The Metagenomic Binning Problem: Clustering Markov Sequences

Grant Greenberg, *Member, IEEE*, and Ilan Shomorony, *Member, IEEE*

***Abstract*—The goal of metagenomics is to study the composition of microbial communities, typically using high-throughput shotgun sequencing. In the metagenomic binning problem, we observe random substrings (called contigs) from a mixture of genomes and aim to cluster them according to their genome of origin. Based on the empirical observation that genomes of different bacterial species can be distinguished based on their *tetranucleotide frequencies*, we model this task as the problem of clustering N sequences generated by M distinct Markov processes, where $M \ll N$. Utilizing the large-deviation principle for Markov processes, we establish the information-theoretic limit for perfect binning. Specifically, we show that the length of the contigs must scale with the inverse of the Chernoff divergence rate between the two most similar species. Furthermore, our result implies that contigs should be binned using the KL divergence rate as a measure of distance, as opposed to the Euclidean distance often used in practice.**

***Index Terms*—Biological information theory, metagenomics, Markov processes, clustering algorithms.**

I. Introduction

In THE last decade, advances in high-throughput DNA sequencing technologies have allowed a vast amount of genomic data to be generated. Countless tasks such as genome assembly, RNA quantification, and genome-wide association studies have become a reality, opening up exciting new research directions within biology and medicine [1].

Spurred by the advancements, several studies have utilized information-theoretic frameworks and principles to provide new insights into certain areas of bioinformatics. In [2], [3], the authors employ a probabilistic model for DNA sequence generation and characterize the fundamental limits of perfect sequence reconstruction as the genome size and number and length of sequencing reads, grow. In [4], [5], a similar procedure is utilized to determine under what conditions it is information-theoretically feasible to detect causal subsequences in GWAS studies. State-of-the-art transcriptome assemblers [6], [7] also draw from information theory principles to optimize the use of transcript abundances to overcome repeat patterns. In the context of this progress, we aim to provide a similar investigation in the area of metagenomics.

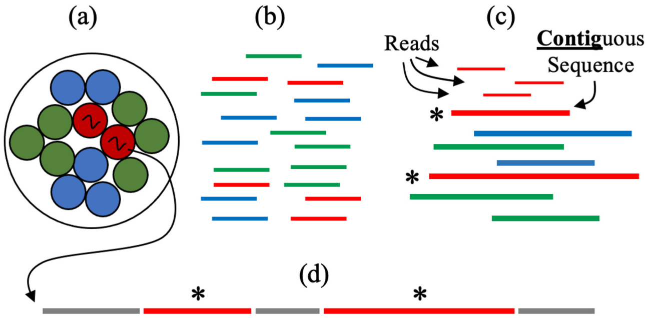

Fig. 1. **(a)** A metagenomic sample containing three species with various abundances. **(b)** The short DNA sequencing reads obtained from the sample. **(c)** The contigs obtained after assembling the reads. **(d)** The genome corresponding to the red species, with the position of the contigs labeled. Note that the contigs are non-overlapping and do not cover the full genome.

Significant attention has recently been given to the analysis of the human microbiome through metagenomics [8]. In metagenomics, a sample is taken from a microbial community, such as the human gut. The genetic material in the sample is then sequenced and analyzed to determine the microbial composition of the community [9]. Recent research, including the Human Microbiome Project [10], has shown that the composition of the microbiome is a "snapshot" of an individual's overall health, providing great potential for personalized medicine.

Full reconstruction of the genomes in a metagenomic sample is generally infeasible due to insufficient coverage and high similarity across species [11]. In the typical analysis pipeline, the millions of reads obtained via high-throughput sequencing are used to create a much smaller number of contiguous sequences, known as *contigs*, by merging reads with large overlaps [12]. This process is illustrated in Fig. 1. The set of resulting contigs typically make up only a small fraction of the full genomes of all species present in the sample and have no significant overlaps with each other.

Metagenomic binning is concerned with the following question: is it possible to group the resulting contigs based on the genome from which they were derived? Somewhat surprisingly, it has been shown that contigs belonging to the same species typically have similar sequence compositions. Specifically, it was empirically verified that the distribution of four-letter strings (e.g., *AGCG*) remains relatively constant across an entire bacterial genome [13], [14], [15]. Even plasmids and bacteriophages corresponding to a bacterial

Manuscript received 5 May 2023; revised 5 October 2023; accepted 15 November 2023. Date of publication 28 November 2023; date of current version 18 March 2024. The work of Grant Greenberg and Ilan Shomorony was supported in part by the National Science Foundation (NSF) under Grant CCF-2007597 and Grant CCF-2046991. The associate editor coordinating the review of this article and approving it for publication was Y. Hori. *(Corresponding author: Grant Greenberg.)*

The authors are with the Department of Electrical and Computer Engineering, University of Illinois at Urbana–Champaign, Urbana, IL 61801 USA (e-mail: gcgreen2@illinois.edu).

Digital Object Identifier 10.1109/TMBMC.2023.3336254

species tend to have similar nucleotide compositions as the chromosome [16], [17]. Hence one can compute for each contig the *tetranucleotide frequency* (TNF) vector, and group together contigs with "similar" TNF vectors. Provided that the underlying TNF distributions are distinct enough, metagenomic binning can thus be performed.

Based on this idea, many different algorithms and software packages have been proposed to perform metagenomic binning [11], [18]. Usually, in addition to the TNF vector, the read coverage of each contig is used as another feature to help with the clustering. However, in this study, we focus on using the TNF alone, which has been shown in previous studies to contain significant statistical power [12]. Moreover, we show in Section V that clustering with TNF alone largely preserves biological information. Other algorithms use a supervised learning approach by comparing the sequence composition of reads to a database of known bacterial genomes [19], [20], [21], [22], or through direct alignment to said database [23], [24].

The fact that the distribution of four-symbol strings is consistent throughout a given genome motivates the modeling of each genome as a *third*-order Markov process. Hence, we assume that a contig is generated by one out of M distinct, *unknown* Markov processes $p_1, \ldots, p_M$ with equal probability, where each p_k corresponds to a certain species. To study the fundamental limits of this problem, we assume all N contigs have length L, and consider an asymptotic regime where $N \to \infty$, and the contig length grows slowly with the number of contigs. Specifically, we set $L = \bar{L} \log N$, which is not based on any natural phenomenon, but which we show is the correct length scaling based on our model (e.g., $L = N^2$ makes the problem too easy, and $L = N$ makes the problem too challenging). Our goal is to characterize how large $\bar{L}$ needs to be to allow perfect binning with high probability. Note that in practice L and N are essentially observed variables, whereas they are controlled parameters in our framework. Nonetheless, the results presented in this work indicate that studying our model provides valuable insights for a practical setting.

To obtain our main result, we establish the equivalent of the Chernoff Information [25, Ch. 11.9] for Markov processes, which gives the error exponent for the Bayesian error probability when testing between two known Markov processes. This result, combined with a scheme to estimate the M Markov distributions, allows us to show that perfect binning is possible if and only if

$$\bar{L} > \frac{1}{\min_{k,\ell} C(p_k, p_\ell)},$$

where $C(p_k, p_\ell)$ is the Chernoff divergence rate between p_k and p_ℓ. To estimate the unknown distributions, we consider building a graph where contigs whose empirical distributions are close are connected. We then show that, with high probability, M large cliques can be found, which can be used to find estimates $\tilde{p}_k$, $k = 1, \ldots, M$, of the Markov distributions. Each contig $\mathbf{x}$ is then placed in bin k given by

$$\underset{k \in \{1,\ldots,M\}}{\arg\min} \quad D(\hat{p}_{\mathbf{x}} \| \tilde{p}_k)$$

where $\hat{p}_{\mathbf{x}}$ is the empirical 4-symbol distribution of $\mathbf{x}$ and $D(\cdot\|\cdot)$ is the KL divergence rate (essentially the KL divergence for random processes) between Markov processes with distributions $\hat{p}_{\mathbf{x}}$ and $\tilde{p}_k$.

