Summer 8-8-2023

Bioinformatics Tools for RNA-seq Data Analysis

Akram Sadat Hosseini

Bioinformatics Tools for RNA-seq Data Analysis

by

Akram Sadat Hosseini

Under the Direction of Alex Zelikovsky, PhD

ABSTRACT

RNA-Seq is a recently developed approach to transcriptome profiling that uses deep-sequencing technologies. The availability of RNA-seq data encouraged computational biologists to develop algorithms to process the data in a statistically disciplinary manner to generate biologically meaningful results. Clustering viral sequences allows us to characterize the composition and structure of intrahost and interhost viral populations, which play a crucial role in disease progression and epidemic spread. In this research we propose and validate a new entropy based method for clustering aligned viral sequences considered as categorical data. The method finds a homogeneous clustering by minimizing information entropy rather than distance between sequences in the same cluster. Moreover in this research, we present a novel pathway analysis method based

on Expectation-Maximization (EM) algorithm to study the enzyme expression and pathway activity using meta-transcriptomic data. We will also discuss our approaches to generating unique gene signatures to understand the role of sensory nerve interference in the anti-melanoma immune response and study the racial disparity in Triple-negative breast cancer. Finally, we present our method to detect the retained introns in RNA-seq data to develop a vaccine against cancer having p53 mutations. In summary, this research provides novel approaches to exploring RNA-seq data and their application to real-world biological research.

INDEX WORDS: Categorical data, Clustering, Entropy, Monte Carlo algorithm, Viral genomic sequences, RNA SEQUENCING, EXPECTATION-MAXIMIZATION (EM), ENZYME EXPRESSION, Minimal spanning network' Genetic relatedness

Bioinformatics Tools for RNA-seq Data Analysis

by

Akram Sadat Hosseini

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

in the College of Arts and Sciences

Georgia State University

2023

Bioinformatics Tools for RNA-seq Data Analysis

by

Akram Sadat Hosseini

Committee Chair:	Alex Zelikovsky
Committee:	Pavel Skums
	Murray Patterson
	Ion Mandoiu

Electronic Version Approved:

Office of Graduate Studies

College of Arts and Sciences

Georgia State University

June 2023

DEDICATION

To my daughter

Nina

ACKNOWLEDGMENTS

Foremost, I would like to express my sincere gratitude to my advisor Dr. Alex Zelikovsky for the continuous support of my Ph.D study and research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I acknowledge Dr. Pavel Skums, Dr. Murray Patterson, and Dr. Ion Mandoiu for their advice and consultations. It was great to work with all the peers in our lab: Dr.Fil Rondel, Dr. Andrew Melnyk, Dr. Sergey Knyazev, Akshay Juyal, Bikram Sahoo, Daniel Novikov, Hafsa Farooq, Dr.Pelin Icer Baykal,and Fatemeh Mohebbi from GSU. I want to express my obligation to all professors and friends I met in the Computer Science Department, GSU.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

1.1 RNA-seq Data Analysis

the emergence of high-throughput next-generation sequencing (NGS) technologies and RNA-sequencing (RNA-Seq) has revolutionized RNA-based research and its applications in various fields. RNA-Seq has provided valuable insights into the transcriptome, enabling the detection of gene expression levels, alternative splicing events, transcript isoforms, gene fusions, single nucleotide variants (SNVs), and insertions/deletions. RNA-Seq technology has been widely adopted due to its ability to provide both quantitative and qualitative information about the transcriptome. The continuous improvement in RNA-Seq methodologies and the decreasing costs have further amplified its usage in diverse studies involving prokaryotes and eukaryotes. This has facilitated the exploration of different RNA species and their biological roles.

However, despite the advancements in RNA-Seq technology, there are challenges associated with data processing, storage, and retrieval that need to be addressed. Computational biologists and bioinformaticians play a crucial role in developing algorithms and data structures to handle these challenges and extract high-quality results from RNA-Seq data. The field of computational biology has made significant contributions to improving the analysis of RNA-Seq data. Researchers in this field continuously work on developing innovative solutions to overcome biases and limitations in data handling. They strive to enhance data processing pipelines, optimize storage and retrieval methods, and refine computational algorithms for accurate and efficient analysis.

Computational biologists and bioinformaticians must stay updated with the latest research and

development in the field. By keeping pace with the continuous advancements, they can address the challenges associated with RNA-Seq data and ensure the generation of reliable and unbiased results.

In summary, RNA-Seq has provided valuable insights into the transcriptome and has become an indispensable tool in various research areas. Computational biologists and bioinformaticians have played a key role in developing advanced algorithms and data structures to process, store, and retrieve high-quality results from RNA-Seq data. Continuous research and development in computational biology are necessary to address biases and improve data handling techniques in RNA-Seq analysis.

In this work, I address three important issues in RNA-seq data analysis.

- Identifying viral variants via clustering is essential for understanding the composition and structure of viral populations within and between hosts, which play a crucial role in disease progression and epidemic spread. The objective is to identify haplotypes in a massively inter-host viral population and utilize them as cluster centers in categorical clustering algorithms, such as k-modes, to identify subtypes. In the absence of ground truth, clustering entropy serves as the measure to assess the effectiveness of the clustering approaches.

- The unprecedented level of genome sequencing during the SARS-CoV-2 pandemic brought about the challenge of processing this genomic data. However, the state-of-the-art phylogenetic methods were mostly designed for analyzing data that are significantly sparser and require extensive subsampling of strains.

- Estimating metabolic pathway activity in planktonic communities using meta-transcriptomic data is challenging due to overlapping pathways and shared enzymes. Accurate pathway assignment through database integration is vital, while pathway enrichment analysis and network-based approaches provide insights into pathway dynamics. Integration with other omics data enhances understanding of complex metabolic processes.

1.2 Contributions

The dissertation describes the following contributions:

- We show that Monte Carlo clustering improves the reconstruction of intra-host viral populations from sequencing data.
- We show that eMSF can accurately detect transmission events and build a genetic network with significantly higher assortativity with respect to Continent and Country attributes of SARS-CoV-2 samples.
- A novel pathway analysis method based on Expectation-Maximization (EM) algorithm to study the enzyme expression and pathway activity using meta-transcriptomic data.

1.3 Articles

1. **Roya Hosseini** , Akshay Juyal , Daniel Novikov1, Mark Grinshpon2, and Alex Zelikovsky. Entropy Based Clustering of Viral Sequences.
2. **Roya Hosseini** , Akshay Juyal , Daniel Novikov1,Mark Grinshpon2, and Alex Zelikovsky. Reconstruction of Viral Variants via Monte Carlo Clustering.

3. Rondel, F.M., **Hosseini, R.**, Sahoo, B., Knyazev, S., Mandric, I., Stewart, F., Măndoiu, I.I., Pasaniuc, B., Porozov, Y. and Zelikovsky, A., 2021. Pipeline for Analyzing Activity of Metabolic Pathways in Planktonic Communities Using Metatranscriptomic Data. Journal of Computational Biology, 28(8), pp.842-855.

4. Rondel, F., **Hosseini, R.**, Sahoo, B., Knyazev, S., Mandric, I., Stewart, F., Măndoiu, I.I., Pasaniuc, B. and Zelikovsky, A., 2020, December. Estimating Enzyme Participation in Metabolic Pathways for Microbial Communities from RNA-seq Data. In International Symposium on Bioinformatics Research and Applications (pp. 335-343). Springer, Cham.

5. F. M. Rondel, **R. Hosseini**, H. Farooq1, B. Bello, A. Juyal, S. Knyazev, B. Pasaniuc, S. Mangul, A. S. Rogovskyy, and A. Zelikovsky. Estimating enzyme expression and metabolic pathway activity in Borreliella-infected and uninfected mice.

6. Melnyk, A., Mohebbi, F., Knyazev, S., Sahoo, B., **Hosseini, R.**, Skums, P., Zelikovsky, A. and Patterson, M., 2020, December. Clustering based identification of SARS-CoV-2 subtypes. In International Conference on Computational Advances in Bio and Medical Sciences (pp. 127-141). LNCS

7. Melnyk, A., Mohebbi, F., Knyazev, S., Sahoo, B., **Hosseini, R.**, Skums, P., Zelikovsky, A. and Patterson, M., 2021. From alpha to zeta: Identifying variants and subtypes of sars-cov-2 via clustering. Journal of Computational Biology, 28(11), pp.1113-1129.

8. Knyazev, S., Novikov, D., Grinshpon, M., Singh, H., Ayyala, R., Sarwal, V., **Hosseini, R.**, Baykal, P.I., Skums, P., Campbell, E. and Mangul, S., 2021, November. A Novel Network

Representation of SARS-CoV-2 Sequencing Data. In International Symposium on Bioinformatics Research and Applications (pp. 165-175).

CHAPTER 2

Entropy Based Clustering of Viral Sequences

2.1 Introduction

Clustering viral sequences allows us to characterize the composition and structure of intrahost and interhost viral populations, which play a crucial role in disease progression and epidemic spread. For intrahost populations, clustering allows us to detect the distinct viral variants present in the patient, including minor low-frequency variants, which can cause immune escape, drug resistance, and an increase of virulence and infectivity[4,17,25,31,58,10,62]. Furthermore, such minor variants are often responsible for transmissions and establishment of infection in new hosts[12,27,64].

In this paper we propose a Monte Carlo entropy minimization method for clustering of viral sequences considered as categorical data. The method finds a homogeneous clustering by minimizing information entropy rather than distance between sequences in the same cluster. We discuss advantages and disadvantages of both entropy and distance based approaches, and further validate the meaningful information content extracted by entropy based clustering. We demonstrate that the proposed method is stable, moving towards the same minimal entropy configuration across multiple runs. We also show that it is fast and scalable to hundreds of thousands of sequences.

By clustering viral populations across different hosts, we determine major strains of closely related viral samples, which is helpful for tracking transmissions and informing public health strategies[5]. For transmission tracking, clustering can identify the source of an outbreak and whether the source is present in the sampled population. It can also determine whether two viral samples belong to the same outbreak, and whether one infected the other[50]. Therefore, using clustering to

obtain an accurate characterization of viral mutation profiles from infected individuals is essential for viral research, therapeutics, and epidemiological investigations.

Viral sequences are strings from a fixed nucleotide alphabet, and hence they can be viewed as vectors of categorical data. In the best possible clustering, sequences in each cluster will be as homogeneous as possible in each site. Typically, this is achieved by minimizing the Hamming distances between sequences within the same cluster or the distances to the cluster's consensus[15]. However, Hamming distance carries the implicit assumption that all mutations at all sites are of equal cost, and also it does not consider the distribution of values in a given category, counting each mismatch equally.

In this paper we propose to use entropy based clustering for viral sequences. Entropy considers the distributions of nucleotides in each site, allowing us to capture different kinds of mismatches. Minimizing entropy instead of distance also avoids the need to introduce the abstraction of equal transition costs, implicit in Hamming distance. Thus, entropy as an objective for clustering makes fewer assumptions on the data and is more informative for clustering categorical data.

We have applied our entropy based clustering method to the viral sequencing data. The unprecedented effort in sequencing its genome has created vast databases of sequences, such as GISAID[39]. Clustering techniques can provide new insights into the evolution of the virus, assist with phylogenetic and phylodynamic analyses, and offer new tools for constructing transmission networks to help with understanding the spread of the pandemic[5].

We validate effectiveness of the entropy based clustering of viral sequences on real datasets. We measure the information content extracted from the sequences by the resulting clustering. We also

demonstrate that our method converges to the same minimum-entropy clustering across different runs, thus marking stability of the method. Finally, we describe a tag selection procedure, which selects the highest entropy sites to represent sequences leading to a significant decrease in runtime without major loss of information.

