøm, D. Zabicka, A. Deptula, W. Hryniewicz, M. Caniça, P. J. Nogueira, P. A. Fernandes, V. Manageiro, G. A. Popescu, R. I. Serban, E. Schréterová, S. Litvová, M. Štefkovicová, J. Kolman, I. Klavs, A. Korošec, B. Aracil, A. Asensio, M. Pérez-Vázquez, H. Billström, S. Larsson, J. S. Reilly, A. Johnson, S. Hopkins, Attributable deaths and disability-adjusted life-years caused by infections with antibiotic-resistant bacteria in the eu and the european economic area in 2015: a population-level modelling analysis, The Lancet Infectious Diseases 19 (1) (2019) 56–66. doi:https://doi.org/10.1016/S1473-3099(18)30605-4.
URL https://www.sciencedirect.com/science/article/pii/S1473309918306054

[4] S. Argimón, S. David, A. Underwood, M. Abrudan, N. E. Wheeler, M. Kekre, K. Abudahab, C. A. Yeats, R. Goater, B. Taylor, H. Harste, D. Muddyman, E. J. Feil, S. Brisse, K. Holt, P. Donado-Godoy, K. L. Ravikumar, I. N. Okeke, C. Carlos, D. M. Aanensen, Rapid genomic characterization and global surveillance of klebsiella using pathogenwatch, Clinical Infectious Diseases 73 (4) (2021) S325–S335. arXiv:https://academic.oup.com/cid/article-pdf/73/Supplement_4/S325/41394889/ciab784.pdf, doi:10.1093/cid/ciab784.
URL https://doi.org/10.1093/cid/ciab784

[5] C.-H. Lee, L.-H. Su, Y.-F. Tang, J.-W. Liu, Treatment of ESBL-producing Klebsiella pneumoniae bacteraemia with carbapenems or flomoxef: a retrospective study and laboratory analysis of the isolates,

Journal of Antimicrobial Chemotherapy 58 (5) (2006) 1074–1077. arXiv:https://academic.oup.com/jac/article-pdf/58/5/1074/2229891/dkl381.pdf, doi:10.1093/jac/dkl381.
URL https://doi.org/10.1093/jac/dkl381

[6] P. Fils, P. Cholley, H. Gbaguidi-Haore, D. Hocquet, M. Sauget, X. Bertrand, Esbl-producing klebsiella pneumoniae in a university hospital: Molecular features, diffusion of epidemic clones and evaluation of cross-transmission, PLOS ONE 16 (2021) e0247875. doi:10.1371/journal.pone.0247875.

[7] A. K. H. HMarit, H. Jane, B. Eva, B. Ragna-Johanne, K. Håkon, I. R. Siren, S. Arnfinn, E. H. Kathryn, H. Iren, Within-patient and global evolutionary dynamics of klebsiella pneumoniae st17, Microb Genom 9 (5). doi:10.1099/mgen.0.001005.

[8] S. Navon-Venezia, K. Kondratyeva, A. Carattoli, Klebsiella pneumoniae: a major worldwide source and shuttle for antibiotic resistance, FEMS Microbiology Reviews 41 (3) (2017) 252–275. arXiv:https://academic.oup.com/femsre/article-pdf/41/3/252/18072899/fux013.pdf, doi:10.1093/femsre/fux013.
URL https://doi.org/10.1093/femsre/fux013

[9] K. Ryan, J. Sara, M. A. B. Jessica, G. J. Iain, Dynamic boolean modelling reveals the influence of energy supply on bacterial efflux pump expression, J R Soc Interface 19 (186) (2022) 20210771. doi:doi:10.1098/rsif.2021.0771.

[10] P. G. Vinayamohan, A. J. Pellissery, K. Venkitanarayanan, Role of horizontal gene transfer in the dissemination of antimicrobial resistance in food animal production, Current Opinion in Food Science 47 (2022) 100882. doi:https://doi.org/10.1016/j.cofs.2022.100882.
URL https://www.sciencedirect.com/science/article/pii/S2214799322000844

[11] E. Commission, A european one health action plan against antimicrobial resistance. brussels: European commission, https://ec.europa.eu/health/sites/health/files/antimicrobial_resistance/docs/amr_2017_action-plan.pdf, accessed: 2023-06-27 (2017).