Our main result suggests that the optimal way to bin metagenomic contigs is to estimate the underlying TNF distributions and then bin contigs using the KL divergence rate as a metric, as opposed to the commonly used Euclidean distance. By simulating contigs from real bacterial genomes, we show that this metric can lead to lower binning error probabilities.

The paper is organized as follows. In Section II we describe the problem formulation in detail and state our main result. In Section III we describe our achievability scheme and the main technical ingredients used to prove it, and in Section IV we describe the converse argument. In Section V we provide preliminary simulation results, and we conclude the paper with a discussion in Section VI.

II. Problem Statement

As shown in [13], the distribution of tetranucleotides (four-letter strings), tends to be stationary across an individual bacterial genome. Hence, it is natural to assume that each of the species in our sample corresponds to a distribution over all possible tetranucleotides[1] $\{AAAA, AAAC, \ldots, TTTT\}$.

Let $\mathcal{P}$ be the $|\mathcal{X}|^4$-dimensional simplex, where $\mathcal{X} = \{A, C, G, T\}$. Notice that not all distributions in $\mathcal{P}$ are valid tetranucleotide distributions, as the tetranucleotides in a sequence overlap with each other. Let $\widetilde{\mathcal{P}}$ be the set of all $p \in \mathcal{P}$ with $p(\mathbf{c}) > 0$, $\forall\, \mathbf{c} \in \mathcal{X}^4$, which, in addition, satisfy for all $\mathbf{a} \in \mathcal{X}^3$

$$\sum_{b \in \mathcal{X}} p(\mathbf{a}b) \;=\; \sum_{b \in \mathcal{X}} p(b\mathbf{a}). \tag{1}$$

Condition (1) ensures that a given $p \in \widetilde{\mathcal{P}}$ corresponds to the tetranucleotide distribution of a specific, *stationary*, irreducible (due to $p(\mathbf{c}) > 0$), third-order Markov chain. More precisely, we can let the induced distribution over 3-letter strings be

$$p(\mathbf{a}) = \sum_{b \in \mathcal{X}} p(\mathbf{a}b). \tag{2}$$

This uniquely determines a stationary Markov process with initial state distributed as (2) and transition probabilities (i.e., conditional distribution)

$$p(b|\mathbf{a}) = \frac{p(\mathbf{a}b)}{p(\mathbf{a})}. \tag{3}$$

Hence, we will model each species in the sample using a distribution $p_k \in \widetilde{\mathcal{P}}$.

A. Metagenomic Binning Problem

We assume that we have M species in our sample (for a known M). In our framework, each species is modeled by a stationary third-order Markov process defined by $p_k \in \widetilde{\mathcal{P}}$, for $k = 1, \ldots, M$. From this genomic mixture, we observe a set of N realizations $\mathcal{Y} = \{\mathbf{x}_i\}_{i=1}^N$, which we call *contigs*. Each $\mathbf{x} \in \mathcal{Y}$ is generated independently by first choosing a species

[1] In practical approaches, *reverse-complementary* tetranucleotides such as *ACAG* and *CTGT* are treated as the same tetranucleotide, but we ignore that fact for the sake of simplicity.

$k \in \{1, \ldots, M\}$ with prior probabilities $\{\pi_k\}_{k=1}^M$, and then generating a length-L sequence according to p_k. The priors are unknown, but we assume they are finite and do not change with N. Furthermore, we assume to know a lower bound on the minimum prior, which we call $\pi_{\min}$; we argue this is a reasonable assumption since, in practice, one can expect to recover species only above a minimum abundance.

For each k, let $\mathcal{C}_k$ be the set of contigs generated according to p_k. We wish to reconstruct $\mathcal{C}_k$, $k = 1, \ldots, M$, by determining which contigs originated from the same genome.

We point out that in real metagenomic experiments, the *coverage depth*, that is, the expected number of contigs containing a specific nucleotide from one of the M genomes, is low [26]. Hence, contigs will have no overlap with high probability, allowing us to model them as independent realizations of the different Markov processes in the sample.

B. Perfect Binning

The goal of the metagenomic binning problem is to cluster the N contigs into M "bins", where each bin k corresponds to a unique species with distribution p_k. More precisely, the goal is to find a decision rule $\delta : \mathcal{X}^L \to \{1, \ldots, M\}$ (using notation from [27]) which correctly maps each contig to its respective genome bin.

Perfect binning would be achieved if for every contig $\mathbf{x}$, $\delta(\mathbf{x})$ chooses the label of the distribution from which it was generated. However, we have the added difficulty that the distributions are unknown. As a result, we can only require the decision rule to be correct up to a consistent relabeling of species indices. Hence, the error event for a decision rule δ is

$$\mathcal{E}_\delta = \{\exists \mathbf{x} \in \mathcal{C}_k, \mathbf{y} \in \mathcal{C}_\ell, k \neq \ell : \delta(\mathbf{x}) = \delta(\mathbf{y})\}. \tag{4}$$

We would like to know under what circumstances we can perfectly bin all N contigs. In order to study the information-theoretic limits of this problem, we analyze an asymptotic regime, similar to [2], in which $N \to \infty$ and

$$L = \bar{L} \log N \tag{5}$$

where $\bar{L}$ is the "normalized contig length". Intuitively, a larger value of $\bar{L}$ should allow one to bin a contig with higher accuracy. This scaling forces the contig length to be small compared to the number of contigs and, as we will show, is a meaningful scaling for the asymptotic problem we consider.

This asymptotic regime allows us to define when species are resolvable as follows:

Definition 1: The M species with distributions $\{p_k\}_{k=1}^M$ are **resolvable** if there exists a sequence of decision rules $\{\delta_N\}$ such that $\Pr(\mathcal{E}_{\delta_N}) \to 0$ as $N \to \infty$.

C. Main Result

Interestingly, the fundamental limit of resolvability relies on the Chernoff divergence rate, which we define next. The Chernoff divergence rate can be thought of as a measure of distance between the distributions. Specifically, it is the KL divergence rate between one of the distributions and the closest distribution that is equidistant (by KL divergence rate) to both distributions.

Definition 2: For two Markov processes p_k and p_ℓ, the **Chernoff divergence rate** between p_k and p_ℓ is given by

$$C(p_k, p_\ell) = D(p^* \| p_k) = D(p^* \| p_\ell) \tag{6}$$

where D is the KL divergence rate (Eq. (36)), and p^* is the solution to the following minimization problem.

$$\begin{aligned} p^* &= \arg\min_{p \in \widetilde{\mathcal{P}}} D(p \| p_k) \\ s.t.\ D(p\|p_k) &= D(p\|p_\ell) \end{aligned} \tag{7}$$

In [27, Sec. 10.1.3], an efficient way to calculate the Chernoff divergence rate is given.

Our main result establishes that the minimum normalized contig length, $\bar{L}$, required for *resolvability* depends exclusively on the minimum Chernoff divergence rate between species distributions.

Theorem 1: Let $C_{\min} = \min_{k \neq \ell} C(p_k, p_\ell)$. The species' distributions $\{p_k\}_{k=1}^M$ are resolvable if and only if

$$\bar{L} > \frac{1}{C_{\min}} \tag{8}$$

Intuitively, this means that the contig length must be large enough to distinguish between the two closest distributions.