2.2 Methods

In this section, first we define the entropy and Hamming distance of clustering of a set of aligned viral sequences. We use each of these two measures as an objective function to be minimized for clustering. Then we describe a Monte Carlo clustering algorithm, which is a modification of an algorithm proposed in[45]. Finally, we describe a tag selection preprocessing step, which significantly reduces the runtime of the algorithm.

2.2.1 Entropy Based Clustering of Viral Sequences

Entropy of a category across a set of categorical vectors quantifies the heterogeneity of values in the category. Entropy is low when a single value is highly frequent, and it is at its highest when all values are equally frequent in the category. Since we are treating viral sequences as vectors of categorical data, the categories here are sites along the sequence, and their values are from the nucleotide alphabet. A clustering with minimal entropy will have the highest possible homogeneity of nucleotides in each site for sequences in the same cluster.

Formally, we have a set S of aligned nucleotide sequences on a set X of genomic sites. Since the sequences $s \in S$ are aligned, they can be viewed as rows of a matrix, and the sites $x \in X$, can be viewed as columns of this matrix. Let the alphabet be the four nucleotides, not counting the gap

() character. Following[45], the entropy $H(C_x)$ of a site $x \in X$ in cluster C is defined as

$$H(C_x) = -\sum_{s \in C} \sum_{a \in} p(s_x = a) \cdot \log p(s_x = a). \tag{2.1}$$

Note that $p(s_x = a)$, the probability that a sequence $s \in C$ has nucleotide $a \in$ at site x, essentially amounts to the *relative frequency* of the nucleotide a in cluster C at site x (ignoring gap characters).

The entropy $H(C)$ of a cluster C of viral sequences on a set X of sites is then defined as

$$H(C) = \sum_{x \in X} H(C_x), \tag{2.2}$$

that is, we simply sum up the entropies at the individual sites.

Finally, given a clustering of the set S, the *entropy* of is defined as follows:

$$H() = \sum_{C \in} \frac{|C|}{|S|} \cdot H(C) = \frac{1}{|S|} \sum_{C \in} |C| \cdot H(C). \tag{2.3}$$

In other words, the entropy of clustering is the sum of cluster entropies weighted by their relative sizes.

In[45], the authors prove that the entropy defined in equation (2.3) is a convex function, allowing any optimization procedure to reach a global minimum. It is because of this property that we can use techniques aimed directly at minimizing clustering entropy as the objective.

2.2.2 Hamming Distance Based Clustering of Viral Sequences

Similarly, we define a different clustering objective as Hamming distance (HD) instead of entropy. This objective is the sum of the Hamming distances from each sequence s to the consensus of the cluster containing s.

Formally, for a cluster C and for a site $x \in X$, the Hamming distance from the consensus letter in this cluster at this site is

$$HD(C_x) = \sum_{a\in} C_x(a) - \max_{a\in} \{C_x(a)\} \tag{2.4}$$

where $C_x(a)$ is the number of occurrences of the letter $a \in$ in site x in cluster C.

Then the Hamming distance $HD(C)$ of a cluster C of viral sequences on a set X of sites is defined as

$$HD(C) = \sum_{x\in X} HD(C_x), \tag{2.5}$$

and the *Hamming distance* of the clustering is defined as

$$HD() = \sum_{C\in} HD(C). \tag{2.6}$$

2.2.3 Algorithm Description

In general, Monte Carlo methods optimize an objective by attempting random changes and accepting a change only if it improves the objective. In our case, the objective is to minimize either the clustering entropy or the clustering Hamming distance, defined in subsections 2.2.1 and 2.2.2 above. A trial step consists of moving a randomly selected sequence to a different randomly selected cluster, and accepting the move only if the objective function is reduced.

Algorithm 1, Monte Carlo based clustering, implements this approach, with several modifications intended to improve its runtime and the quality of its outputs.

The algorithm takes an existing clustering as its starting point. In our experiments, we use clusterings generated by the CliqueSNV tool[43] from the datasets described in subsection 2.3.1.

Algorithm 1 Monte Carlo based clustering

Input:
Initial clustering (by default, from CliqueSNV)
Number of rejected moves: I (by default, $I = 800$)
Relative difference: K (by default, $K = 0.00001$)
Output:
Clustering with reduced entropy or Hamming distance
Initializations:
Compute nucleotide counts for each column in each cluster
Compute entropy (resp., Hamming disatnce) for each cluster and H, clustering entropy (resp., clustering Hamming distance)
Initialize number of rejected moves $T = 0$
Iteration:

while $T \leq I$ **do** Pick a random sequence s
Move s from its cluster A to a randomly selected cluster B, $B \neq A$
Update the nucleotide counts for A and B
Compute Δ, overall entropy (resp., Hamming distance) reduction after moving s from A to B
$\Delta/H \geq K$ Accept the move
$H = H - \Delta$
$T = 0$ $T = T + 1$
Move s from B back to A

We supply such clustering as an input to the algorithm in order to generate a new clustering with reduced clustering entropy $H()$ or reduced Hamming distance $HD()$. Additional inputs to the algorithm are two parameters I and K, where I defines the number of consecutive rejected trials before stopping, and K defines the relative objective function reduction threshold for accepting a move.

To achieve the goal of finding a new clustering with reduced entropy or Hamming distance as the objective, the algorithm applies Monte Carlo optimization by repeatedly trying to move a randomly selected sequence from its current cluster to another randomly selected cluster; any such move is accepted only if the relative improvement to the objective function is higher than the

threshold value K.

In the initialization phase, lines 1–3 of the algorithm, it starts by computing nucleotide counts for each column in each cluster, which are then used to compute the values of the entropy and the Hamming distance for each cluster, as well as the overall clustering entropy or Hamming distance, H.

2.2.4 Tag Selection

To improve runtime, we apply a preprocessing tag selection step that allows us to represent sequences by a smaller subset of sites. Preferring tags with highest variability, the procedure chooses the n sites with highest entropy, where n is some predefined value. Then clustering proceeds with each sequence now of length n corresponding to the selected sites.

2.3 Settings for Validation of Clustering Methods

We validate entropy based clustering by estimating improvement over an existing clustering technique. To that end, we apply this Monte Carlo based algorithm to the clustering obtained by the CliqueSNV tool[43].

2.3.1 Datasets

For validation, we use two of the datasets of sequences that were also used by Melnyk et al in[52]. For both datasets, an initial clustering was obtained by the CliqueSNV-based method.

D1: This dataset includes all sequences submitted to the global GISAID viral database[39] from the beginning of the pandemic up until the beginning of March 2020. It consists of 3688 aligned

sequences, all sequences 29891 nucleotides long. CliqueSNV produced an initial clustering of this dataset consisting of 28 clusters.

D2: This dataset includes all sequences submitted to the UK-based EMBL-EBI database from the end of January 2020 to the end of December 2020[19]. It consists of 148000 aligned sequences, all sequences 29903 nucleotides long. CliqueSNV produced an initial clustering of this dataset consisting of 15 clusters.

2.3.2 Tag Selection Effects on Runtime

To measure the effects of tag selection on runtime, the proposed method was run on the D1 dataset of 3688 sequences, using the initial clustering generated by CliqueSNV as a starting point. The tag selection procedure was employed to produce four subdatasets consisting of the same sequences, but of reduced lengths of 100, 1000, 3000, and 5100 tags. We expect that the Monte Carlo method, when applied to a dataset consisting of shorter sequences, will be able to take more trial steps in the same amount of time, and thus reduce its clustering entropy quicker. The total number of SNPs in the input data was exactly 5100; all other positions did not mutate. Thus, this largest number of tags contains information equivalent to the full length sequences for the purposes of clustering.

The program was run on each length of sequences. Every hour, the current clusterings of each run were evaluated by their entropy on the full-length sequences. These hourly entropy values are shown on an entropy-over-time graph, Figure 3.5, which compares the speed of entropy reduction for different sequence lengths.

2.3.3 Discerning Signal from Noise with Monte Carlo Based Clustering Optimization

We estimate the amount of meaningful information extracted by the clusterings obtained by our entropy minimization method. To distinguish between sample-specific noise and meaningfully extracted information, we run our method on a perturbed version of the input with the same starting entropy. For this experiment we use the D1 dataset with 3688 sequences, alongside the initial clustering from CliqueSNV of 28 clusters.

The permutation procedure is as follows. Within each cluster, every site is shuffled into a random permutation. Importantly, by respecting clusters during permutation, the initial nucleotide frequencies within each site in each cluster stay the same. Thus, the permuted input has the same starting entropy as the original input. What changed is the haplotypes being clustered.

We run the program on both of these inputs for exactly 100000 Monte Carlo trials each, accepting all moves that reduce entropy. We compare the resulting entropy reductions between the two runs. Any entropy reduction present in the permuted data is sample-specific noise extracted by our method, while the difference in resulting entropies between the original and permuted inputs corresponds to the amount of meaningful information extracted by our method.

2.3.4 Stability of Optimized Clustering

Now we evaluate the robustness of our method against slight permutations of the input data as well as changes in random seed. Rather than completely shuffling each site as in 2.3.3, we only shuffle a small percentage p of nucleotides at each site. We still respect clusters when permuting the data, to ensure that nucleotide frequencies in each site in each cluster remain unchanged.

We chose two values of p to create slightly permuted data sets for validation, $p = 1\%$ and $p = 5\%$, to be compared with the original data with 0% permutation. For each of the three datasets (two permuted and one original), we run our method three times, on two different objectives: first minimizing entropy, and second minimizing Hamming distance between sequences and their cluster consensus. As a result, for each degree of permutation and for each Monte Carlo objective, we obtain three minimum entropy clusterings.

The Rand index, measuring the degree of agreement between two clusterings, is measured between the initial clustering and all resulting clusterings, to get a sense of how far away the resulting clusterings have moved from the initial one under varying degrees of permutation. Further, we also measure the Rand index between the resulting clusterings, to determine whether the proposed method converges to similar clusterings across multiple runs.

2.4 Validation Results

We ran the proposed method on the cluster hardware consisting of 128 cores IntelXeonCPU E7-4850 v4 CPU @ 2.10GHz, with 3 TB of RAM, running Ubuntu 16.04.7 LTS.

We ran our entropy based Monte Carlo clustering algorithm on different selections of tags (selected based on highest entropy contribution). Figure 3.5 shows the results for different numbers of tags. When the number of tags is 1000, using tag selection yields better results of reducing entropy for some time durations (up to 3 hours). But in general, the effect of using tags on reducing entropy is not significant.