[12] K. L. Wyres, M. M. C. Lam, K. E. Holt, Population genomics of klebsiella pneumoniae, Nature reviews. Microbiology 18 (6) (2020) 344—359. doi:10.1038/s41579-019-0315-1.
URL https://doi.org/10.1038/s41579-019-0315-1

[13] B. Aslam, M. Khurshid, M. I. Arshad, S. Muzammil, M. Rasool, N. Yasmeen, T. Shah, T. H. Chaudhry, M. H. Rasool, A. Shahid, X. Xueshan, Z. Baloch, Antibiotic resistance: One health one world outlook, Frontiers in cellular and infection microbiology 11 (2021) 771510. doi:10.3389/fcimb.2021.771510.
URL https://europepmc.org/articles/PMC8656695

[14] K. S. Baker, E. Jauneikaite, K. L. Hopkins, S. W. Lo, L. Sánchez-Busó, M. Getino, B. P. Howden, K. E. Holt, L. A. Musila, R. S. Hendriksen, D. G. Amoako, D. M. Aanensen, I. N. Okeke, B. Egyir, J. G. Nunn, J. T. Midega, N. A. Feasey, S. J. Peacock, Genomics for public health and international surveillance of antimicrobial resistance, The Lancet Microbe 4 (12) (2023) e1047–e1055. doi:https://doi.org/10.1016/S2666-5247(23)00283-5.
URL https://www.sciencedirect.com/science/article/pii/S2666524723002835

[15] S. F. Greenbury, M. Barahona, I. G. Johnston, Hypertraps: Inferring probabilistic patterns of trait acquisition in evolutionary and disease progression pathways, Cell Systems 10 (1) (2020) 39–51.e10. doi:https://doi.org/10.1016/j.cels.2019.10.009.
URL https://www.sciencedirect.com/science/article/pii/S2405471219303850

[16] F. Baquero, J. L. Martínez, V. F. Lanza, J. Rodríguez-Beltrán, J. C. Galán, A. S. Millán, R. Cantón, T. M. Coque, Evolutionary pathways and trajectories in antibiotic resistance, Clinical Microbiology Reviews 34 (4) (2021) e00050–19. arXiv:https://journals.asm.org/doi/pdf/10.1128/CMR.00050-19, doi:10.1128/CMR.00050-19.
URL https://journals.asm.org/doi/abs/10.1128/CMR.00050-19

[17] R. Diaz-Uriarte, A picture guide to cancer progression and monotonic accumulation models: evolutionary assumptions, plausible interpretations, and alternative uses (2023). arXiv:2312.06824.

[18] R. DESPER, F. JIANG, O.-P. KALLIONIEMI, H. MOCH, C. H. PAPADIMITRIOU, A. A. SCHÄFFER, Inferring tree models for oncogenesis from comparative genome hybridization data, Journal of Computational Biology 6 (1) (1999) 37–51, pMID: 10223663. arXiv:https://doi.org/10.1089/cmb.1999.6.37, doi:10.1089/cmb.1999.6.37.
URL https://doi.org/10.1089/cmb.1999.6.37

[19] A. Szabo, K. M. Boucher, Oncogenetic trees. In W.-Y. Tan and L. Hanin, editors, Handbook of Cancer Models with Applications, World Scientific Publishing Company, 2008.
URL http://www.worldscibooks.com/lifesci/6677.html.

[20] P. B. Nicol, K. R. Coombes, C. Deaver, O. Chkrebtii, S. Paul, A. E. Toland, A. Asiaee, Oncogenetic network estimation with disjunctive bayesian networks, Computational and Systems Oncology 1 (2) (2021) e1027. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/cso2.1027, doi:https://doi.org/10.1002/cso2.1027.
URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cso2.1027

[21] M. Gerstung, M. Baudis, H. Moch, N. Beerenwinkel, Quantifying cancer progression with conjunctive Bayesian networks, Bioinformatics 25 (21) (2009) 2809–2815. arXiv:https://academic.oup.com/bioinformatics/article-pdf/25/21/2809/48996991/bioinformatics_25_21_2809.pdf, doi:

10.1093/bioinformatics/btp505.
URL https://doi.org/10.1093/bioinformatics/btp505

[22] H. Montazeri, J. Kuipers, R. Kouyos, J. Böni, S. Yerly, T. Klimkait, V. Aubert, H. F. Günthard, N. Beerenwinkel, T. S. H. C. Study, Large-scale inference of conjunctive Bayesian networks, Bioinformatics 32 (17) (2016) i727–i735. arXiv:https://academic.oup.com/bioinformatics/article-pdf/32/17/i727/49023594/bioinformatics_32_17_i727.pdf, doi:10.1093/bioinformatics/btw459.
URL https://doi.org/10.1093/bioinformatics/btw459