III. Achievability

The achievability proof of Theorem 1 is described in the form of an algorithm to highlight the algorithmic nature of metagenomic binning. Algorithm 1 first estimates the species distributions by finding large cliques in the distance-thresholded graph, then averaging the empirical distributions of the contigs in large cliques. Finally, it bins the contigs based on the estimates. Note that the algorithm as described is not computationally efficient (specifically, finding large cliques) and is used only to establish the achievability of Thm. 1. Given a contig $\mathbf{x} \in \mathcal{Y}$, we define the empirical fourth-order distribution of $\mathbf{x}$ as $\hat{p}_\mathbf{x}$ and we use d as the L_1 distance between distributions, i.e., $d(p, q) = \sum_{\mathbf{c} \in \mathcal{X}^4} |p(\mathbf{c}) - q(\mathbf{c})|$. We will let $d_{\min} \triangleq \min_{k \neq \ell} d(p_k, p_\ell)$ be the minimum L_1 distance between any pair of the M species distributions.

A. Estimating Distributions

Recall that $\mathcal{C}_k$ is the set of contigs generated by p_k. We expect the empirical distribution of the majority of contigs in $\mathcal{C}_k$ to be near p_k. To identify those "good contigs", let

$$\mathcal{C}_{k,\epsilon} = \{\mathbf{x} \in \mathcal{C}_k : d(\hat{p}_\mathbf{x}, p_k) \leq \frac{\epsilon}{2}\}.$$

To prove that the distribution estimates $\{\tilde{p}_k\}_{k=1}^M$ are close to the true distributions $\{p_k\}_{k=1}^M$ (after proper reindexing), we use three lemmas, which demonstrate that 1) sufficiently large cliques exist in the graph that 2) are each close to a different species distribution and 3) are not "contaminated" by a significant number of contigs from a different species. The proofs of each lemma can be found in Appendix.

We begin by establishing that M sufficiently large cliques will exist in $\mathcal{G}_\epsilon$ with high probability using the following

Algorithm 1: Decision Rule to Bin All Contigs. The Algorithm Searches for the Smallest Threshold ϵ Which Gives Rise to Sufficiently Large Cliques. It Then Uses the Cliques to Estimate the Corresponding Species Distributions, Which It Uses to Bin the Contigs

```
Result: Decision Rule $\delta(\mathbf{x})$
Input: Contigs $\mathcal{Y}$, Parameter $\alpha \in (0,1)$
begin
    $\mathcal{D} \longleftarrow \text{sortinascendingorder}\{d(\hat{p}_{\mathbf{x}}, \hat{p}_{\mathbf{y}}), \forall \mathbf{x}, \mathbf{y} \in \mathcal{Y}\}$
    for $\epsilon$ in $\mathcal{D}$
        $\mathcal{G}_\epsilon \longleftarrow \big(V = \mathcal{Y}, E_\epsilon = \{(\mathbf{x},\mathbf{y}) : d(\hat{p}_{\mathbf{x}}, \hat{p}_{\mathbf{y}}) \le \epsilon\}\big)$
        if $\mathcal{G}_\epsilon$ has disjoint cliques $\{\mathcal{K}_k\}_{k=1}^M : |\mathcal{K}_k| \ge (1-\alpha)N\pi_{\min},\ \sum_{k=1}^M |\mathcal{K}_k| \ge (1-\alpha)N$
            for $k \longleftarrow 1$ to $M$
                $\tilde{p}_k \longleftarrow \frac{1}{|\mathcal{K}_k|}\sum_{\mathbf{x}\in\mathcal{K}_k} \hat{p}_{\mathbf{x}}$
            break
    for $\mathbf{x} \in \mathcal{Y}$
        $\delta(\mathbf{x}) \longleftarrow \arg\min_{k\in\{1,\ldots,M\}} D(\hat{p}_{\mathbf{x}} \| \tilde{p}_k)$
```

lemma, which says that a large fraction of the contigs will be close to their respective generating distributions.

Lemma 1: For $\epsilon > 0$, $k \in \{1, \ldots, M\}$ and N large enough,

$$\Pr\big(|\mathcal{C}_{k,\epsilon}| < (1-\alpha)N\pi_k\big) \le e^{-\gamma\alpha^2 L - \log\alpha}, \tag{9}$$

where γ is a positive constant. Moreover, by the triangle inequality, any two contigs $\mathbf{x}$, $\mathbf{y} \in \mathcal{C}_{k,\epsilon}$ will be at a distance ϵ or less, and thus, $\mathcal{C}_{k,\epsilon}$ forms a clique in $\mathcal{G}_\epsilon$.

Fixing $\epsilon < \frac{d_{\min}}{2}$, Lemma 1 guarantees that for a reasonably chosen α, and large enough N, the right side of (9) will be small, and thus, we will have $|\mathcal{C}_{k,\epsilon}| \ge (1-\alpha)N\pi_k$ with high probability.

Next, we present Lemma 2, which establishes that the cliques and species have a one-to-one mapping. Recall that $\pi_{\min}$ is a lower bound on the minimum species prior.

Lemma 2: For $\epsilon < \frac{d_{\min}}{2}$, if Algorithm 1 finds cliques $\{\mathcal{K}_k\}_{k=1}^M$, then there exists a bijection $\sigma : \{1, \ldots, M\} \to \{1, \ldots, M\}$ such that, for each k,

$$\mathcal{K}_k \cap \mathcal{C}_{\sigma(k),\epsilon} \neq \emptyset, \tag{10}$$

$$\forall \ell \neq \sigma(k) : \mathcal{K}_k \cap \mathcal{C}_{\ell,\epsilon} = \emptyset \tag{11}$$

with probability $1 - o(1)$.

Notice that $d_{\min}$ is not known, so the algorithm cannot restrict its search to $\epsilon < \frac{d_{\min}}{2}$. However, since the algorithm considers different values of ϵ in increasing order, for some $\epsilon < \frac{d_{\min}}{2}$, M cliques satisfying the constraints in Algorithm 1 will exist with probability $1 - o(1)$ (meaning that for any probability arbitrarily close to 1, one can find an N_0 such that all $N > N_0$ achieves that probability or greater).

Now that we have established that sufficiently large cliques will exist covering all species distributions with high probability, we finally use Lemma 3, which says that fraction of "good contigs" in each clique goes to one.

Lemma 3: Let σ be the bijection from Lemma 2 which maps the clique index to the species index. If $\alpha \to 0$ as $N \to \infty$, then

$$\frac{|\mathcal{K}_k \cap \mathcal{C}_{\sigma(k),\epsilon}|}{|\mathcal{K}_k|} \to 1 \tag{12}$$

with probability $1 - o$ (1).

If we set $\alpha = \frac{1}{\log L}$, (12) holds and (9) converges to 0 as $N \to \infty$. Thus, with high probability, a vanishing fraction of the contigs in $\mathcal{K}_k$ does not belong to $\mathcal{C}_{\ell,\epsilon}$. Since distribution vectors are bounded, the impact of wrong contigs in $\mathcal{K}_k$ on $\tilde{p}_k$ also vanishes, and we conclude that the distribution estimate $\tilde{p}_k = \frac{1}{|\mathcal{K}_k|}\sum_{\mathbf{x}\in\mathcal{K}_k} \hat{p}_{\mathbf{x}} \to p_{\sigma(k)}$ as $N \to \infty$.

B. Binning Contigs

In Section III-A, we established that we can construct estimates of the underlying distributions $\{p_k\}_{k=1}^M$ that are arbitrarily accurate as $N \to \infty$. Next, we show that, binning the contigs based on the KL divergence rate using the underlying distributions achieves (8) in the limit.