Compared to previous work[52], we were able to reduce the runtime by 95.83%. For accom-

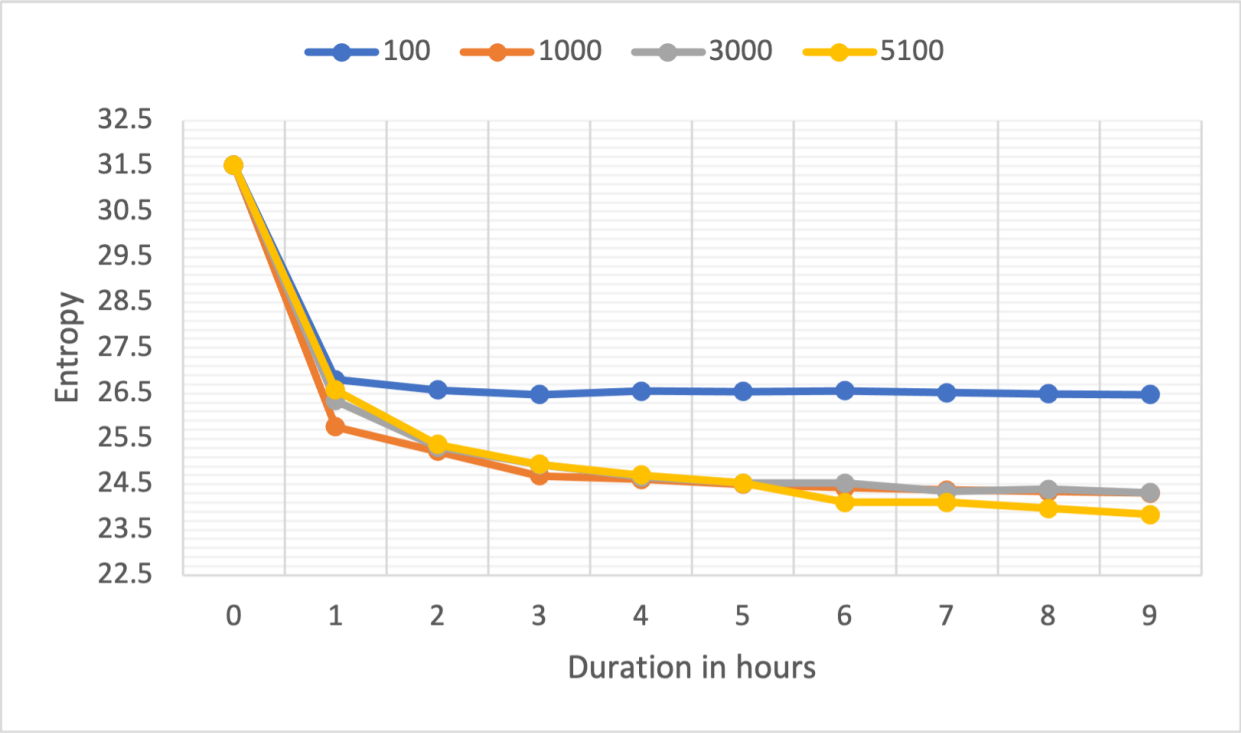

Figure 2.1 Entropy reduction over runtime for different numbers of selected tags. After 1 hour, the 1000 tags representation was able to reach the lowest entropy. The maximum number of tags, 5100, corresponds to all SNP positions present in the input data. All other sites were homogeneous. Homogeneous sites have zero entropy and thus can be ignored. Therefore, the yellow line corresponds to using all sites.

plishing this, we initially stored the counts of each nucleotide across a given tag in a cluster of sequences.

After further improving our entropy based Monte Carlo clustering algorithm and implementing parallel computing, we made it scalable for running on large data sets. This version performs 12K–13K iterations on average per hour, which is approximately **10** times faster than all our previous implementations.

Interestingly, after running the algorithm for **83K** iterations on the D2 dataset (which contains 148000 aligned sequences each 29903 nucleotides long) originally distributed across 15 clusters,

we came to the conclusion that in instances where sequence data is clustered into a smaller number of clusters there are some clusters where sequences are more dense than in others. For instance, after this run, in the output clustering the biggest cluster consisted of 34995 sequences, while the smallest had only 3571 sequences.

Therefore, moving one sequence at a time is not always beneficial for overall entropy reduction. We tried to tackle this problem by updating our move acceptance threshold to accept even smaller positive changes to entropy, from our previous relative difference threshold of $K = 10^{-5}$ to $K = 10^{-7}$. Although this change made our algorithm run for many more iterations before stopping (recall that the algorithm stops when it reaches the stopping threshold of $I = 800$ unsuccessful moves), the reduction to the overall entropy was still very small.

We believe that moving similar sequences together within clusters rather than moving just one could be a possible way of overcoming the problem we face here.

2.4.1 Picking signal over noise in clustering

By minimizing entropy on the permuted data, we find that the method reduces entropy to 29.6, while on the original, unshuffled data the method reaches a much lower entropy of 24 (see Table 2.1). This difference in entropies, $29.6 - 24 = 5.6$, of resulting clusterings accounts for the amount of meaningful information, which is not noise, that our method was able to extract from the real data.

Entropy MC reduced					Hamming distance MC reduced				
Initial	Original		Permuted		Initial	Original		Permuted	
31.524	Avg	Min	Avg	Min	1008.41	Avg	Min	Avg	Min
	24.77	24.7	29.62	28.65		373.14	369.61	770.39	689.97

Table 2.1 Results after running Monte Carlo for 1000 tags selected in decreasing order of entropy across 100 datasets obtained by applying the random permutation procedure described in section 2.3.3 to the D1 dataset 100 times. *Average Iterations in Monte Carlo*: 53804.39. *Average successful moves*: 615.31.

2.4.2 Stability of Monte Carlo Output

Table 2.2 shows the results of stability validation, in which we compare clustering similarity for varying degrees of permutation of the input data (see subsection 2.3.4).

The first column compares resulting clusterings to the initial clustering. Without any permutations, the resultant clustering moves significantly further away from the initial one, giving a Rand index of 0.93. As the permutation degree increases, we observed that the clusterings produced by the Monte Carlo algorithm do not move as far away from the initial clustering; in other words, even after Monte Carlo was applied, the resulting clusterings had high degree of agreement with the initial clustering.

The second column in Table 2.2 gives the average Rand index between multiple runs of Monte Carlo for a given permutation. We see that for all degrees of permutation the method stably converges towards similar clusterings, with Rand index scores of 0.97–0.98. The same trends can be observed when using Hamming distance to cluster consensus as the objective.

2.4.3 Results for Large Datasets

Running our entropy based Monte Carlo method on the large dataset D2, which consist of 143000 aligned sequences, we get the initial entropy for our initial clustering from CliqueSNV of 80.2750171. After 82786 iterations, the final entropy for the resultant clustering was 79.509444, and the total

% permutation	Cluster similarity(Rand index)			
	Entropy		Hamming	
	With original	With runs	With original	With runs
0	0.936476	0.970195	0.936114	0.970135
1	0.978898	0.979435	0.936126	0.970374
5	0.980458	0.980688	0.936180	0.970616

Table 2.2 Clustering similarity (Rand index) across three choices of degree of permutation. The proposed method was run three times for each permuted instance, each run consisting of $100,000$ Monte Carlo trials. Reported are average Rand index similarity of the resulting clusterings to the initial clustering, as well as between resulting clusterings.

runtime for this was 9 hours 15 minutes.

2.5 Conclusions

We have developed a scalable method to find minimum-entropy clusterings of datasets viral genomic sequences. The method is scalable to hundreds of thousands of sequences, and is made even faster without significant loss of accuracy by picking a subset of tags with maximum entropy to represent the sequences. We estimate the amount of meaningful information extracted by the method. We also show that our method converges toward similar minimum-entropy clusterings across multiple runs, demonstrating its stability.

For future directions, we believe the Monte Carlo entropy minimization approach can be improved by using simulated annealing, whose tolerance of suboptimal moves can allow us to escape local minima. We are also going to add Monte Carlo entropy minimization method to CliqueSNV's clustering of intrahost populations.

CHAPTER 3

A Novel Network Representation of SARS-CoV-2 Sequencing Data

3.1 Introduction

The tendency of RNA viruses to rapidly change their genomes is the major reason for vast inter-host and intra-host viral diversity and fast viral evolution[59]. This evolution can be tracked with high precision thanks to the rapid growth of capacity for viral genome sequencing that has been occurring over the past decade[32]. To see what that data can reveal about viral evolution and transmission, numerous analytical methods have been proposed[20,3,57,11,26,63,70,41,49,42,51,7]. One of the essential tasks in analysis of viral genomic data is representation of genetic relatedness between viral samples. For this purpose, standard phylogenetic methods as well as network-based methods that were initially applied only to specific viruses such as HIV[56] and HCV[46] have been proposed for SARS-CoV-2 analysis too.

The standard approach for representation of viral genomic data is based on phylogenetic trees. In general, finding an optimal phylogenetic tree that fits a biologically relevant model of evolution, e.g., using maximum likelihood approaches, is an NP-hard problem[22]. As a result, tools for phylogeny reconstruction are too slow or inaccurate on large datasets. Indeed, the quality of advanced phylogeny tools' outputs significantly decreases with the growth of the number of input sequences. For this reason, these methods rely heavily on subsampling to maintain the acceptable quality of phylogeny reconstruction.

Network-based methods, designed to represent the most likely pairs of connected viral genomes rather than viral evolution, offer an alternative approach to phylogenetic tree reconstruction[20,3].

This is convenient because establishing genetic relatedness between viral samples is the basic step in viral outbreak investigations. Furthermore, network-based methods are more promising in the era of big genomic data because they are much more simple and scalable than phylogenetic methods. The success of using networks-based method for HCV and HIV outbreak investigations[9,6,1,2] motivated the Centers of Disease Control and Prevention (CDC) to adopt them for wider use.

Currently, the CDC is actively advancing the following two alternative approaches for tracing genetic relatedness:

1. The Global Hepatitis Outbreak and Surveillance Technology (GHOST) is a cloud-based system that allows users to analyze and visualize data regardless of computational expertise. It uses the intra-host viral haplotypes for tracing HCV epidemics, the Hamming distance between genomic sequences as a metric for genetic relatedness, and k-step networks (introduced as minimal spanning networks in[20]) for choosing genetically related vertices[9,13,46].

2. For HIV outbreaks, the current tool of choice is HIV-TRACE, which performs high-scale analysis of genomes in HIV surveillance systems[56,29]. It identifies groups of closely related genomes using the Tamura-Nei 93 (TN93) genetic distance metric[68]. HIV-TRACE accepts either a Sanger sequence or a consensus sequence from NGS sequencing experiment for each individual. These sequences are then used to look for evidence of relatedness between them. Relatedness is suspected when the genetic distance between sequences is below a certain threshold (we will refer to this construction as τ-networks). This simple approach has demonstrated high reliability for detecting rapidly growing outbreak clusters[55].

In this paper we introduce a novel network-based method for constructing genetic similarity networks, which generalizes both GHOST and HIV-TRACE.

First, we present , an algorithm that generalizes the methodology implemented in GHOST. Given a set of genomic sequences along with distances between them, builds a network that includes the union of all minimal spanning trees (MST), as well as some additional edges: those that were close to being included in an MST, i.e., edges whose weights are just a little (within a specified parameter ε) more than the largest edge weight in a path connecting its endpoints in an MST.

We further improve by incorporating the τ-networks of HIV-TRACE. This second tool, , effectively builds an first, but then removes all edges that are too heavy, i.e. all edges whose weight is more than a specified parameter τ.

We compare our tool with minimum spanning networks (MSN) and τ-networks[28,56] and demonstrate that and results are of better quality in terms of attribute assortativity, recall, and precision. Our method allows us to construct phylogeny networks and report results for SARS-CoV-2 sequence datasets having up to more than hundred thousand strains, with a potential of being scalable to much larger datasets.

3.2 Methods

When analyzing evolutionary relationships in a set of genomes, the first step may be to create the complete graph of the genomes and measure the distances between them. Such a complete graph incorporates all those relationships that we would like to see, but it has too many edges,

making it impossible to discern the useful information from it. In fact, as a complete graph, it contains all possible edges, including those between really distant genomes, i.e., edges whose weights represent very long genetic distances. To extract edges connecting related genomes only, we introduce a threshold parameter τ and remove edges with length greater than τ, thus getting rid of edges that are too long.

Another idea for improving the graph is to use minimal spanning trees (MST). Generally, the standard greedy algorithms for building MST's always pick the shortest of available edges towards the closest neighbor genomes. However, if one neighbor of a vertex (genome) is just slightly closer than some others, the MST would include the shortest edge only and leave the rest out, even though all of them have a close enough genetic relationship with each other. Thus we introduce another parameter ε that allows us to include edges that are only slightly longer than the closest neighbors.

In summary, our goal is to build a graph $G = (V, E)$, where each vertex $v \in V$ represents a viral genomic sequence, and where an edge $e \in E$ connects two vertices u and v whenever u and v represent genetically related viral genomes. Previous studies proposed to take G as the union of all possible MST's in the weighted graph whose nodes are the viral genomes and whose edges are weighted by a genetic distance between these genomes[20,9]. We extend this approach and propose to include some additional edges.