[23] F. Angaroni, K. Chen, C. Damiani, G. Caravagna, A. Graudenzi, D. Ramazzotti, PMCE: efficient inference of expressive models of cancer evolution with high prognostic power, Bioinformatics 38 (3) (2021) 754–762. arXiv:https://academic.oup.com/bioinformatics/article-pdf/38/3/754/49007460/btab717.pdf, doi:10.1093/bioinformatics/btab717.
URL https://doi.org/10.1093/bioinformatics/btab717

[24] R. Schill, S. Solbrig, T. Wettig, R. Spang, Modelling cancer progression using Mutual Hazard Networks, Bioinformatics 36 (1) (2019) 241–249. arXiv:https://academic.oup.com/bioinformatics/article-pdf/36/1/241/48981322/bioinformatics_36_1_241.pdf, doi:10.1093/bioinformatics/btz513.
URL https://doi.org/10.1093/bioinformatics/btz513

[25] I. G. Johnston, B. P. Williams, Evolutionary inference across eukaryotes identifies specific pressures favoring mitochondrial gene retention, Cell systems 2 (2) (2016) 101––111. doi:10.1016/j.cels.2016.01.013.
URL https://doi.org/10.1016/j.cels.2016.01.013

[26] M. T. Moen, I. G. Johnston, HyperHMM: efficient inference of evolutionary and progressive dynamics on hypercubic transition graphs, Bioinformatics 39 (1), btac803. arXiv:https://academic.oup.com/bioinformatics/article-pdf/39/1/btac803/48763706/btac803_supplementary_data.pdf, doi:10.1093/bioinformatics/btac803.
URL https://doi.org/10.1093/bioinformatics/btac803

[27] A. v. d. Bosch, Hidden Markov Models, Springer US, Boston, MA, 2010, pp. 493–495. doi:10.1007/978-0-387-30164-8_362.
URL https://doi.org/10.1007/978-0-387-30164-8_362

[28] A. Gotovos, R. Burkholz, J. Quackenbush, S. Jegelka, Scaling up continuous-time markov chains helps resolve underspecification, in: M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, J. W. Vaughan (Eds.), Advances in Neural Information Processing Systems, Vol. 34, Curran Associates, Inc., 2021, pp. 14580–14592.

URL https://proceedings.neurips.cc/paper_files/paper/2021/file/7a50d83a1e70e9d96c3357438aed7a44-Paper.pdf

[29] R. Diaz-Uriarte, P. Herrera-Nieto, EvAM-Tools: tools for evolutionary accumulation and cancer progression models, Bioinformatics 38 (24) (2022) 5457–5459. arXiv:https://academic.oup.com/bioinformatics/article-pdf/38/24/5457/47886950/btac710.pdf, doi:10.1093/bioinformatics/btac710.
URL https://doi.org/10.1093/bioinformatics/btac710

[30] J. Diaz-Colunga, R. Diaz-Uriarte, Conditional prediction of consecutive tumor evolution using cancer progression models: What genotype comes next?, PLOS Computational Biology 17 (12) (2021) 1–23. doi:10.1371/journal.pcbi.1009.

[31] Y. Ren, T. Chakraborty, S. Doijad, L. Falgenhauer, J. Falgenhauer, A. Goesmann, A.-C. Hauschild, O. Schwengers, D. Heider, Prediction of antimicrobial resistance based on whole-genome sequencing and machine learning, Bioinformatics 38 (2) (2021) 325–334. arXiv:https://academic.oup.com/bioinformatics/article-pdf/38/2/325/49006417/btab681.pdf, doi:10.1093/bioinformatics/btab681.
URL https://doi.org/10.1093/bioinformatics/btab681

[32] J. Botelho, L. Tüffers, J. Fuss, F. Buchholz, C. Utpatel, J. Klockgether, S. Niemann, B. Tümmler, H. Schulenburg, Phylogroup-specific variation shapes the clustering of antimicrobial resistance genes and defence systems across regions of genome plasticity in pseudomonas aeruginosa, eBioMedicine 90 (2023) 104532. doi:https://doi.org/10.1016/j.ebiom.2023.104532.
URL https://www.sciencedirect.com/science/article/pii/S235239642300097X

[33] Clustering with the average silhouette width, Computational Statistics & Data Analysis 158 (2021) 107190. doi:https://doi.org/10.1016/j.csda.2021.107190.
URL https://www.sciencedirect.com/science/article/pii/S0167947321000244

[34] P. Patel, B. Sivaiah, R. Patel, Approaches for finding optimal number of clusters using k-means and agglomerative hierarchical clustering techniques, in: 2022 International Conference on Intelligent Controller and Computing for Smart Power (ICICCSP), 2022, pp. 1–6. doi:10.1109/ICICCSP53532.2022.9862439.