Consider the hypothesis test between two Markov processes p_k and p_ℓ (assumed to be known). Given prior probabilities π_k and π_ℓ, the Bayesian probability of error is

$$\pi_k \Pr(\text{choose } \ell | k \text{ true}) + \pi_\ell \Pr(\text{choose } k | \ell \text{ true})$$

for the decision rule on a contig generated by either p_k or p_ℓ.

Theorem 2: Let $\mathcal{E}_{k,\ell}^{(L)}$ be the error event for the decision rule which minimizes the Bayesian probability of error. Then

$$\lim_{L\to\infty} \frac{1}{L} \log \Pr\left(\mathcal{E}_{k,\ell}^{(L)}\right) = -C(p_k, p_\ell), \tag{13}$$

i.e., $C(p_k, p_\ell)$ is the optimal error exponent.

The proof of Theorem 2 is given in Appendix. For a given contig, the last step of Algorithm 1 can be thought of as $M - 1$ binary hypothesis tests between the true distribution and each of the remaining distributions. Thus, we will use Theorem 2 to bound the overall error probability, $\Pr(\mathcal{E}_{\delta_N})$, by considering the two closest distributions.

$$\Pr\left(\mathcal{E}_{\delta_N}\right) \le \sum_{i=1}^N \sum_{k=1}^M \pi_k \sum_{\ell\neq k} \Pr\left(\mathcal{E}_{k,\ell}^{(L)}\right) \tag{14}$$

$$\le N(M-1) \max_{k\neq\ell} \Pr\left(\mathcal{E}_{k,\ell}^{(L)}\right) \tag{15}$$

$$\le M 2^{L\left(1/\bar{L} + \max_{k\neq\ell}(1/L)\log\Pr(\mathcal{E}_{k,\ell}^{(L)})\right)} \tag{16}$$

where (14) follows from the union bound. By Theorem 2,

$$\max_{k\neq\ell} \frac{1}{L} \log \Pr\left(\mathcal{E}_{k,\ell}^{(L)}\right) \to -\min_{k\neq\ell} C(p_k, p_\ell) = -C_{\min}$$

as $N \to \infty$. Hence, if $\bar{L} > \frac{1}{C_{\min}}$, $\Pr(\mathcal{E}_{\delta_N}) \to 0$. Consider the case when instead we have estimates of the true distributions. The decision boundary for the optimal assignment of contigs is continuous on $\{\tilde{p}_k\}_{k=1}^{M}$. Since each $\tilde{p}_k \to p_\ell$, a continuity argument can be used to show that the probability of error of the binary hypothesis test converges to the same value as the distributions converge to the true ones. Then, it follows that the overall error probability converges to (16). This concludes the achievability proof of Theorem 1.

IV. Converse

Without loss of generality, let p_1 and p_2 be such that $C_{\min} = C(p_1, p_2)$. Given the decision rule δ_N, contigs $\mathbf{x}_1 \in \mathcal{C}_1$ and $\mathbf{x}_2 \in \mathcal{C}_2$, and a contig $\mathbf{x} \in \mathcal{Y}$, let

$$\begin{aligned} \widetilde{\mathcal{E}}_{1,2,\mathbf{x}} = \{&\mathbf{x} \in \mathcal{C}_1, \delta_N(\mathbf{x}) \neq \delta_N(\mathbf{x}_1)\} \\ &\cup \{\mathbf{x} \in \mathcal{C}_2, \delta_N(\mathbf{x}) \neq \delta_N(\mathbf{x}_2)\} \end{aligned}$$

i.e., the event that $\mathbf{x}$ was generated by either p_1 or p_2 and incorrectly binned. Note that $\Pr(\widetilde{\mathcal{E}}_{1,2,\mathbf{x}}) \geq 2\pi_{\min}\Pr(\mathcal{E}_{1,2})$. Then

$$\begin{aligned} \Pr(\mathcal{E}_{\delta_N}) &\geq \Pr\left(\bigcup_{i=1}^{N} \widetilde{\mathcal{E}}_{1,2,\mathbf{x}_i}\right) \\ &= 1 - \left(1 - \Pr(\widetilde{\mathcal{E}}_{1,2,\mathbf{x}_1})\right)^N \\ &\geq 1 - \left[\left(1 - 2\pi_{\min}\Pr(\mathcal{E}_{1,2})\right)^{1/\Pr(\mathcal{E}_{1,2})}\right]^{N\Pr(\mathcal{E}_{1,2})} \\ &\geq 1 - e^{-2\pi_{\min} N \Pr(\mathcal{E}_{1,2})} \end{aligned} \tag{17}$$

where (17) follows from the bound $(1 - ap)^{1/p} \leq e^{-a}$ for $p \in (0, 1]$, $a \in \mathbb{R}$. We see that, if $N\Pr(\mathcal{E}_{1,2}) \not\to 0$, then $\Pr(\mathcal{E}) \not\to 0$. Since

$$N\Pr(\mathcal{E}_{1,2}) = 2^{L\left(1/\bar{L} + (1/L)\log\Pr(\mathcal{E}_{1,2})\right)},$$

then by Theorem 2, $N\Pr(\mathcal{E}_{1,2}) \not\to 0$ when $\bar{L} \leq \frac{1}{C_{\min}}$. This concludes the converse proof for Theorem 1.

V. Experimental Results

From the point of view of practical metagenomic binning algorithms, our main result suggests that:

1) the KL divergence rate is a good metric for binning contigs,
2) the Chernoff divergence rate can be used as a measure of how difficult it is to distinguish two species.

In this section, we provide preliminary empirical evidence of these claims through two sets of experiments. For the first set of experiments shown in Figures 2a-c, we utilized several previously sequenced and assembled bacterial genomes, available at NCBI [28]. For each bacterial species k, we numerically computed its fourth-order distribution p_k (i.e., the overall tetranucleotide frequency vector). We were able to simulate contigs of a desired length L by sampling from all length-L substrings from the genome. For each experiment, we assume $N = 10^6$ for concreteness (thus $L = \bar{L}\log 10^6$). Note that the results are not meaningfully affected by the choice of N. For example, setting $N = 10^4$ would “stretch out” the plots in Fig. 2a and 2c compared to $N = 10^6$, since a larger $\bar{L}$ is needed to obtain the same contig length, but the actual error values and structure of the graphs will not change.

To verify the usefulness of the KL divergence rate and compare it to the Euclidean distance (used in state-of-the-art tools such as [11], [18]), we considered the following experiment: we extracted random contigs from a species p_1 and then tested whether it was closer to species p_1 or to another species p_2 based on both the Euclidean distance and the KL divergence rate. In Figure 2a, the KL divergence rate metric[2] consistently outperforms the Euclidean metric as we vary $\bar{L}$ in the test between the species *Alistipes obesi* and *Megamonas funiformi*.

We performed this experiment for 45 different choices of pairs of bacterial genomes from NCBI. For each pair (k, ℓ), we considered a fixed normalized contig length given by $\bar{L} = C(p_k, p_\ell)^{-1}$. As shown in Figure 2b, the conditional divergence improves the error compared to the Euclidean distance in almost 90% of cases.

Theorem 1 implies that an appropriate similarity measure is the inverse of the Chernoff divergence rate since it characterizes how long the extracted contigs need to be in order for two species to be reliably distinguishable. To verify that, we calculated $\bar{L}_{5\%}$, the minimum normalized contig length required to guarantee a 5% error rate in the Bayesian hypothesis test between species p_k and p_ℓ with equal priors. In Figure 2c, we plot $\bar{L}_{5\%}$ vs $C^{-1}(p_k, p_\ell)$ for many such pairs and observe a roughly linear relationship between these two quantities. Such a linear relationship agrees with the relationship suggested by Theorem 1. Moreover, it provides support to the claim that $C^{-1}(p_k, p_\ell)$ is a measure of how difficult it is to distinguish contigs from two species based on tetranucleotide frequencies.