[] Given an $\varepsilon \geq 0$, is a graph in which two vertices u, v are connected if $d(u, v) \leq (1 + \varepsilon) \cdot d(x, y)$, where $d(x, y)$ is the weight of the heaviest edge on the u-v path in an MST.

An efficient algorithm for constructing is given in Algorithm 2.

Following the τ-network methodology, we allow setting a threshold τ for additionally filtering

out edges that are too long.

[] Given an $\varepsilon \geq 0$ and a $\tau > 0$, is a graph in which two vertices u, v are connected if they are connected in and $d(u, v) \leq \tau$.

is a generalization of both the MSN and τ-network methods. Indeed, MSN is a special case of if we set the parameters to $\varepsilon = 0$ and $\tau = \infty$. Similarly, τ-network is a special case of when $\varepsilon = \infty$.

An implementation of the algorithm as a software tool is freely available on GitHub at `https://github.com/Sergey-Knyazev/eMST`. The software can accept sequences in FASTA format and compute using either of the two genetic metrics of choice, Hamming distance or TN93. The user can also provide their own distance matrix in the list of edges format.

For efficiently computing Hamming distance between sequences, we implemented the following speed up technique[54]. Initially, we infer a consensus of all input sequences. Then, for each sequence in the input, we determine a set of positions where each sequence has mutated from the consensus. Finally, for each pair of sequences the Hamming distance is computed in two steps. First, we initialize the value of Hamming distance to be the size of the symmetric difference between the two sets. Second, for each position in the intersection of the two sets, we check if the sequences differ at this position, and if they do, we increment the value of Hamming distance by one.

3.3 Results

To demonstrate the usability of the methodology, we benchmarked it against other methods.

Algorithm 2 ()

MSA: Multiple Sequence Alignment of Strains
G: Fully connected distance graph obtained from strains
ε: $\varepsilon \geq 0$, parameter for
function ADDEPSILONEDGES(MST, $LongestEdge$, E, ε)
    for $(x, y) \in E$ **do**
        if $d(x, y) \leq (1 + \varepsilon) \cdot (LongestEdge(x, y))$ **then**
            `add` (x, y) `to` MST
    return MST
function GETLONGESTEDGES(MST, E)
    for $(x, y) \in E$ **do**
        $LongestEdge(x, y) \leftarrow \max(e_i)\ \forall e_i \in MST_{x \to y}$
    return $LongestEdge$
procedure EMSN($A = MSA$ or G, ε)
    `If` $A = MSA$, `obtain` $G(V, E)$ `using a distance metric (e.g., Hamming, TN93, etc)`
    $MST \leftarrow getMST(G)$
    $LongestEdge \leftarrow getLongestEdges(MST, E)$
    $eMST \leftarrow addEpsilonEdges(MST, LongestEdge, E, \varepsilon)$
=0

First, we compared and with the two state-of-the-art methods for constructing τ-networks (used in HIV-TRACE) and minimum spanning networks (used in GHOST) on COVID-19 sequences available from GISAID using assortativity analysis.

Second, we examined , , τ-networks, and MSN on their ability to infer transmission events and compared them with other available tools including CS-phylogeny[65], NETWORK5011CS[23], RAxML[66], outbreaker[8], and phybreak[40]. For this test, we used a SARS-CoV-2 sequencing dataset with known ground truth about infective transmission events, and we measured precision and recall of each of the methods when applied to infer these events from the sequencing data.

Third, we showed the scalability potential of to process networks up to a size of more than hundred thousand of sequences.

3.3.1 Datasets

1. For comparison of the methods via assortativity analysis, we used the coast-to-coast (C2C) dataset, which contains 168 SARS-CoV-2 sequences collected from different countries, including 9 sequences from COVID-19 patients identified in Connecticut[21]. Each sample in this dataset has geographical attributes named Continent, Country, and Division.

2. For comparison of precision and recall of the methods in inferring transmission links we used the Early Transmission Links (ETL) dataset, which consists of 293 global SARS-CoV-2 sequences collected before March 9th, 2020. Each sequence has a known country of origin. This dataset was constructed to match the 25 known country-to-country transmission links that were collected from news articles detailing transmissions prior to the pandemic declaration, in the MIDAS 2019 Novel Coronavirus Repository[65].

3. For scalability analysis, we created datasets consisting of the initial 100, 200, 500, 1000, 2000, 5000, 10000, 20000, 50000, and 100000 SARS-CoV-2 sequences from the masked multiple sequence alignment from GISAID. To generate these datasets, we ordered sequences by date, and picked the earliest date when the number of sequences exceeds the desired number.

3.3.2 Assortativity analysis

We ran MSN, , τ-network, and on the C2C dataset using TN93 as measurement of genetic relatedness, and we evaluated attribute assortativity for continent, country, and division.

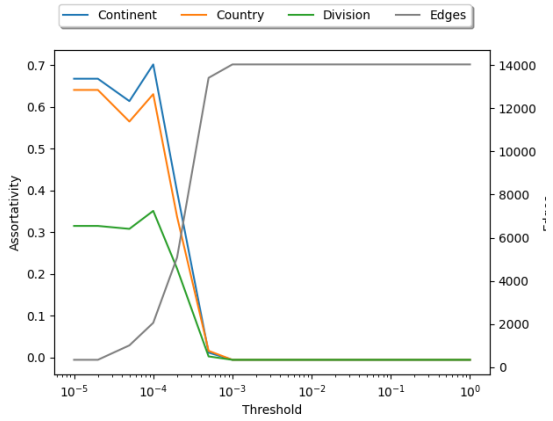

Figure 3.1 Attribute assortatitivity on the C2C dataset for different values of edge threshold τ, using τ-network with TN93 distance.

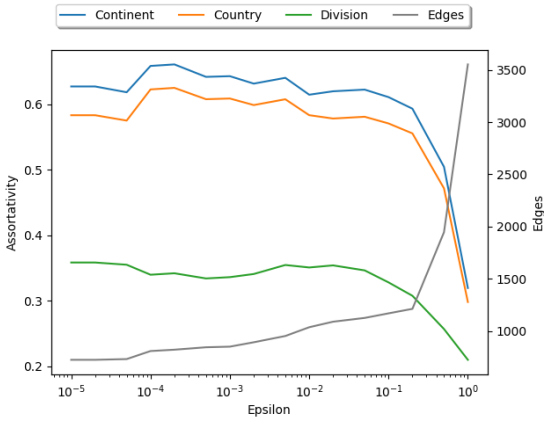

Figure 3.2 Attribute assortativity on the C2C dataset for different values of ε, using with TN93 distance and edge threshold $\tau = \infty$.

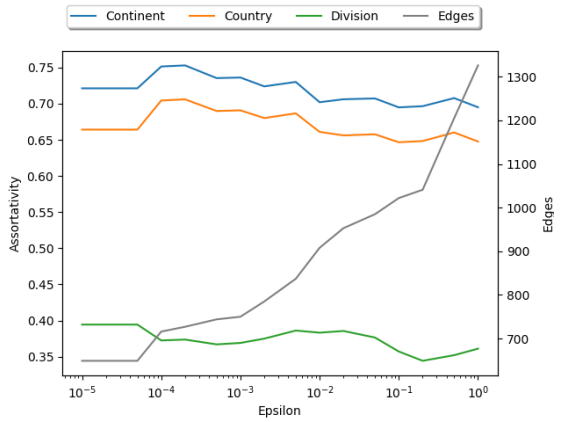

Figure 3.3 Attribute assortatitivity on the C2C dataset for different values of ε, using with TN93 distance and edge threshold $\tau = 0.0001$. The maximum assortativity occurs when $\varepsilon = 0.0002$.

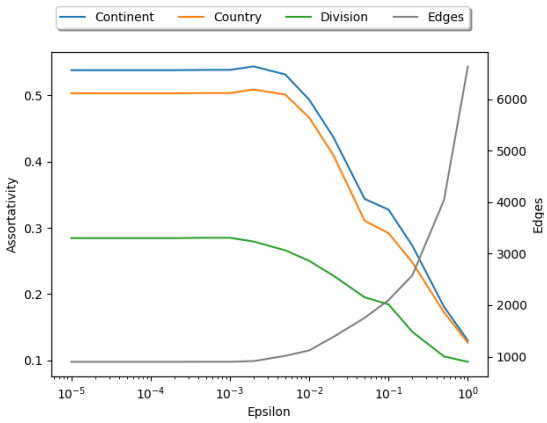

Figure 3.4 Attribute assortatitivity on the C2C dataset for different values of ε, using with Hamming distance.

Figure 3.1 shows the dependence of attribute assortativity on $\tau \in [0, 1]$ for τ-network. The optimal value for continent assortativity is 0.702 when τ is 0.0001. Figure 3.2 shows the dependence of attribute assortativity on $\varepsilon \in [0, 1]$ for . The optimal value for continent assortativity is 0.661 when ε is 0.0002. Figure 3.3 shows the dependence of attribute assortativity on $\varepsilon \in [0, 1]$ for , with fixed threshold $\tau = 0.0001$ that maximized assortativity in the τ-network analysis from Figure 3.1. The optimal value for continent assortativity is 0.7573 when ε is 0.0002. Figure 3.4 shows the dependence of attribute assortativity on $\varepsilon \in [0, 1]$ for with Hamming distance instead of TN93. The optimal value for continent assortativity is 0.546 when ε is 0.002.

Table 3.1 shows that the maximum assortativity on the C2C dataset was achieved by 's mixture of both parameters, with $\varepsilon = 0.0002$ and $\tau = 0.0001$. The resulting continent assortativity value of 0.753 is higher than the other methods, and the same is seen for country and division assortativity.

Method	ε	τ	No. of edges	Assortativity		
				Continent	Country	Division
MSN	0	∞	717	0.626	0.581	0.360
τ-network	∞	0.0001	2056	0.702	0.631	0.351
	0.0002	∞	821	0.661	0.625	0.342
	0.0002	0.0001	727	**0.753**	**0.706**	**0.374**

Table 3.1 This table shows the attribute assortativity values for optimal choices of ε and τ for MSN, , , and threshold-based network, each using TN93 distance.

We find that performs the best in terms of country, continent, and division assortativity values across all four methods.

3.3.3 Transmission network analysis

We evaluated the precision and recall of on the ETL dataset. To evaluate the transmission network quality of , we define an undirected transmission link to be the pair of sequence locations of the two vertices connected by an edge in . The set of all unique undirected transmission links forms the transmission network.

For each method shown in Table 3.2, we produced its transmission network and evaluated the precision and recall against the known links provided in the ETL dataset. We calculate precision as the ratio of the number of known true links predicted by the method and the total number of predicted links, and we calculate recall as the ratio of the number of known true links predicted and the total number of known true links in the ETL dataset.

Table 3.2 shows that MSN performed best in Precision and F1-Score. τ-network performed best in Recall but not as well in precision or F1-score. and performed comparably well to MSN, and together these network based methods all outperformed the other standard methods being compared.

3.3.4 Scalability analysis

To examine scalability of the proposed methods, we applied the tool to datasets of increasing sizes of up to several hundred thousand sequences. For each of these datasets, we ran in TN93 mode and Hamming distance mode separately and recorded the running times. Figure 3.5 shows the results of the analysis. We see that has a quadratic runtime in both modes, but that Hamming distance is significantly faster because of its efficient implementation.