[35] A. E. Ezugwu, A. M. Ikotun, O. O. Oyelade, L. Abualigah, J. O. Agushaka, C. I. Eke, A. A. Akinyelu, A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects, Engineering Applications of Artificial Intelligence 110 (2022) 104743. doi:https://doi.org/10.1016/j.engappai.2022.104743.
URL https://www.sciencedirect.com/science/article/pii/S095219762200046X

[36] L. M. Abualigah, A. T. Khader, E. S. Hanandeh, A new feature selection method to improve the document clustering using particle swarm optimization algorithm, Journal of Computational Science 25 (2018) 456–466. doi:https://doi.org/10.1016/j.jocs.2017.07.018.
URL https://www.sciencedirect.com/science/article/pii/S1877750316305002

[37] A. M. Ikotun, A. E. Ezugwu, Boosting k-means clustering with symbiotic organisms search for automatic clustering problems, PLOS ONE 17 (8) (2022) 1–33. doi:10.1371/journal.pone.0272861.
URL https://doi.org/10.1371/journal.pone.0272861

[38] J. B. MacQueen, Some methods for classification and analysis of multivariate observations, Proc. fifth Berkeley Symp. Math. Stat. Probab. 1 (14) (1967) 281–297. doi:10.1007/s00500-019-04628-6.
URL https://doi.org/10.1007/s00500-019-04628-6

[39] R. Tibshirani, G. Walther, T. Hastie, Estimating the Number of Clusters in a Data Set Via the Gap Statistic, Journal of the Royal Statistical Society Series B: Statistical Methodology 63 (2) (2002) 411–423. arXiv:https://academic.oup.com/jrsssb/article-pdf/63/2/411/49590410/jrsssb_63_2_411.pdf, doi:10.1111/1467-9868.00293.
URL https://doi.org/10.1111/1467-9868.00293

[40] M. Lever, J.and Krzywinski, N. Altman, Principal component analysis, Nat. Methods 14 (2017) 641–642.

[41] L. E. Baum, T. Petrie, G. Soules, N. Weiss, A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains, The Annals of Mathematical Statistics 41 (1) (1970) 164 – 171. doi:10.1214/aoms/1177697196.
URL https://doi.org/10.1214/aoms/1177697196

[42] L. Rabiner, B. Juang, An introduction to hidden markov models, IEEE ASSP Magazine 3 (1) (1986) 4–16. doi:10.1109/MASSP.1986.1165342.

[43] I. Pasteur, Klebsiella pneumoniae amr genetic data, data retrieved from Institut Pasteur, https://bigsdb.pasteur.fr/klebsiella (2023).