In the second experiment, shown in Figures 2d-f, we use metagenomic datasets simulated by CAMISIM [29] for the first Critical Assessment of Metagenome Interpretation (CAMI) challenge [30]. The datasets are simulated to mimic the taxonomic profiles of the Human Microbiome Project [31] for the Gastrointestinal (samples 0 and 1) and Oral (samples 6 and 7) microbiomes. We preprocess each dataset by removing contigs shorter than 1kbp, as well as all contigs from families that represent less than 0.1% of the dataset (most of these were complete or nearly complete metagenomes).

We aim to strengthen the evidence that KL divergence rate is a more effective measure compared to Euclidean distance when binning contigs. To this end, we perform k-means clustering for each dataset, $\mathcal{Y}$. In the first step, we initialize a set of cluster means $\{\mu_i^{(0)}\}_{i=1}^{k}$, $\mu_i \in \mathbb{R}^2 56$, by choosing k contigs at random and calculating their TNFs. In k-means we iterate between an assignment step, where we partition the dataset into clusters

$$S_i^{(t)} = \{x \in \mathcal{Y} : i = \operatorname*{arg\,min}_{j=1,\ldots,k} d(x, \mu_j)\},$$

[2] $D(\cdot||\cdot)$ is not technically a metric as it is not symmetric.

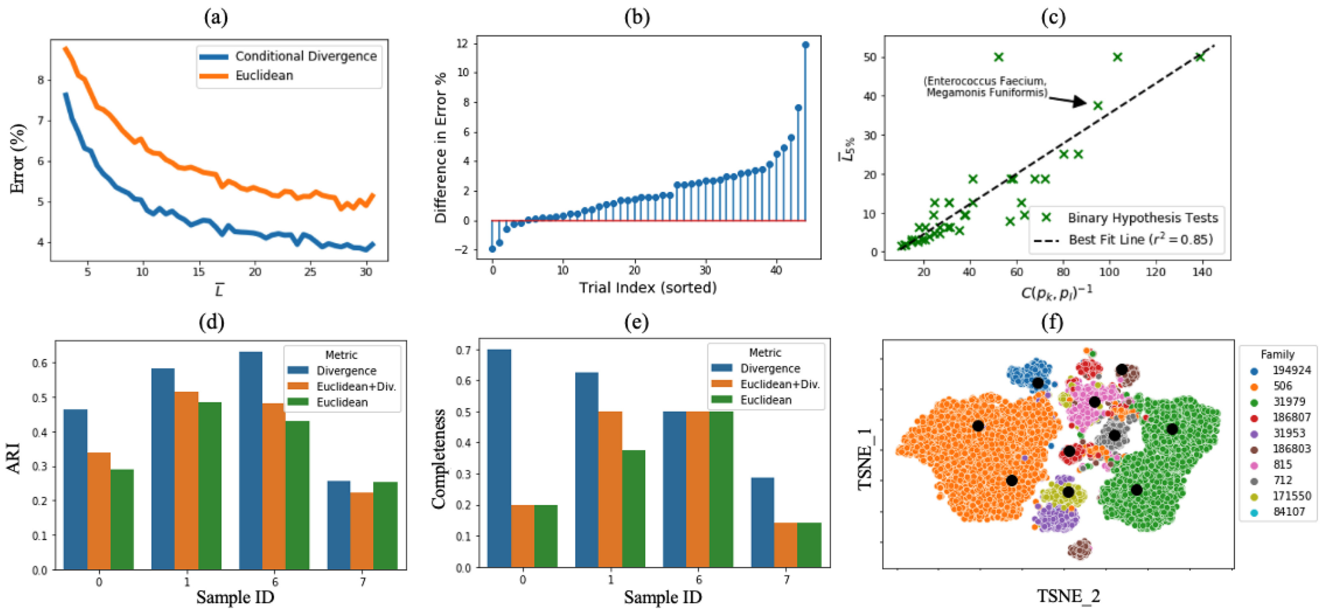

Fig. 2. **(a)** Comparison of KL divergence rate (conditional divergence) and Euclidean distance for a hypothesis test between *Alistipes obesi* and *Megamonas funiformis*; **(b)** A stem plot of the (sorted) improvement in performance from Euclidean distance to KL divergence rate with $\bar{L} = C(p_k, p_\ell)^{-1}$; **(c)** Normalized contig length required for 5% error $(\bar{L}_{5\%})$ vs species similarity as measured by the inverse of the Chernoff divergence rate for several pairs of species. **(d,e)** Bar graphs of adjusted Rand index and completeness, respectively, for k-means clustering on several CAMI samples using three different measures for the assignment step. **(f)** t-SNE plot of the first HMP CAMI sample, colored by family (with taxonomic IDs in the legend).

and an update step,

$$\mu_i^{(t)} = \frac{1}{|S_i^{(t)}|} \sum_{x \in S_i^{(t)}} x.$$

We use three different distance measures, $d(\cdot,\cdot)$, for the assignment of contigs to means[3]: 1) KL divergence rate, 2) Euclidean distance, and 3) KL divergence rate for only the final assignment, much like Algorithm 1 of the achievability proof. Note that the KL divergence rate requires the additional calculation of the (stationary) third-order distribution to determine. For each sample, the k chosen is equal to the number of families in the sample. In practice, this quantity cannot be known, but there exist several heuristics to choose an appropriate number of clusters, including the "elbow method", AIC [32], and BIC [33]. Furthermore, methods such as [11] use a greedy method for binning, obviating the need to choose a k-value.

To evaluate our simple binning scheme, we calculate purity and completeness, which are typical binning metrics [11], [12], as well as adjusted Rand index (ARI). For each predicted bin, corresponding to family f (i.e., by majority vote of ground-truth labels in the bin), purity measures the contamination by families other than f, and completeness, the fraction of all contigs from f the bin contains. ARI measures the fraction of all pairs of contigs in either the same or different clusters in both the predicted clustering (bins) and the ground truth clustering (families). In Figure 2e, we see that the fraction of complete clusters (greater than 50%) significantly increases when using KL divergence rate for three of four samples. Similarly, the KL divergence greatly improves the ARI for three of four samples, indicating that KL divergence rate is indeed the correct measure to capture the biology of bacterial reads. Additionally, note that performing the final assignment step using KL divergence rate generally improves the performance metrics as well. We chose to provide the purity values in Fig. 3a in the Appendix since the results were similar across binning methods, with KL divergence outperforming Euclidean distance in one dataset, and matching in the remaining three.

VI. Discussion

In this paper, we modeled the metagenomic binning problem as the problem of clustering sequences generated by distinct Markov processes. While overly simplistic, this model allowed us to establish the Chernoff divergence rate as a measure of how easy it is to distinguish contigs generated by two species.

The algorithm used to prove the achievability suggests that a good "metric" for binning is the KL divergence rate between a contig and an estimate of a species TNF. Through experiments, we provided preliminary evidence that this metric often outperforms the Euclidean metric in the problem of assigning a contig to a species bin. However, this assumes knowledge of the overall TNF of a genome, which is not known in practical settings. Therefore, a natural direction for future investigation is how to efficiently estimate the TNF distribution for the species present in the sample.

Furthermore, it is unclear whether estimating the underlying TNF distributions is necessary to achieve the fundamental limit. Alternatively, one could consider an approach that directly clusters the contigs based on their pairwise distances or based on a graph obtained by thresholding the distances

[3] Though k-means clustering method is designed to minimize the sum of Euclidean distances, we believe that it is nonetheless natural to use KL divergence rate for the assignment step despite no convergence guarantees.