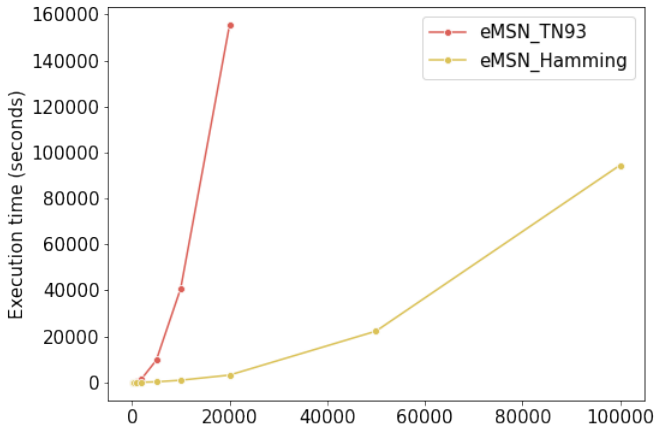

Figure 3.5 Runtime analysis of on increasing input sizes. is a quadratic algorithm in both TN93 and Hamming distance modes, although Hamming distance, with its efficient implementation, is much faster.

3.4 Conclusion

We have developed two versions of a new network-based tool, and , which generalize the minimal spanning networks and τ-network approaches to representing genetic relationships.

We compared the proposed tools with other network-based methods using attribute assortativity values. The experiments show that Hamming distance does not perform as well as with TN93 distance. With TN93, outperforms all the other methods in continent, country, and division attribute assortativity.

Further, we validated multiple tools, including the proposed ones, on known transmission networks. We evaluated recall, precision and F1-score for each tool. We found that network-based tools perform better than the others, including those that are phylogeny-based.

The results validated the network and showed that the structure of the network correlates with phylogenetic trees. is interpretable, integrable and scalable.

Tool	Recall*	Precision*	F1-Score*
MSN ($\varepsilon = 0, \tau = \infty$, TN93)	80%	**7.6%**	**0.139**
τ-network ($\varepsilon = \infty, \tau = 0.0001$, TN93)	**96%**	2.5%	0.049
($\varepsilon = 0.0002, \tau = \infty$, TN93)	80%	7.4%	0.135
($\varepsilon = 0.0002, \tau = 0.0001$, TN93)	72%	6.6%	0.121
CS-phylogeny	80%	4.76%	0.090
NETWORK5011CS	72%	4.99%	0.093
RAxML	64%	4.26%	0.080
Bitrugs	52%	3.38%	0.063
outbreaker	28%	5.83%	0.097
phybreak	4%	0.83%	0.076

Table 3.2 Recall and precision comparison across different methods ran on the ETL dataset. MSN methods were ran using the TN93 distance metric. Recall is defined as the ratio of known true links formed by the tool to the total number of known true links. Precision is defined as the ratio of known true links formed by the tool to the total number of links formed by the tool. F1-Score is defined as the twice the product of precision and recall divided by the sum of precision and recall. * The ground truth is only partially known.

Users of our proposed tools can fit the parameters ε and τ to any dataset using the same methodology we used in our analysis, namely, fixing one parameter and varying the other, then fixing the other and varying the first.

Our methodology is implemented in MicrobeTrace[7], a tool currently in use by the CDC for viral outbreak investigation.

CHAPTER 4

Bioinformatics Tools for RNA-seq Data Analysis

4.1 Introduction

With the arrival of next-generation sequencing technology, there are significant methodologies developed by bioinformatics communities to study the activity and interaction of metabolic pathways in microbial communities. Despite many advances in methods to process the RNA-seq data, there are still challenges in retrieving information from the community-level data, which often generates unstable pathway activity information. Apart from standard pathway activity information, there is always a need for a single unit (e.g., enzyme contribution in case of metabolic pathway activity)[67,18,53,61,53,69,16,35,44,71,60].

In this research, we proposed a Maximum Likelihood-based model considering the annotation information from KEGG and MAP databases to predict the metabolic pathway activity using enzyme expression and participation coefficients[37,34]. Our model works as follows; first, it accepts the meta-transcriptomic data, an organism-specific pathway list from KEGG. Next, it merges the enzymes sharing the same contigs, estimates the individual enzyme participation level in each pathway using the Expectation-Maximization algorithm, and uses the calculated values for a more accurate prediction of pathway activity. Finally, the model computes the correlation between pathways and environmental parameters to generate the final results. We tested the model on meta-transcriptomic data from a marine microbial community collected at different time points with different environmental parameters.

4.2 Dataset

This study considers a meta-transcriptomic dataset with 26 samples of a bacterioplankton community from surface waters of the Northern Gulf of Mexico. The data was accompanied by environmental parameters collected in July 2015 at two different depths, 2 and 18 meters, every four hours for 48 hours. The six environmental parameters are PAR (photosynthetic active radiation), seawater dissolved oxygen concentration, density, salinity, temperature, and chlorophyll concentration.

4.2.1 Metabolic pathway database

We considered a microbial community-specific metabolic pathway list from KEGG. We also removed high-level metabolic pathways such as ec01100, ec01110, ec01120, and ec01130 from our list. Finally, we only considered 69 micro-organism-specific metabolic pathways for the current research[33,24,36].

4.2.2 Algorithm

In this research, we proposed an enhanced algorithm published in[48] to infer the enzyme expression and participation levels in a metabolic pathway by repeatedly applying the maximum likelihood model. The solutions for these models were achieved using the Expectation-Maximization (EM) algorithm Figure 4.2.2.

Step 1: Estimation of the abundance of the assembled contigs.

The abundances inferred by RNA-seq quantification tool (here, we used IsoEM[47].)

Step 2: Estimation of the enzyme expression

The enzyme expression estimated using contigs abundance and mapping of contigs onto en-

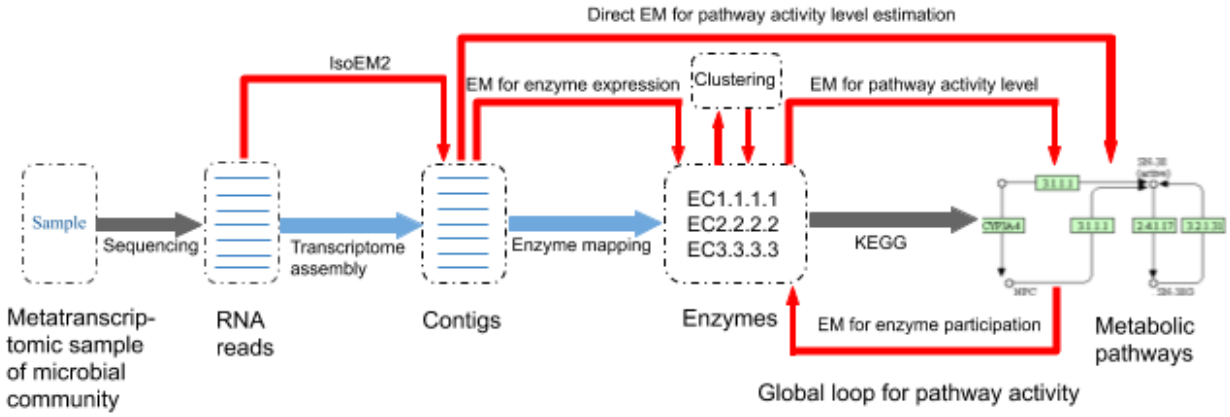

Figure 4.1 Pipeline of metabolic pathway analysis for a microbial community sample. The metatranscriptomic data obtained from microbial community samples are sequenced, and raw reads are assembled into contigs. The genes containing obtained contigs are further mapped into the enzyme-pathway database. Contig frequencies are obtained using IsoEM2[47]. The direct EM estimates pathway activity levels using directly contig frequencies. Alternatively, we first estimate the enzyme expressions, then cluster enzymes, and simultaneously estimate enzyme participation in each pathway and pathway activity levels.

zymes.

Step 3: Calculation of metabolic pathway activity

The Estimation-Maximization algorithm used for pathway activity levels calculation based on inferred enzyme expression and metabolic pathway annotation. Here we proposed two different way to calculate metabolic pathway activity.

a) Direct EM for pathway level estimation

b) The Global loop for pathway activity

Step 4: Direct EM for pathway activity level estimation

The EM for pathway activity levels is based on inferred enzyme expression and metabolic pathway annotation. Each enzyme is initialize a participation level of $1/|w|$. $|w|$ is the total amount

of enzymes in the pathway w.

Step 5: The Global loop for pathway activity

The global loop for pathway activity updates the enzyme participation level by fitting expected enzyme expressions to the expressions estimated by EM for enzyme expression. The global loop computes the pathway activity and replaces the Pathway activity estimated by Direct EM. Which directly estimates pathway activity from contig abundances bypassing enzyme expression and participation coefficients.

4.3 Results

Our results consist of empirical and statistical validation of estimated enzyme expression, enzyme participation, and pathway activity estimations. We first analyze the stability of enzyme participation levels and then check the number of enzyme expressions and pathway activities that correlate with environmental parameters.

4.3.1 Enzyme participation coefficients

We estimate the participation level of individual enzymes in each pathway separately for each data point. We found that the participation level does not significantly change from one data point to another, i.e. the standard deviation is significantly smaller than the mean for all enzymes. Table 4.1 presents the participation level of all expressed enzymes in the pathway ec00561.

ec00561	D1:12	D1:16	D1:20	D2:00	D2:04	D2:08	D2:12	D2:16	D3:00	D3:04	D3:12	AVE	STD
EC:1.1.1.2	56.43	52.43	43.04	46.68	51.86	41.29	65.25	39.34	46.54	37.81	0.00	48.07	8.10
EC:1.2.1.3	98.96	136.18	95.66	69.85	79.35	69.48	112.58	60.23	80.19	83.30	208.15	99.45	40.11
EC:1.1.1.21	61.63	62.54	48.77	46.64	50.41	39.57	55.37	34.04	49.16	40.73	77.37	51.47	11.72
EC:2.7.7.9	60.17	55.14	50.26	39.73	44.57	45.06	62.68	38.50	45.22	41.12	131.96	55.86	25.27
EC:3.2.1.22	47.41	47.43	38.91	41.32	42.99	35.23	0.00	30.73	41.31	38.90	0.00	40.47	5.07
EC:2.3.1.20	90.96	77.07	0.00	61.58	59.41	61.75	57.87	52.78	56.58	76.46	0.00	66.05	11.86
EC:2.7.1.31	94.47	131.20	99.82	119.04	122.46	0.00	0.00	0.00	119.22	122.29	0.00	115.50	12.28
EC:2.3.1.51	58.79	90.73	75.29	75.97	69.66	59.05	79.23	59.45	80.74	65.96	0.00	71.49	10.22
EC:3.13.1.1	0.00	90.59	81.77	59.46	69.55	68.97	76.55	59.72	69.07	75.58	0.00	72.36	9.46
EC:1.1.1.156	90.96	0.00	0.00	0.00	0.00	0.00	0.00	52.78	0.00	0.00	0.00	71.87	19.09
EC:1.1.1.6	0.00	0.00	0.00	0.00	0.00	0.00	57.87	52.78	0.00	0.00	0.00	55.33	2.54
EC:2.3.1.15	50.80	66.09	55.42	55.00	49.67	49.59	65.75	44.20	60.25	54.64	149.56	63.72	27.90
EC:2.3.1.22	90.96	77.07	63.98	61.58	59.41	0.00	0.00	52.78	0.00	0.00	0.00	67.63	12.72
EC:2.4.1.241	0.00	0.00	0.00	61.58	0.00	61.75	57.87	52.78	56.58	0.00	0.00	58.11	3.35
EC:2.4.1.315	0.00	0.00	0.00	61.58	59.41	0.00	0.00	0.00	0.00	0.00	0.00	60.49	1.08
EC:2.4.1.336	0.00	0.00	0.00	0.00	0.00	61.75	0.00	0.00	0.00	0.00	0.00	61.75	0.00
EC:2.4.1.337	0.00	0.00	0.00	0.00	59.41	0.00	0.00	0.00	0.00	0.00	0.00	59.41	0.00
EC:2.4.1.46	0.00	77.07	63.98	0.00	0.00	61.75	57.87	52.78	56.58	0.00	0.00	61.67	7.77
EC:2.7.1.107	50.80	66.09	55.42	55.00	49.67	49.59	65.75	44.20	60.25	54.64	0.00	55.14	6.77
EC:2.7.1.29	0.00	0.00	108.83	78.85	74.39	82.41	77.92	65.58	75.45	78.86	0.00	80.29	11.74
EC:2.7.1.30	90.96	77.07	63.98	61.58	59.41	61.75	57.87	52.78	56.58	76.46	0.00	65.84	11.27
EC:2.7.8.20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	56.58	0.00	0.00	56.58	0.00
EC:3.1.1.23	0.00	0.00	63.98	61.58	59.41	61.75	57.87	52.78	56.58	76.46	0.00	61.30	6.59
EC:3.1.1.3	90.96	77.07	63.98	61.58	59.41	61.75	57.87	52.78	56.58	76.46	390.20	95.33	93.86
EC:3.1.1.34	0.00	0.00	63.98	0.00	0.00	0.00	0.00	0.00	0.00	76.46	0.00	70.22	6.24
EC:3.1.3.4	43.19	50.83	42.05	44.04	43.19	37.42	64.52	33.09	46.99	40.50	51.36	45.20	7.95
EC:3.1.3.81	0.00	0.00	0.00	0.00	0.00	49.59	0.00	0.00	0.00	54.64	0.00	52.12	2.53

Table 4.1 Enzyme participation levels for all enzymes across all data points for 2m depth in the metabolic pathway ec00561.