6. Appendix: Supplementary Material

A.1: Synthetic Examples

Example 1

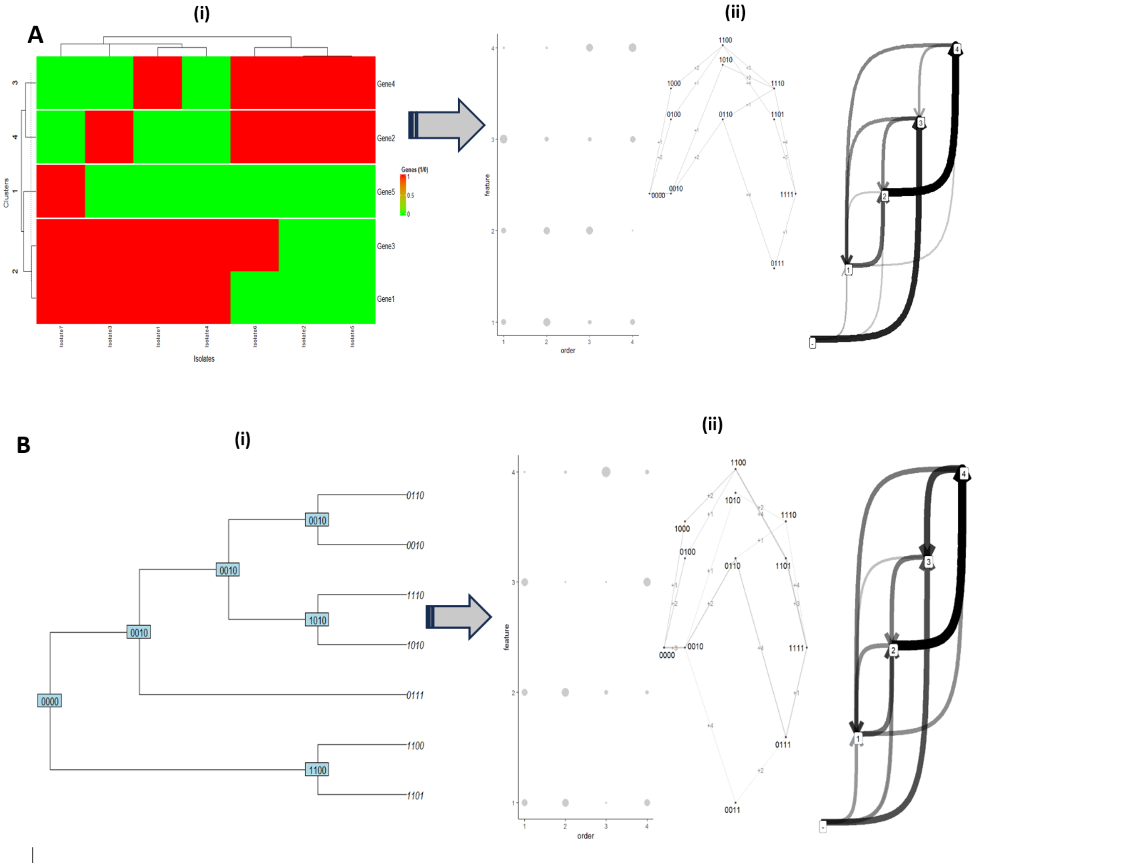


Figure S1:**Inferred evolutionary dynamics pathways from Synthetic Data.** (A) Cross-sectionally observed barcode was considered with four varying clusters representing the genetic trait as an example. (i) The heatmap that explains the virtual relationship between the clusters, isolates, and genes. (ii) The bubble plot unveils the trajectories of feature acquisition, illustrating the probability of obtaining trait 3 (Cluster 3) at the time order 1 and following the acquisition of other features. Other visualizations showcase the inferential pathways of the features. (B) Phylogenetically coupled data for the binary barcode for four varying traits. (i) The pair of observations were observed as ancestor-descendant, with the tip of the tree viewed as a descendant, and the internode is considered to be the common ancestor. The ancestors were estimated using the concept of maximum parsimony described in algorithm 1. (ii) The bubble plot illustrates the dynamics of feature acquisition, specifically highlighting the probability of acquiring trait 2 (Cluster 2) at time order 1, and it further captures the sequential acquisition of additional features. Additional visualizations elucidate the inferential pathways of the features.

```
Algorithm 1: Infer Trait Acquisition Patterns using Minimum Evolution Approach
Input  : Set of descendants, phylogenetic Tree, ensemble
Output: Inferred trait acquisition patterns
Initialize matrix $A$ to represent trait presence/absence in ancestors;
Initialize list of phylogenetic trees if ensemble;
foreach descendant in descendants do
    Traverse phylogenetic Tree from root to descendant;
    foreach internal node in the traversal do
        Mark traits present in descendant as present in ancestor;
    end
end
if ensemble then
    foreach tree in ensemble do
        Apply Minimum Evolution Approach to infer ancestor traits;
        Accumulate inferred trait patterns from each tree;
    end
    Summarize inferred trait patterns from ensemble;
end
return Inferred trait acquisition patterns;
```

Example 2

We generate synthetic data with 5 Isolates and 4 Genes as follows:

$$\text{N1} = \begin{array}{lcccc} & \text{Gene 1} & \text{Gene 2} & \text{Gene 3} & \text{Gene 4} \\ \text{Isolate 1} & 1 & 0 & 0 & 1 \\ \text{Isolate 2} & 1 & 0 & 0 & 0 \\ \text{Isolate 3} & 0 & 1 & 1 & 0 \\ \text{Isolate 4} & 0 & 0 & 0 & 1 \\ \text{Isolate 5} & 1 & 0 & 1 & 0 \end{array}$$

Let $N1$ be 5×4 matrix generated from the above data:

$$\text{N1} = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix}.$$

The second matrix $N2$ is based on 4 Genes and 3 Clusters

$$\text{N2} = \begin{array}{cccc} & \text{Cluster 1} & \text{Cluster 2} & \text{Cluster 3} \\ \text{Gene 1} & 1 & 0 & 0 \\ \text{Gene 2} & 0 & 0 & 1 \\ \text{Gene 3} & 1 & 0 & 1 \\ \text{Gene 4} & 0 & 1 & 0 \end{array}$$

$$\text{N2} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

We then generate matrix M_{ij} from the two matrices $N1$ and $N2$ as follows:
if any of $N1[i,]$ and $N2[,j]$, $\forall i, j$ is one (i.e. row- and column-wise have equal element ”1”), then M_{ij} is 1, else 0. Thus, we demonstrated how the matrix M_{ij} is generated here manually as follows: when $i = 1$ and $j = 1$