(similar to our $\mathcal{G}_\epsilon$). We point out that, for such a graph, the problem becomes a community detection problem, and bears similarities with the stochastic block model [34], since for each species there is a given probability that an edge is placed among two of its contigs, and for each pair of distinct species, there is another probability that an edge is placed between their contigs. These probabilities would in general depend on the species TNF distributions (or the Markov processes generating the contigs). Notice that, unlike in the standard stochastic block model, here the placing of the edges would not be independent events.

Finally, we point out that in most approaches to metagenomic binning, the read coverage, or *abundance*, is used to compare contigs in addition to the TNF. The read coverage of a contig is essentially the average number of reads that cover any given base in the contig. Intuitively, this number is proportional to the abundance of the corresponding species in the mixture. Hence, one expects contigs from the same species to have similar read coverages, which can be used to improve metagenomic binning. Another direction for future work is thus to consider the metagenomic binning problem where the different species have different abundances and, for each contig, one observes a read coverage value that is related to the species abundance.

Appendix

A. Proof of Lemma 1

Recall from Section II that $\mathcal{C}_k$ is the set of contigs generated according to p_k. Let $\mathcal{E}_k$ be the event of interest, $\{|\mathcal{C}_{k,\epsilon}| < (1-\alpha)N\pi_k\}$, and let $\mathcal{A}_k = \{|\mathcal{C}_k| < (1 - \frac{\alpha}{2})N\pi_k\}$. Note that we use $\frac{\alpha}{2}$ for $\mathcal{A}_k$ as opposed to α because we need $|\mathcal{C}_k|$ to be larger than $|\mathcal{C}_{k,\epsilon}|$. By Hoeffding's inequality,

$$\Pr(\mathcal{A}_k) \leq \exp\left(-2\frac{\left(\frac{\alpha}{2}N\pi_k\right)^2}{N}\right) = \exp\left(-N\frac{\alpha^2\pi_k^2}{2}\right)$$

This means, with high probability, p_k will generate enough contigs.

Let $\mathcal{F}_k$ be the set of distributions "far" from p_k:

$$\mathcal{F}_k = \left\{p \in \widetilde{\mathcal{P}} : d(p_k, p) \geq \frac{\epsilon}{2}\right\}. \tag{18}$$

By a version of Sanov's theorem for Markov chains, given in Theorem 4 in Appendix D, for any $\mathbf{x} \in \mathcal{C}_k$,

$$\Pr(\hat{p}_\mathbf{x} \in \mathcal{F}_k) \leq (L+1)^4 2^{-LD(p^*\|p_k)} \tag{19}$$

where $p^* = \arg\inf_{p\in\mathcal{F}_k} D(p\|p_k)$; i.e., p^* is the distribution in $\mathcal{F}_k$ closest to p_k in KL divergence rate. Notice that $\mathcal{E}_k$ occurs when more than $|\mathcal{C}_k| - (1-\alpha)N\pi_k$ contigs lie in $\mathcal{F}_k$, leaving an insufficient number of "good" contigs. Letting $\mathbf{x}_0 \in \mathcal{C}_k$ be some contig generated by p_k,

$$\begin{aligned}
&\Pr(\mathcal{E}_k|\mathcal{A}_k^\mathsf{c}) \\
&\quad = \Pr\left(\sum_{\mathbf{x}\in\mathcal{C}_k} 1\{\hat{p}_\mathbf{x} \in \mathcal{F}_k\} > |\mathcal{C}_k| - (1-\alpha)N\pi_k \middle| \mathcal{A}_k^\mathsf{c}\right) && (20)\\
&\quad \leq \Pr\left(\sum_{\mathbf{x}\in\mathcal{C}_k} 1\{\hat{p}_\mathbf{x} \in \mathcal{F}_k\} \geq \frac{\alpha}{2}N\pi_k \middle| \mathcal{A}_k^\mathsf{c}\right) && (21)\\
&\quad \leq \frac{2}{\alpha\pi_k}\cdot\Pr(\hat{p}_{\mathbf{x}_0} \in \mathcal{F}_k|\mathcal{A}_k^\mathsf{c}) && (22)
\end{aligned}$$

where (21) follows the definition of $\mathcal{A}_k$, and (22) from Markov's inequality and symmetry across contigs. Combining the probabilities,

$$\begin{aligned}
\Pr(\mathcal{E}_k) &= \Pr(\mathcal{E}_k|\mathcal{A}_k^\mathsf{c})\Pr(\mathcal{A}_k^\mathsf{c}) + \Pr(\mathcal{E}_k|\mathcal{A}_k)\Pr(\mathcal{A}_k) && (23)\\
&\leq \frac{2}{\alpha\pi_k}\cdot\Pr(\hat{p}_{\mathbf{x}_0} \in \mathcal{F}_k) + \Pr(\mathcal{A}_k) && (24)\\
&\leq \frac{2}{\alpha\pi_k}(L+1)^4 2^{-LD(p^*\|p_k)} + \exp\left(-N\frac{\alpha^2\pi_k^2}{2}\right) && (25)\\
&\leq e^{-\gamma\alpha^2 L - \log\alpha} && (26)
\end{aligned}$$

where $\gamma > 0$ is a constant that does not depend on α or L, guaranteed to exist such that (26) holds for large N. γ can be found by manipulating (25) using simple algebraic operations.

B. Proof of Lemma 2

Given the constraints from Algorithm 1,

$$\begin{aligned}
|\mathcal{K}_k| &\geq (1-\alpha)N\pi_{\min}, && (27)\\
\sum k &= 1^M|\mathcal{K}_k| \geq (1-\alpha)N, && (28)
\end{aligned}$$

we aim to show that σ indeed represents the one-to-one mapping from cliques to "good contigs". We first show that there is a function $f : \{1,\ldots,M\} \to \{1,\ldots,M,\text{err}\}$ which maps the cliques to at most one species' "good contigs" (or none). Suppose by contradiction that $\mathbf{x},\mathbf{y} \in \mathcal{K}_j$, $\mathbf{x} \in \mathcal{C}_{k,\epsilon}$ and $\mathbf{y} \in \mathcal{C}_{\ell,\epsilon}$, for $k \neq \ell$. Then $d(\mathbf{x},\mathbf{y}) < \epsilon$, and we have

$$\begin{aligned}
d(p_k,p_\ell) &\leq d(p_k,\hat{p}_\mathbf{x}) + d(\hat{p}_\mathbf{x},p_\ell)\\
&\leq d(p_k,\hat{p}_\mathbf{x}) + d\big(\hat{p}_\mathbf{x},\hat{p}_\mathbf{y}\big) + d\big(\hat{p}_\mathbf{y},p_l\big)\\
&< \frac{\epsilon}{2} + \epsilon + \frac{\epsilon}{2} < d_{\min},
\end{aligned}$$

which is a contradiction to the definition of $d_{\min}$. Hence, any $\mathcal{K}_k$ may only contain contigs from one $\mathcal{C}_{\ell,\epsilon}$, and we can define the function f described above.

Second, we show that all species must be covered. Suppose, again by contradiction, that f maps some k to "err" or that there exists j, k such that $f(j) = f(k)$. Then, there should exist some ℓ, such that, for no k, $f(k) = \ell$. Hence, for a sufficiently small α,

$$\begin{aligned}
\left|\bigcup_{k=1}^M \mathcal{K}_k\right| &\leq N - \sum_{\ell:\nexists k, f(k)=\ell} |C_{\ell,\epsilon}|\\
&\leq N - (1-\alpha)N\pi_{\min}\\
&= N(1-\pi_{\min} - \alpha\pi_{\min} && (29)\\
&< N(1-\alpha), && (30)
\end{aligned}$$

with probability $1 - o(1)$ from Lemma 1. Note that the inequality in (30) results from (29 tending toward $N(1 - \pi_{\min})$ and (30) toward N as α decreases. This shows that if at least one species is not covered, then Algorithm 1 cannot satisfy (27) and (28), which is a contradiction.