4.3.2 Correlation of pathway activity levels

The goal of regression-based validation is to check our hypothesis that there exist enzymes and pathways whose expression and activity level variation across data points can be explained (i.e. correlate with) certain environmental parameters. For each environmental parameter, we check whether it significantly correlates ($P < 5\%$) with each enzyme across 11 data points for the 2-meter depth (see Table 4.2). In the row 2, we give 95% CI for the number of significantly correlated enzymes with a randomly permuted parameter. Since the upper bound of 95% CI for salinity is 190

(row 2), we conclude that there is no evidence of enzymes significantly correlated with salinity. We also report the enzyme that correlates the most with salinity, i.e. EC 1.2.1.59. From Table 4.2 we see that most parameters do not correlate well with enzymes, except perhaps PAR.

Table 4.3 is the same as Table 4.2 but reports correlation significance of pathway activities instead of enzyme expressions. In contrast to enzymes it is clear that the many metabolic pathways correlate with each environmental parameter and this correlation is not by chance. Indeed, pathway activity is supposed to be more stable than enzyme expression since generally metabolism is much less affected by the current. For each environmental parameter, we also cross-check the PUBMED database whether the most correlated pathway is known to depend on this parameter. For instance, fatty acid degradation is well correlated with salinity, and several studies reported that fatty acid degradation is often altered by salinity at sea surface environments[30,38,14]. The citric acid pathway's role is to provide the energy required for the growth and division of microorganisms by breaking organic molecules in the presence of oxygen[33]. Additionally, it plays a central role in regulating other metabolic processes in microorganisms. The occurrence of fatty acid biosynthesis is diverse in the microbial community, which controls lipid homeostasis and biogenesis. Fatty acid biosynthesis supports the membrane biogenesis and controls the usages of ATP, crucial for microbial metabolism[24,36].

	Salinity	Temp	Oxygen	Chl	PAR	Density	MLR
1. # enzymes	146	110	117	93	97	138	156
2. 95% CI	80-190	79-114	62-94	58-92	36-63	82-123	70-107
3. EC number	1.2.1.59	2.6.1.1	3.1.3.11	2.2.1.7	3.5.1.16	2.4.1.16	1.1.1.136

Table 4.2 1. The number of enzymes significantly correlated with each of 6 environmental parameters and their linear combination (via multiple linear regression (MLR)). 2. The number of enzymes strongly correlated with randomly permuted parameter values (95% CI). 3. The EC number of the metabolic enzyme which is the most strongly correlated with the corresponding parameter.

	Salinity	Temp	Oxygen	Chl	PAR	Density	MLR
1. # pathways	31	22	19	18	14	30	22
2. 95% CI	1-8	0-8	0-6	0-6	0-6	1-8	0-7
3. Pathway	ec00071	ec00195	ec00622	ec00460	ec00360	ec00071	ec00626

Table 4.3 **Global Loop EM.** 1. The number of pathways significantly correlated with each of 6 environmental parameters and correlated via multiple linear regression. 2. The number of pathways strongly correlated with randomly permuted parameter values (95% CI). 3. The EC number of the metabolic pathway which is the most strongly correlated with the corresponding parameter.

Table 4.4 is the same as Table 4.3. The only exception for this table being Direct EM used to compute metabolic pathway activity directly from contigs, as opposed to Global Loop EM, which uses enzyme expression and enzyme participation coefficients to compute pathway activity. While there is significant correlation between metabolic pathway activity and temperature, chlorophyll, as well as all environmental parameters bundled together, some other pathways may have correlated with the rest of the environmental parameters by chance. The statistical regression validation

used to evaluate our model clearly demonstrates Global Loop EM's ability to calculate metabolic pathway activity more accurately than Direct EM.

	Salinity	Temp	Oxygen	Chl	PAR	Density	MLR
1. # pathways	5	14	5	8	1	4	10
2. 95% CI	1-10	1-11	1-8	0-7	0-6	1-8	0-8
3. Pathway	ec00364	ec00310	ec00281	ec00281	ec00740	ec00623	ec00623

Table 4.4 **Direct EM.** Similarly to Table 4.3 this table presents the results of the statistical validation, the only difference is the Direct EM from contigs to pathway activity being used here.

4.3.3 Cyclic changes of enzyme expressions and pathway activities

We hypothesize that we will be able to observe the cyclic changes in enzyme expression and pathway activity level during 36 hours from 00:00 am on day 2 until 12:00 am on day 3. The cyclic changes should manifest themselves as a higher similarity between two respectively mid-days and mid-nights which are 24 hours apart than the similarity between two data points that are 12 hours apart. We measure similarity between two data points by the correlation between all estimated enzyme expressions or, alternatively, all estimated pathway activity levels. Figure 4.3.3.(a) (respectively, Figure 4.3.3.(b)) shows the correlation between enzyme expressions in 3 time points at the depth of 2m (respectively, 18 m). Similarly, Fig.4.3.3.c, d show the correlations between pathway activity levels. For the enzyme expressions and the pathway activity levels, the correlation between midnight samples (24 hours gap) is higher than the correlation between midnight and noon samples (just 12 hour gap). It is also important to notice that as expected pathway activity levels are more stable than enzyme expressions. Indeed, correlations between enzymes expression are significantly lower than correlations between pathways activity levels.

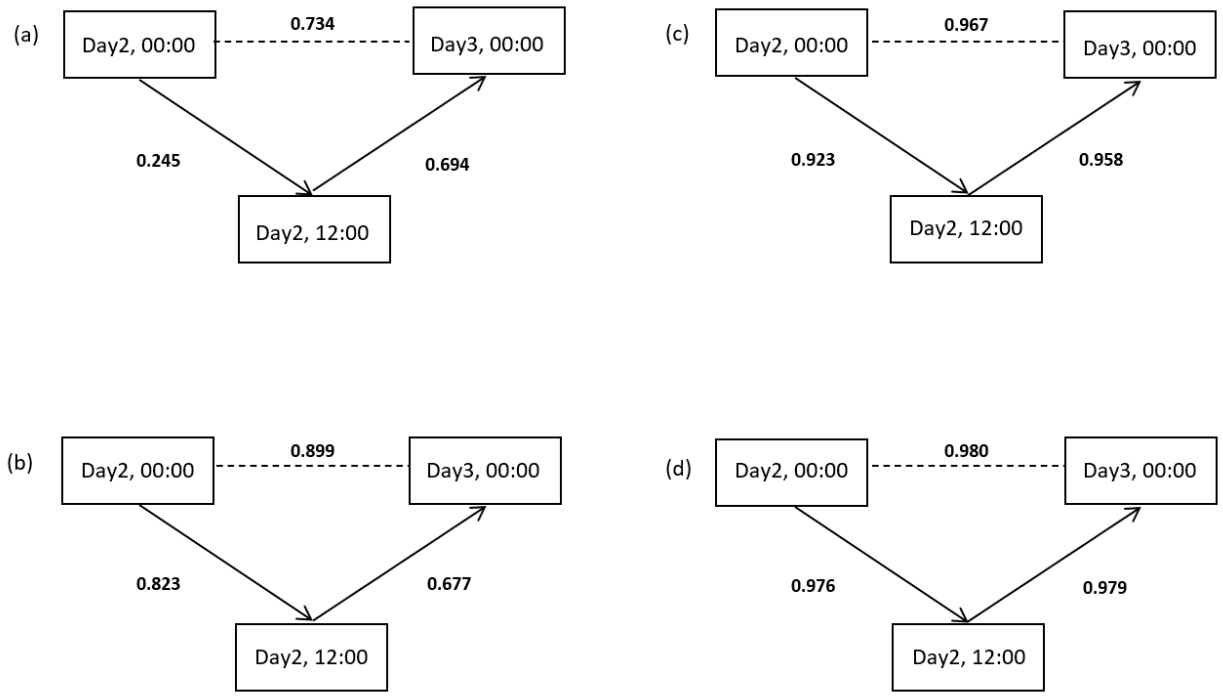

Figure 4.2 Correlations between enzyme expressions for 3 time points (time 00:00 of the day 2, 00:00 of the day 3, and 12:00 of the day 2) at 2 m-depth (a) and, respectively, at 18 m depth (b). Correlations between pathway activity levels for 3 time points (time 00:00 of day 2, 00:00 day 3, and 12:00 of day 2) at 2 m-depth (c) and, respectively, at 18 m depth (d).

4.4 Conclusion

The proposed Estimation-Maximization (EM) based metabolic pathway enrichment analysis methodology significantly predicts the participation of enzyme levels that do not vary across the data samples. The model validates the enzyme expression and metabolic pathway activity with environmental parameters with regression. The results show that pathway activity levels significantly correlate with environmental parameters compared to the enzyme expression. The 3-way metabolic pathway correlation analysis predicts the daytime samples are more closely related than the night samples.

REFERENCES

1. I. Alexiev, E. M. Campbell, S. Knyazev, Y. Pan, L. Grigorova, R. Dimitrova, A. Partsuneva, A. Gancheva, A. Kostadinova, C. Seguin-Devaux, and W. M. Switzer. Molecular epidemiology of the HIV-1 subtype b sub-epidemic in bulgaria. *Viruses*, 12(4):441, Apr. 2020. doi: 10.3390/v12040441. URL https://doi.org/10.3390/v12040441.

2. I. Alexiev, E. M. Campbell, S. Knyazev, Y. Pan, L. Grigorova, R. Dimitrova, A. Partsuneva, A. Gancheva, A. Kostadinova, C. Seguin-Devaux, I. Elenkov, N. Yancheva, and W. M. Switzer. Molecular epidemiological analysis of the origin and transmission dynamics of the HIV-1 CRF01_AE sub-epidemic in bulgaria. *Viruses*, 13(1):116, Jan. 2021. doi: 10.3390/v13010116. URL https://doi.org/10.3390/v13010116.

3. H. J. Bandelt, P. Forster, and A. Rohl. Median-joining networks for inferring intraspecific phylogenies. *Molecular Biology and Evolution*, 16(1):37–48, Jan. 1999. doi: 10.1093/oxfordjournals.molbev.a026036. URL https://doi.org/10.1093/oxfordjournals.molbev.a026036.