$N1[1,] = (1001)$ and $N2[,1] = (1010)$, therefore $M_{1,1} = 1$

when $i = 2$ and $j = 1$

$N1[2,] = (1000)$ and $N2[,1] = (1010)$, therefore $M_{2,1} = 1$

when $i = 3$ and $j = 1$

$N1[3,] = (0110)$ and $N2[,1] = (1010)$, therefore $M_{3,1} = 1$

when $i = 4$ and $j = 1$

$N1[4,] = (0001)$ and $N2[,1] = (1010)$, therefore $M_{4,1} = 0$

when $i = 5$ and $j = 1$

$N1[5,] = (1010)$ and $N2[,1] = (1010)$, therefore $M_{5,1} = 1$

when $i = 1$ and $j = 2$

$N1[1,] = (1001)$ and $N2[,2] = (0001)$, therefore $M_{1,2} = 1$

when $i = 2$ and $j = 1$

$N1[2,] = (1000)$ and $N2[,2] = (0001)$, therefore $M_{2,2} = 0$

when $i = 3$ and $j = 1$

$N1[3,] = (0110)$ and $N2[,2] = (0001)$, therefore $M_{3,2} = 0$

when $i = 4$ and $j = 1$

$N1[4,] = (0001)$ and $N2[,2] = (0001)$, therefore $M_{4,2} = 1$

when $i = 5$ and $j = 1$

$N1[5,] = (1010)$ and $N2[,2] = (0001)$, therefore $M_{5,2} = 0$

when $i = 1$ and $j = 3$

$N1[1,] = (1001)$ and $N2[,3] = (0110)$, therefore $M_{1,3} = 0$

when $i = 2$ and $j = 1$

$N1[2,] = (1000)$ and $N2[,3] = (0110)$, therefore $M_{2,3} = 0$

when $i = 3$ and $j = 1$

$N1[3,] = (0110)$ and $N2[,3] = (0110)$, therefore $M_{3,3} = 1$

when $i = 4$ and $j = 1$

$N1[4,] = (0001)$ and $N2[,3] = (0110)$, therefore $M_{4,3} = 0$

when $i = 5$ and $j = 1$

$N1[5,] = (1010)$ and $N2[,3] = (0110)$, therefore $M_{5,3} = 1$.

The resulting final matrix M_{ij} or M is then presented as follows:

$$\mathrm{M} = \begin{array}{ccccc} & \text{Cluster.1} & \text{Cluster.2} & \text{Cluster.3} & \text{Binary\_barcode} \\ \text{Isolate 1} & 1 & 1 & 0 & 110 \\ \text{Isolate 2} & 1 & 0 & 0 & 100 \\ \text{Isolate 3} & 1 & 0 & 1 & 101 \\ \text{Isolate 4} & 0 & 1 & 0 & 010 \\ \text{Isolate 5} & 1 & 0 & 1 & 101 \end{array}.$$

The binary barcode derived from matrix M is considered an independent observation (cross-sectional) because the data was independently generated for each isolate. Nevertheless, it is apparent that isolates may share certain characteristics, prompting considerations of evolutionary relationships or phylogeny among them. Most importantly, evolutionary biologists are keen on understanding the dynamic evolution of genetic traits, necessitating an exploration of the functional associations between isolates. The virtual representation of the functional relationships among the clusters, barcode, and isolates is illustrated in figure S2.

Following this, we apply our proposed Cluster-HyperHMM method to the binary barcode in matrix M, assuming independent relationships among the isolates (110, 100, 101, 010, 101). The outcomes are visualized in Figure S3, illustrating the order of acquisition of traits and the likely transition from one state to another.

Subsequently, we implement our proposed method Cluster-HyperHMM on the binary barcode in M by assuming evolutionary dynamic trajectories between the isolates (000-010, 000-100, 100-101, 100-110), and the results are displayed in figure S4.