C. Proof of Lemma 3

Consider a clique $\mathcal{K}_k$ and let ℓ be such that $\mathcal{K}_k \cap \mathcal{C}_{\ell,\epsilon} \neq \emptyset$. By Lemma 1, the fraction of "good" contigs in $\mathcal{K}_k$ will be

$$\begin{aligned}\frac{|\mathcal{K}_k \cap \mathcal{C}_{\ell,\epsilon}|}{|\mathcal{K}_k|} &\geq 1 - \frac{N - M \cdot (1-\alpha)N\pi_k}{(1-\alpha)N\pi_k} \\ &= 1 - \frac{\alpha M}{1-\alpha} \qquad (31)\end{aligned}$$

with probability $1 - o(1)$ from Lemma 1. The lower bound results from dividing the maximum number of contigs *not* in any $\mathcal{C}_{k,\epsilon}$ by the minimum number of contigs in $\mathcal{K}_k$.

D. Proof of Theorem 2

We define the *type* of a contig $\mathbf{x}$ to be its empirical fourth-order distribution, denoted $\hat{p}_{\mathbf{x}}$. Let the set of all possible types of length-L, stationary, third-order Markov sequences be $\mathcal{P}_L$. The cardinality of $\mathcal{P}_L$ is upper-bounded by $(L+1)^4$ as shown in [35]. The *type class*, T_L, of a given type, $p \in \mathcal{P}_L$, is then defined as the set of all length-L sequences whose types are equal to p:

$$T_L(p) = \left\{\mathbf{x} \in \mathcal{X}^L : \hat{p}_{\mathbf{x}} = p\right\}. \qquad (32)$$

To facilitate analysis, we use a *cyclical* Markov model, where three artificial transitions are added from the end of the sequence to the beginning. This model ensures that $\hat{p}_{\mathbf{x}}$ is *consistent*. More precisely, for $\mathbf{a} \in \mathcal{X}^3$,

$$\sum_{b \in \mathcal{X}} \hat{p}_{\mathbf{x}}(\mathbf{a}b) = \sum_{b \in \mathcal{X}} \hat{p}_{\mathbf{x}}(b\mathbf{a}). \qquad (33)$$

Note that this implies $\hat{p}_{\mathbf{x}} \in \widetilde{P}$ as defined in Section II. Furthermore, the third-order and conditional empirical distributions can be derived from $\hat{p}_{\mathbf{x}}$ as follows, for any $b \in \mathcal{X}$,

$$\hat{p}_{\mathbf{x}}(\mathbf{a}) = \sum_{b \in \mathcal{X}} \hat{p}_{\mathbf{x}}(\mathbf{a}b) \qquad (34)$$

and

$$\hat{p}_{\mathbf{x}}(b|\mathbf{a}) = \frac{\hat{p}_{\mathbf{x}}(\mathbf{a}b)}{\hat{p}_{\mathbf{x}}(\mathbf{a})}. \qquad (35)$$

We now use some Large Deviations theory to make an argument about the probability of error in the hypothesis test.

1) Large Deviations Principle: Vidyasagar [35] provides an extensive analysis of large deviations theory for Markov processes. Theorems 3 and 4, shown below, utilize this analysis along with [36, Lemma 1], which allows us to make an argument about the probability of error for the subsequent hypothesis test. For the proofs of Theorems 3 and 4, the reader is referred to [35, Th. 7] and [25, Ch. 11].

The results in [35] show that a Markov process $\mathbf{X} = (X_1, \ldots, X_L)$ with type p and generated by q satisfies the large deviations property with rate function $D(p\|q)$. Here, D is the KL divergence rate defined as the KL divergences averaged over p:

$$\begin{aligned}D(p\|q) &= \sum_{\mathbf{a}\in\mathcal{X}^3} p(\mathbf{a}) \sum_{b\in\mathcal{X}} p(b|\mathbf{a}) \log\left(\frac{p(b|\mathbf{a})}{q(b|\mathbf{a})}\right) \\ &= \sum_{\mathbf{a}\in\mathcal{X}^3, b\in\mathcal{X}} p(\mathbf{a}b) \log \frac{p(\mathbf{a}b)}{q(\mathbf{a}b)} \\ &\quad - \sum_{\mathbf{a}\in\mathcal{X}^3} p(\mathbf{a}) \log \frac{p(\mathbf{a})}{q(\mathbf{a})} \\ &= D^{(4)}(p\|q) - D^{(3)}(p\|q) \qquad (36)\end{aligned}$$

i.e., the divergence between the fourth-order distributions minus the divergence between the third-order distributions. Similarly, the "Markov conditional entropy" can be written as

$$\begin{aligned}H(p) &= \sum_{\mathbf{a}\in\mathcal{X}^3} p(\mathbf{a}) \sum_{b\in\mathcal{X}} p(b|\mathbf{a}) \log p(b|\mathbf{a}) \\ &= H^{(4)}(p) - H^{(3)}(p)\end{aligned}$$

We use D and H without superscript so as to distinguish between the divergence and entropy *rates*.

Theorem 3: The probability of $\mathbf{x}$ under q depends only on its type $\hat{p}_{\mathbf{x}}$ and is given by

$$q^{(L)}(\mathbf{x}) = 2^{-L[D(\hat{p}_{\mathbf{x}}\|q) + H(\hat{p}_{\mathbf{x}})] + \log\alpha} \qquad (37)$$

where $\alpha = q(x_1x_2x_3)$, i.e., the probability of the initial state of $\mathbf{x}$.

Theorem 4 (Sanov's Theorem for Markov Processes): Let $\mathbf{X} = (X_1, X_2, \ldots, X_L)$ be a Markov process q, and let $\mathcal{F} \subseteq \widetilde{\mathcal{P}}$. The probability that the empirical distribution of $\mathbf{X}$ is contained in $\mathcal{F}$, denoted $q^{(L)}(\mathcal{F})$, is upper-bounded as

$$q^{(L)}(\mathcal{F}) \leq |\mathcal{P}_L| 2^{-LD(p^*\|q) + \log\alpha} \qquad (38)$$

where p^* is the information projection of q onto $\mathcal{F}$:

$$p^* = \arg\inf_{p\in\mathcal{F}} D(p\|q). \qquad (39)$$

If, in addition, the closure of $\mathcal{F}$ is equal to the closure of its interior ($\bar{\mathcal{F}} = \overline{\mathcal{F}^o}$), then

$$\lim_{L\to\infty} \frac{1}{L} \log q^{(L)}(\mathcal{F}) = -D(p^*\|q). \qquad (40)$$

2) Hypothesis Test: In the binary hypothesis test, there are two candidate models for $q : p_1$ and p_2, where $p_1 \neq p_2$. We decide between the two hypotheses:

- $\mathrm{H}_1 : q = p_1$
- $\mathrm{H}_2 : q = p_2$

Let $\mathcal{P}_1$ and $\mathcal{P}_2$ be the decision regions for H_1 and H_2, respectively. The sets $\mathcal{P}_1$ and $\mathcal{P}_2$ form a partition of $\widetilde{\mathcal{P}}$ ($\mathcal{P}_1 \cup \mathcal{P}_2 = \widetilde{\mathcal{P}}$). As a result, given any $\mathbf{x} \in \mathcal{X}^L$, $\delta_N(\mathbf{x})$ decides H_1 if $\hat{p}_{\mathbf{x}} \in \mathcal{P}_1$ and H_2 if $\hat{p}_{\mathbf{x}} \in \mathcal{P}_2$. The Bayesian probability of error, P_e, for the binary hypothesis test with priors π_1 and π_2 is given by

$$P_e = \pi_1 p_1^{(L)}(\mathcal{P}_2) + \pi_2 p_2^{(L)}(\mathcal{P}_1). \qquad (41)$$