4. N. Beerenwinkel, T. Sing, T. Lengauer, J. Rahnenführer, K. Roomp, I. Savenkov, R. Fischer, D. Hoffmann, J. Selbig, K. Korn, H. Walter, T. Berg, P. Braun, G. Fätkenheuer, M. Oette, J. Rockstroh, B. Kupfer, R. Kaiser, and M. Däumer. Computational methods for the design of effective therapies against drug resistant HIV strains. *Bioinformatics*, 21(21):3943–3950, September 2005. doi: 10.1093/bioinformatics/bti654.

5. M. Bousali, A. Dimadi, E.-G. Kostaki, S. Tsiodras, G. K. Nikolopoulos, D. N. Sgouras, G. Ma-

giorkinis, G. Papatheodoridis, V. Pogka, G. Lourida, A. Argyraki, E. Angelakis, G. Sourvinos, A. Beloukas, D. Paraskevis, and T. Karamitros. SARS-CoV-2 molecular transmission clusters and containment measures in ten european regions during the first pandemic wave. *Life*, 11 (3), 2021. doi: 10.3390/life11030219.

6. E. M. Campbell, H. Jia, A. Shankar, D. Hanson, W. Luo, S. Masciotra, S. M. Owen, A. M. Oster, R. R. Galang, M. W. Spiller, S. J. Blosser, E. Chapman, J. C. Roseberry, J. Gentry, P. Pontones, J. Duwve, P. Peyrani, R. M. Kagan, J. M. Whitcomb, P. J. Peters, W. Heneine, J. T. Brooks, and W. M. Switzer. Detailed transmission network analysis of a large opiate-driven outbreak of HIV infection in the united states. *The Journal of Infectious Diseases*, 216(9):1053–1062, Oct. 2017. doi: 10.1093/infdis/jix307. URL `https://doi.org/10.1093/infdis/jix307`.

7. E. M. Campbell, A. Boyles, A. Shankar, J. Kim, S. Knyazev, R. Cintron, and W. M. Switzer. MicrobeTrace: Retooling molecular epidemiology for rapid public health response. *PLOS Computational Biology*, 17(9):e1009300, Sept. 2021. doi: 10.1371/journal.pcbi.1009300. URL `https://doi.org/10.1371/journal.pcbi.1009300`.

8. F. Campbell, X. Didelot, R. Fitzjohn, N. Ferguson, A. Cori, and T. Jombart. outbreaker2: a modular platform for outbreak reconstruction. *BMC Bioinformatics*, 19(S11), Oct. 2018. doi: 10.1186/s12859-018-2330-z. URL `https://doi.org/10.1186/s12859-018-2330-z`.

9. D. S. Campo, Z. Dimitrova, L. Yamasaki, P. Skums, D. T. Lau, G. Vaughan, J. C. Forbi, C.-G. Teo, and Y. Khudyakov. Next-generation sequencing reveals large connected net-

works of intra-host HCV variants. *BMC Genomics*, 15(S5), July 2014. doi: 10.1186/1471-2164-15-s5-s4. URL https://doi.org/10.1186/1471-2164-15-s5-s4.

10. D. S. Campo, P. Skums, Z. Dimitrova, G. Vaughan, J. C. Forbi, C.-G. Teo, Y. Khudyakov, and D. T.-Y. Lau. Drug resistance of a viral population and its individual intrahost variants during the first 48 hours of therapy. *Clinical Pharmacology and Therapeutics*, 95(6):627–635, June 2014. doi: 10.1038/clpt.2014.20.

11. D. S. Campo, G.-L. Xia, Z. Dimitrova, Y. Lin, J. C. Forbi, L. Ganova-Raeva, L. Punkova, S. Ramachandran, H. Thai, P. Skums, S. Sims, I. Rytsareva, G. Vaughan, H.-J. Roh, M. A. Purdy, A. Sue, and Y. Khudyakov. Accurate genetic detection of hepatitis c virus transmissions in outbreak settings. *Journal of Infectious Diseases*, 213(6):957–965, Nov. 2015. doi: 10.1093/infdis/jiv542. URL https://doi.org/10.1093/infdis/jiv542.

12. D. S. Campo, G.-L. Xia, Z. Dimitrova, Y. Lin, J. C. Forbi, L. Ganova-Raeva, L. Punkova, S. Ramachandran, H. Thai, P. Skums, S. Sims, I. Rytsareva, G. Vaughan, H.-J. Roh, M. A. Purdy, A. Sue, and Y. Khudyakov. Accurate genetic detection of hepatitis C virus transmissions in outbreak settings. *The Journal of Infectious Diseases*, 213(6):957–965, March 2016. doi: 10.1093/infdis/jiv542.

13. D. S. Campo, J. Zhang, S. Ramachandran, and Y. Khudyakov. Transmissibility of intra-host hepatitis c virus variants. *BMC Genomics*, 18(S10), Dec. 2017. doi: 10.1186/s12864-017-4267-4. URL https://doi.org/10.1186/s12864-017-4267-4.

14. C. Carvalho and M. Caramujo. The various roles of fatty acids, 2018.

15. W. F. de la Vega, M. Karpinski, C. Kenyon, and Y. Rabani. Approximation schemes for

clustering problems. In *Proceedings of the Thirty-Fifth Annual ACM Symposium on Theory of Computing*, STOC '03, pages 50—-58. Association for Computing Machinery, 2003. doi: 10.1145/780542.780550.

16. M. Donato, Z. Xu, A. Tomoiaga, J. G. Granneman, R. G. Mackenzie, R. Bao, N. G. Than, P. H. Westfall, R. Romero, and S. Draghici. Analysis and correction of crosstalk effects in pathway analysis. *Genome Res.*, 23(11):1885–1893, Nov. 2013.

17. D. C. Douek, P. D. Kwong, and G. J. Nabel. The rational design of an AIDS vaccine. *Cell*, 124(4):677–681, 2006. doi: 10.1016/j.cell.2006.02.005.

18. B. Efron and R. Tibshirani. On testing the significance of sets of genes. *The Annals of Applied Statistics*, 1(1):107–129, 2007.

19. EMBL-EBI. EMBL's European Bioinformatics Institute. URL `https://www.ebi.ac.uk/`.

20. L. Excoffier and P. E. Smouse. Using allele frequencies and geographic subdivision to reconstruct gene trees within a species: molecular variance parsimony. *Genetics*, 136(1):343–359, Jan. 1994. doi: 10.1093/genetics/136.1.343. URL `https://doi.org/10.1093/genetics/136.1.343`.

21. J. R. Fauver, M. E. Petrone, E. B. Hodcroft, K. Shioda, H. Y. Ehrlich, A. G. Watts, C. B. Vogels, A. F. Brito, T. Alpert, A. Muyombwe, J. Razeq, R. Downing, N. R. Cheemarla, A. L. Wyllie, C. C. Kalinich, I. Ott, J. Quick, N. J. Loman, K. M. Neugebauer, A. L. Greninger, K. R. Jerome, P. Roychoudhury, H. Xie, L. Shrestha, M.-L. Huang, V. E. Pitzer, A. Iwasaki, S. B. Omer, K. Khan, I. I. Bogoch, R. A. Martinello, E. F. Foxman, M. L. Landry, R. A.

Neher, A. I. Ko, and N. D. Grubaugh. Coast-to-coast spread of SARS-CoV-2 in the united states revealed by genomic epidemiology. Mar. 2020. doi: 10.1101/2020.03.25.20043828. URL https://doi.org/10.1101/2020.03.25.20043828.

22. J. Felsenstein. *Inferring Phylogenies*. Sinauer Associates is an imprint of Oxford University Press, paperback edition, 9 2003. ISBN 978-0878931774. URL https://lead.to/amazon/com/?op=bt&la=en&cu=usd&key=0878931775.

23. P. Forster, L. Forster, C. Renfrew, and M. Forster. Phylogenetic network analysis of SARS-CoV-2 genomes. *Proceedings of the National Academy of Sciences*, 117(17):9241–9243, Apr. 2020. doi: 10.1073/pnas.2004999117. URL https://doi.org/10.1073/pnas.2004999117.

24. G. Gago, L. Diacovich, A. Arabolaza, S.-C. Tsai, and H. Gramajo. Fatty acid biosynthesis in actinomycetes. *FEMS Microbiol. Rev.*, 35(3):475–497, May 2011.

25. B. Gaschen, J. Taylor, K. Yusim, B. Foley, F. Gao, D. Lang, V. Novitsky, B. Haynes, B. H. Hahn, T. Bhattacharya, and B. Korber. Diversity considerations in HIV-1 vaccine selection. *Science*, 296(5577):2354–2360, 2002. doi: 10.1126/science.1070441.

26. O. Glebova, S. Knyazev, A. Melnyk, A. Artyomenko, Y. Khudyakov, A. Zelikovsky, and P. Skums. Inference of genetic relatedness between viral quasispecies from sequencing data. *BMC Genomics*, 18(S10), Dec. 2017. doi: 10.1186/s12864-017-4274-5. URL https://doi.org/10.1186/s12864-017-4274-5.

27. O. Glebova, S. Knyazev, A. Melnyk, A. Artyomenko, Y. Khudyakov, A. Zelikovsky, and P. Skums. Inference of genetic relatedness between viral quasispecies from sequencing data.

BMC Genomics, 2017. doi: 10.1186/s12864-017-4274-5.

28. A. S. Gonzalez-Reiche, M. M. Hernandez, M. J. Sullivan, B. Ciferri, H. Alshammary, A. Obla, S. Fabre, G. Kleiner, J. Polanco, Z. Khan, B. Alburquerque, A. van de Guchte, J. Dutta, N. Francoeur, B. S. Melo, I. Oussenko, G. Deikus, J. Soto, S. H. Sridhar, Y.-C. Wang, K. Twyman, A. Kasarskis, D. R. Altman, M. Smith, R. Sebra, J. Aberg, F. Krammer, A. García-Sastre, M. Luksza, G. Patel, A. Paniz-Mondolfi, M. Gitman, E. M. Sordillo, V. Simon, and H. van Bakel. Introductions and early spread of sars-cov-2 in the new york city area. *Science*, 369(6501):297–301, May 2020. doi: 10.1126/science.abc1917. URL `https://doi.org/10.1126/science.abc1917`.

29. K. M. Grande, C. L. Schumann, M. C. B. Ocfemia, J. M. Vergeront, J. O. Wertheim, and A. M. Oster. Transmission patterns in a low HIV-morbidity state — wisconsin, 2014–2017. *MMWR. Morbidity and Mortality Weekly Report*, 68(6):149–152, Feb. 2019. doi: 10.15585/mmwr.mm6806a5. URL `https://doi.org/10.15585/mmwr.mm6806a5`.

30. S. M. Heinzelmann, D. Chivall, D. M'Boule, D. Sinke-Schoen, L. Villanueva, J. S. Sinninghe Damsté, S. Schouten, and M. T. J. van der Meer. Comparison of the effect of salinity on the D/H ratio of fatty acids of heterotrophic and photoautotrophic microorganisms. *FEMS Microbiology Letters*, 362(10), 2015.

31. J. Holland, J. De La Torre, and D. Steinhauer. RNA virus populations as quasispecies. *Current Topics in Microbiology and Immunology*, pages 1–20, 1992.

32. C. J. Houldcroft, M. A. Beale, and J. Breuer. Clinical and biological insights from viral genome sequencing. *Nature Reviews Microbiology*, 15(3):183–192, jan 2017. doi: 10.1038/

nrmicro.2016.182. URL https://doi.org/10.1038/nrmicro.2016.182.

33. Y. Hu and J. F. Holden. Citric acid cycle in the hyperthermophilic archaeon pyrobaculum islandicum grown autotrophically, heterotrophically, and mixotrophically with acetate. *J. Bacteriol.*, 188(12):4350–4355, 2006.