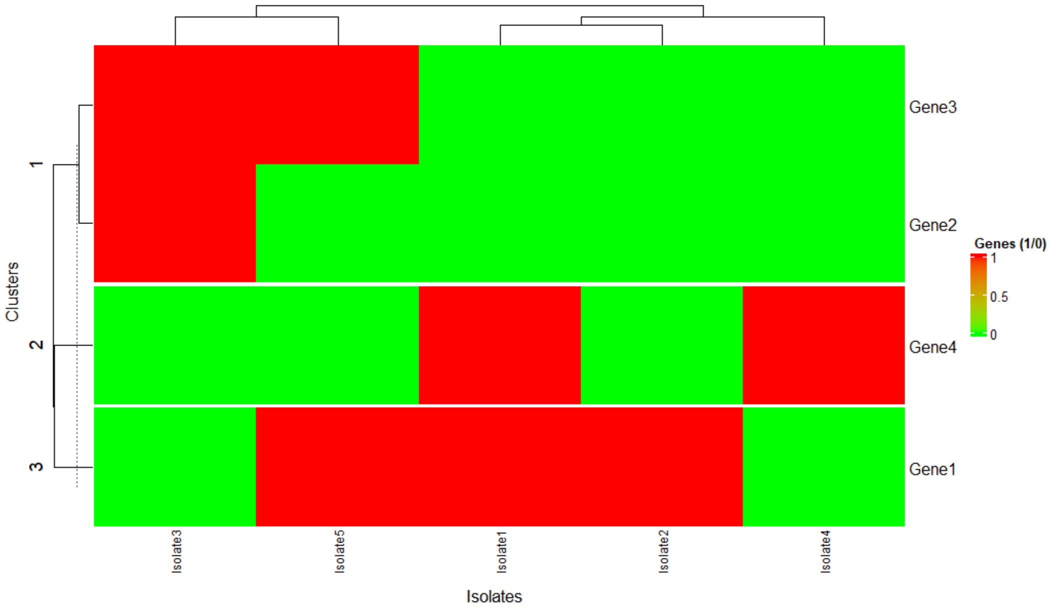


Figure S2:The clustering heatmap of the binary barcode

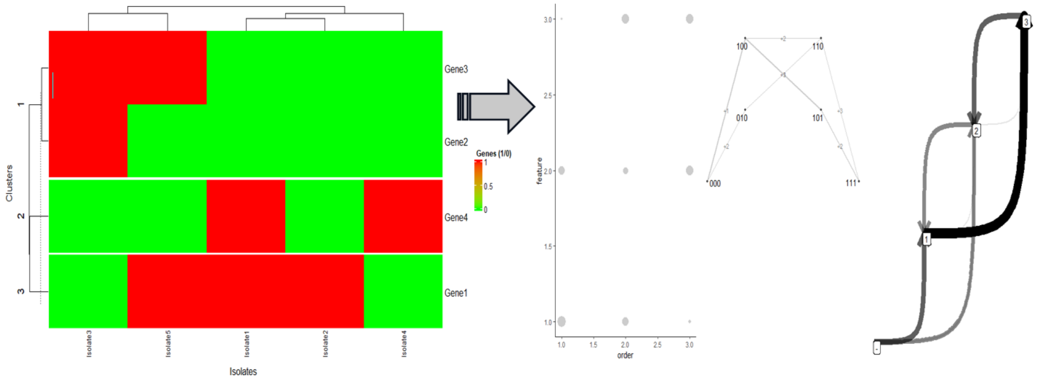


Figure S3: Illustration of Cluter-HyperHMM pathways associated with cross-sectionally observed data

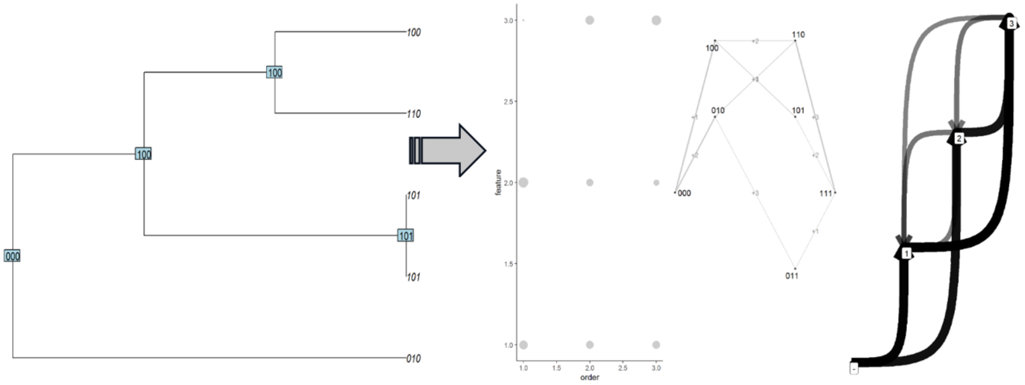
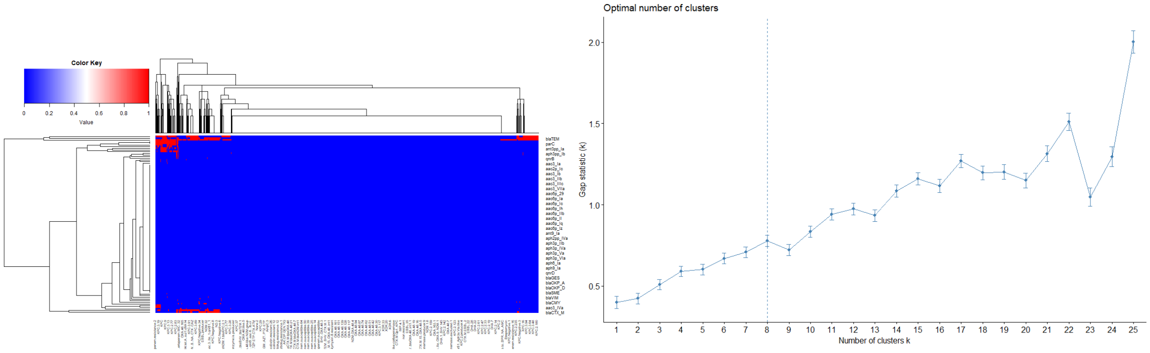