34. M. Huntemann, N. N. Ivanova, K. Mavromatis, H. J. Tripp, D. Paez-Espino, K. Tennessen, K. Palaniappan, E. Szeto, M. Pillay, I.-M. A. Chen, A. Pati, T. Nielsen, V. M. Markowitz, and N. C. Kyrpides. The standard operating procedure of the DOE-JGI metagenome annotation pipeline (MAP v.4). *Stand. Genomic Sci.*, 11:17, Feb. 2016.

35. D. H. Huson, S. Mitra, H.-J. Ruscheweyh, N. Weber, and S. C. Schuster. Integrative analysis of environmental sequences using MEGAN4. *Genome Res.*, 21(9):1552–1560, Sept. 2011.

36. H. J. Janßen and A. Steinbüchel. Fatty acid synthesis in escherichia coli and its applications towards the production of fatty acid based biofuels. *Biotechnol. Biofuels*, 7(1):7, Jan. 2014.

37. M. Kanehisa. KEGG: Kyoto encyclopedia of genes and genomes, 2000.

38. J. Z. Kaye. Halomonas neptunia sp. nov., halomonas sulfidaeris sp. nov., halomonas axialensis sp. nov. and halomonas hydrothermalis sp. nov.: halophilic bacteria isolated from deep-sea hydrothermal-vent environments, 2004.

39. S. Khare, C. Gurry, L. Freitas, M. B. Schultz, G. Bach, A. Diallo, N. Akite, J. Ho, R. T. Lee, W. Yeo, G. Core Curation Team, and S. Maurer-Stroh. GISAID's role in pandemic response. *China CDC weekly*, 3(49):1049—1051, 2021. doi: 10.46234/ccdcw2021.255.

40. D. Klinkenberg, J. Backer, X. Didelot, C. Colijn, and J. Wallinga. New method to reconstruct phylogenetic and transmission trees with sequence data from infectious disease outbreaks.

Aug. 2016. doi: 10.1101/069195. URL https://doi.org/10.1101/069195.

41. S. Knyazev, L. Hughes, P. Skums, and A. Zelikovsky. Epidemiological data analysis of viral quasispecies in the next-generation sequencing era. *Briefings in Bioinformatics*, 22(1):96–108, June 2020. doi: 10.1093/bib/bbaa101. URL https://doi.org/10.1093/bib/bbaa101.

42. S. Knyazev, V. Tsyvina, A. Shankar, A. Melnyk, A. Artyomenko, T. Malygina, Y. B. Porozov, E. M. Campbell, W. M. Switzer, P. Skums, S. Mangul, and A. Zelikovsky. Accurate assembly of minority viral haplotypes from next-generation sequencing through efficient noise reduction. *Nucleic Acids Research*, July 2021. doi: 10.1093/nar/gkab576. URL https://doi.org/10.1093/nar/gkab576.

43. S. Knyazev, V. Tsyvina, A. Shankar, A. Melnyk, A. Artyomenko, T. Malygina, Y. B. Porozov, E. M. Campbell, W. M. Switzer, P. Skums, S. Mangul, and A. Zelikovsky. Accurate assembly of minority viral haplotypes from next-generation sequencing through efficient noise reduction. *Nucleic Acids Research*, 49(17):e102–e102, July 2021. ISSN 0305-1048. doi: 10.1093/nar/gkab576.

44. K. M. Konwar, N. W. Hanson, A. P. Pagé, and S. J. Hallam. MetaPathways: a modular pipeline for constructing pathway/genome databases from environmental sequence information. *BMC Bioinformatics*, 14:202, June 2013.

45. T. Li, S. Ma, and M. Ogihara. Entropy-based criterion in categorical clustering. In *Proceedings, Twenty-First International Conference on Machine Learning, ICML 2004*, volume 3, pages 536–543, 2004. doi: 10.1145/1015330.1015404.

46. A. G. Longmire, S. Sims, I. Rytsareva, D. S. Campo, P. Skums, Z. Dimitrova, S. Ramachandran, M. Medrzycki, H. Thai, L. Ganova-Raeva, Y. Lin, L. T. Punkova, A. Sue, M. Mirabito, S. Wang, R. Tracy, V. Bolet, T. Sukalac, C. Lynberg, and Y. Khudyakov. Ghost: global hepatitis outbreak and surveillance technology. *BMC Genomics*, 18(S10), dec 2017. doi: 10.1186/s12864-017-4268-3. URL https://doi.org/10.1186/s12864-017-4268-3.

47. I. Mandric, S. Knyazev, C. Padilla, F. Stewart, I. I. Măndoiu, and A. Zelikovsky. Metabolic analysis of metatranscriptomic data from planktonic communities. pages 396–402, 2017.

48. I. Mandric, Y. Temate-Tiagueu, T. Shcheglova, S. Al Seesi, A. Zelikovsky, and I. I. Mandoiu. Fast bootstrapping-based estimation of confidence intervals of expression levels and differential expression from RNA-Seq data. *Bioinformatics*, 33(20):3302–3304, Oct. 2017.

49. A. Melnyk, S. Knyazev, F. Vannberg, L. Bunimovich, P. Skums, and A. Zelikovsky. Using earth mover's distance for viral outbreak investigations. *BMC Genomics*, 21(S5), dec 2020. doi: 10.1186/s12864-020-06982-4. URL https://doi.org/10.1186/s12864-020-06982-4.

50. A. Melnyk, S. Knyazev, F. Vannberg, L. Bunimovich, P. Skums, and A. Zelikovsky. Using earth mover's distance for viral outbreak investigations. *BMC Genomics*, 21(582), 2020. doi: 10.1186/s12864-020-06982-4.

51. A. Melnyk, F. Mohebbi, S. Knyazev, B. Sahoo, R. Hosseini, P. Skums, A. Zelikovsky, and M. Patterson. Clustering based identification of SARS-CoV-2 subtypes. In *Computational Advances in Bio and Medical Sciences*, pages 127–141. Springer International Publishing, 2021. doi: 10.1007/978-3-030-79290-9_11. URL https://doi.org/10.1007/

978-3-030-79290-9_11.

52. A. Melnyk, F. Mohebbi, S. Knyazev, B. Sahoo, R. Hosseini, P. Skums, A. Zelikovsky, and M. Patterson. From Alpha to Zeta: Identifying variants and subtypes of SARS-CoV-2 via clustering. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, 28(11):1113–1129, 2021. doi: 10.1089/cmb.2021.0302.

53. C. Mitrea, Z. Taghavi, B. Bokanizad, S. Hanoudi, R. Tagett, M. Donato, C. Voichiţa, and S. Drăghici. Methods and approaches in the topology-based analysis of biological pathways, 2013.

54. D. Novikov, S. Knyazev, M. Grinshpon, P. I. Baykal, P. Skums, and A. Zelikovsky. Scalable reconstruction of SARS-CoV-2 phylogeny with recurrent mutations. *Journal of Computational Biology*, to appear.

55. A. M. Oster, A. M. France, N. Panneer, M. C. B. Ocfemia, E. Campbell, S. Dasgupta, W. M. Switzer, J. O. Wertheim, and A. L. Hernandez. Identifying clusters of recent and rapid HIV transmission through analysis of molecular surveillance data. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 79(5):543–550, Dec. 2018. doi: 10.1097/qai.0000000000001856. URL https://doi.org/10.1097/qai.0000000000001856.

56. S. L. K. Pond, S. Weaver, A. J. L. Brown, and J. O. Wertheim. HIV-TRACE (TRAnsmission cluster engine): a tool for large scale molecular epidemiology of HIV-1 and other rapidly evolving pathogens. *Molecular Biology and Evolution*, 35(7):1812–1819, Jan. 2018. doi: 10.1093/molbev/msy016. URL https://doi.org/10.1093/molbev/msy016.

57. S. Prabhakaran, M. Rey, O. Zagordi, N. Beerenwinkel, and V. Roth. HIV haplotype inference

using a propagating dirichlet process mixture model. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 11(1):182–191, Jan. 2014. doi: 10.1109/tcbb.2013.145. URL https://doi.org/10.1109/tcbb.2013.145.

58. S.-Y. Rhee, T. F. Liu, S. P. Holmes, and R. W. Shafer. HIV-1 subtype B protease and reverse transcriptase amino acid covariation. *PLOS Computational Biology*, 3(5):1–8, May 2007. doi: 10.1371/journal.pcbi.0030087.

59. R. Sanjuán and P. Domingo-Calap. Mechanisms of viral mutation. *Cellular and Molecular Life Sciences*, 73(23):4433–4448, July 2016. doi: 10.1007/s00018-016-2299-6. URL https://doi.org/10.1007/s00018-016-2299-6.

60. I. Sharon, S. Bercovici, R. Y. Pinter, and T. Shlomi. Pathway-based functional analysis of metagenomes. *J. Comput. Biol.*, 18(3):495–505, Mar. 2011.

61. M. Shen, Q. Li, M. Ren, Y. Lin, J. Wang, L. Chen, T. Li, and J. Zhao. Trophic status is associated with community structure and metabolic potential of planktonic microbiota in plateau lakes. *Front. Microbiol.*, 10:2560, Nov. 2019.

62. P. Skums, L. Bunimovich, and Y. Khudyakov. Antigenic cooperation among intrahost HCV variants organized into a complex network of cross-immunoreactivity. *Proceedings of the National Academy of Sciences*, 112(21):6653–6658, 2015. doi: 10.1073/pnas.1422942112.

63. P. Skums, A. Zelikovsky, R. Singh, W. Gussler, Z. Dimitrova, S. Knyazev, I. Mandric, S. Ramachandran, D. Campo, D. Jha, L. Bunimovich, E. Costenbader, C. Sexton, S. O'Connor, G.-L. Xia, and Y. Khudyakov. QUENTIN: reconstruction of disease transmissions from viral quasispecies genomic data. *Bioinformatics*, 34(1):163–170, June 2017. doi: 10.

1093/bioinformatics/btx402. URL https://doi.org/10.1093/bioinformatics/btx402.

64. P. Skums, A. Zelikovsky, R. Singh, W. Gussler, Z. Dimitrova, S. Knyazev, I. Mandric, S. Ramachandran, D. Campo, D. Jha, L. Bunimovich, E. Costenbader, C. Sexton, S. O'Connor, G.-L. Xia, and Y. Khudyakov. QUENTIN: Reconstruction of disease transmissions from viral quasispecies genomic data. *Bioinformatics*, 34(1):163–170, June 2017. doi: 10.1093/bioinformatics/btx402.

65. P. Skums, A. Kirpich, P. I. Baykal, A. Zelikovsky, and G. Chowell. Global transmission network of SARS-CoV-2: from outbreak to pandemic. Mar. 2020. doi: 10.1101/2020.03.22.20041145. URL https://doi.org/10.1101/2020.03.22.20041145.

66. A. Stamatakis. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313, Jan. 2014. doi: 10.1093/bioinformatics/btu033. URL https://doi.org/10.1093/bioinformatics/btu033.

67. A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.*, 102(43):15545–15550, Oct. 2005.

68. K. Tamura and M. Nei. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution*, May 1993. doi: 10.1093/oxfordjournals.molbev.a040023. URL https://doi.org/10.1093/oxfordjournals.molbev.a040023.

69. A. L. Tarca, S. Draghici, G. Bhatti, and R. Romero. Down-weighting overlapping genes improves gene set analysis. *BMC Bioinformatics*, 13:136, June 2012.

70. C. Wymant, M. Hall, O. Ratmann, D. Bonsall, T. Golubchik, M. de Cesare, A. Gall, M. Cornelissen, and C. F. and. PHYLOSCANNER: Inferring transmission from within- and between-host pathogen genetic diversity. *Molecular Biology and Evolution*, 35(3):719–733, Nov. 2017. doi: 10.1093/molbev/msx304. URL `https://doi.org/10.1093/molbev/msx304`.

71. Y. Ye and T. G. Doak. A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes. *PLoS Comput. Biol.*, 5(8):e1000465, Aug. 2009.