Figure S4: Illustration of Cluter-HyperHMM pathways associated with Phylogenetically observed data

A.2: Clustering Details

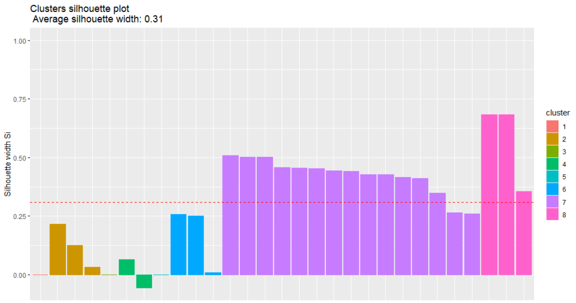


(a) Klebsiella pneumoniae raw data

(b) Optimum number of clusters

Cluster	Size	Gene Name
8	3	parC, parE, gyrA
7	15	qnrA, $aac3_{Ia}$, blaVEB, $ant2pp_{Ia}$, blaVIM, $ant3pp_{Ia}$, blaIMP, $aph3pp_{Ib}$, $aph6_{Id}$, qnrB, blaCARB, blaCMY, $blaCTX_M$, blaAMPC, blaNDM
6	3	$aac3_{IVa}$, $aph4_{Ia}$, gyrB
2	3	$aac6p_{Ib}, ant3pp_{Ia}, aph3p_{Ia}$
4	2	$blaCTX_M$, blaOXA
5	1	blaSHV
3	1	blaTEM
1	1	blaKPC

(c) Clusters and Genes

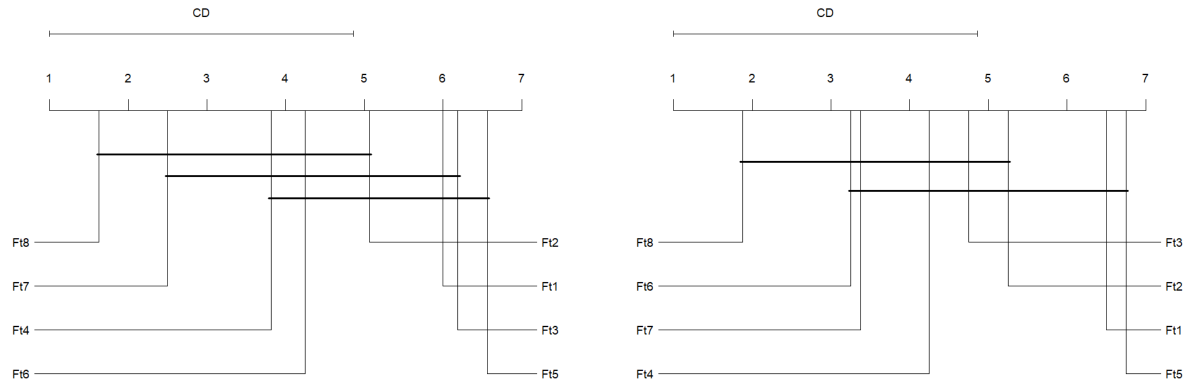


(d) Average Silhouette Width Cluster plot

Figure S5:Clustering details

A.3: Real-life Data: Evolutionary pathways of each content

6.1. Comparing cross-sectional and phylogeny Result



(a) Cross-sectionally observed

(b) Phylogenetically observed

Figure S6:Critical Difference (CD) diagrams showing the results of a statistical comparison of the performance of all the features from the Friedman-Nemenyi test. The solid bars are used to represent cliques. Within these cliques, there's an assumption that there are no significant differences between the features. The diagrams displayed information including the average ranking of each feature, where a rank of 1 signifies the method with the highest acquired feature. In **(a)** feature 8 is ranked first followed by 7, 4, and 6, wherein **(b)** feature 8 is ranked first followed by 6, 7, and 4.