To minimize the error, the decision rule[4] uses a Neyman-Pearson test

$$\delta_N(\hat{p}_{\mathbf{x}}) = \begin{cases} \mathrm{H}_1 \text{ if } \mathcal{L}(\mathbf{x}) \geq \frac{\pi_2}{\pi_1} \\ \mathrm{H}_2 \text{ if } \mathcal{L}(\mathbf{x}) < \frac{\pi_2}{\pi_1} \end{cases} \qquad (42)$$

[4] The decision rule δ_L uses overloaded notation with the decision rule for the main problem.

where the likelihood ratio, $\mathcal{L}$, is defined as:

$$\mathcal{L}(\mathbf{x}) = \frac{\Pr(\mathbf{x}|\mathrm{H}_1\mathrm{true})}{\Pr(\mathbf{x}|\mathrm{H}_2\mathrm{true})} = \frac{p_1^{(L)}(\mathbf{x})}{p_2^{(L)}(\mathbf{x})}. \tag{43}$$

Using Theorem 3, the normalized log-likelihood ratio is

$$\begin{aligned}\frac{1}{L}\log\mathcal{L}(\mathbf{x}) =& -[D(\hat{p}_{\mathbf{x}}\|p_1) + H(\hat{p}_{\mathbf{x}})] + \frac{\log\alpha_1}{L}\\ &+ [D(\hat{p}_{\mathbf{x}}\|p_2) + H(\hat{p}_{\mathbf{x}})] - \frac{\log\alpha_2}{L}\\ =& \, D(\hat{p}_{\mathbf{x}}\|p_2) - D(\hat{p}_{\mathbf{x}}\|p_1) + \frac{1}{L}\log\frac{\alpha_1}{\alpha_2}\end{aligned}$$

Again, α_1 and α_2 represent the probabilities of the initial states of $\mathbf{x}$ under p_1 and p_2, respectively. Notice that as $L \to \infty$, the optimal decision rule simply chooses $\arg\min_{k\in\{1,2\}} D(p\|p_k)$ because the effect of the priors washes out with L, along with the probability of the initial states. We will show that, by using the decision regions, $\mathcal{P}_1$ and $\mathcal{P}_2$, given by

$$\mathcal{P}_1 = \{p \in \widetilde{\mathcal{P}} : D(p\|p_2) - D(p\|p_1) \geq 0\} \tag{44}$$

$$\mathcal{P}_2 = \{p \in \widetilde{\mathcal{P}} : D(p\|p_2) - D(p\|p_1) < 0\} \tag{45}$$

the optimal error exponent is achieved in the limit. First, we will prove Lemmas 4 and 5, which allow for the use of (40) in Theorem 4.

Lemma 4: $\mathcal{P}_1$ and $\mathcal{P}_2$ are convex.

Proof: Let $p_a, p_b \in \mathcal{P}_1$ and let

$$p_{ab} = \lambda p_a + (1-\lambda)p_b, \quad \lambda \in (0,1)$$

be a convex combination of p_a and p_b. Then

$$\begin{aligned}&D(p_{ab}\|p_2) - D(p_{ab}\|p_1)\\ &\quad= \sum_{\mathbf{a}\in\mathcal{X}^3, b\in\mathcal{X}} p_{ab}(\mathbf{a}b)\log\frac{p_{ab}(b|\mathbf{a})}{p_2(b|\mathbf{a})}\\ &\qquad - \sum_{\mathbf{a}\in\mathcal{X}^3, b\in\mathcal{X}} p_{ab}(\mathbf{a}b)\log\frac{p_{ab}(b|\mathbf{a})}{p_1(b|\mathbf{a})}\\ &\quad= \sum_{\mathbf{a}\in\mathcal{X}^3, b\in\mathcal{X}} p_{ab}(\mathbf{a}b)\log\frac{p_1(b|\mathbf{a})}{p_2(b|\mathbf{a})}\\ &\quad= \lambda[D(p_a\|p_2) - D(p_a\|p_1)]\\ &\qquad + (1-\lambda)[D(p_b\|p_2) - D(p_b\|p_1)] \geq 0\end{aligned}$$

so $p_{ab} \in \mathcal{P}_1$ and therefore, $\mathcal{P}_1$ is a convex set. A similar argument can be made for the set $\mathcal{P}_2$. Note that since $\mathcal{P}_1$ and $\mathcal{P}_2$ are both convex, this implies that the boundary linearly divides the set of stationary fourth-order distributions, $\widetilde{\mathcal{P}}$. ■

Lemma 5:

$$\overline{\mathcal{P}_1} = \overline{\mathcal{P}_1^o} \quad \text{and} \quad \overline{\mathcal{P}_2} = \overline{\mathcal{P}_2^o} \tag{46}$$

Proof: The boundary between $\mathcal{P}_1$ and $\mathcal{P}_2$ consists of the set of distributions $p \in \widetilde{\mathcal{P}}$ for which $D(p\|p_2) - D(p\|p_1) = 0$. We see that p_1 does not lie on this boundary because

$$D(p_1\|p_2) - D(p_1\|p_1) = D(p_1\|p_2) > 0 \tag{47}$$

by the non-negativity of the KL-divergence. Furthermore, p_1 cannot lie on any other boundary of $\mathcal{P}_1$ because all the elements of p_1 are nonzero. Thus, p_1 is an interior point of $\mathcal{P}_1$ as it does not lie on any of the boundaries.

Finally, we need to show that convexity and a non-empty interior imply (46).

Take a point $p \in \overline{\mathcal{P}_1}$. Then either $p \in \mathcal{P}_1^o$ or $p \in \partial\mathcal{P}_1$, the boundary of $\mathcal{P}_1$. If $p \in \mathcal{P}_1^o$, then $p \in \overline{\mathcal{P}_1^o}$, trivially. If $p \in \partial\mathcal{P}_1$, we must prove that p is a limit point of $\mathcal{P}_1^o$. Since p_1 is an interior point of $\mathcal{P}_1$, then there exists an open ball U_1 centered at p_1 which is completely contained in $\mathcal{P}_1$. We define V_1 as the set of distributions that result from a convex combination of p and U_1

$$V_1 = \{\alpha U_1 + (1-\alpha)p : 0 < \alpha \leq 1\} \tag{48}$$

using Minkowski addition. The set V_1 clearly has non-zero volume (by Lebesgue measure) and all of its points are interior points of $\mathcal{P}_1$ due to Lemma 4. Therefore, there exists a sequence of interior points $\{p_t\}, p_t \in V_1$ such that $p_t \to p$. Thus, $p \in \overline{\mathcal{P}_1^o}$.

A similar argument can be made for $\mathcal{P}_2$. Hence, the proof of Lemma 5 is complete. ■

Now, by Theorem 4 and Lemmas 4 and 5, the error exponents are

$$\lim_{L\to\infty}\frac{1}{L}\log p_1^{(L)}(\mathcal{P}_2) = -D(p_1^*\|p_1) \tag{49}$$

$$\lim_{L\to\infty}\frac{1}{L}\log p_2^{(L)}(\mathcal{P}_1) = -D(p_2^*\|p_2). \tag{50}$$

Distribution p_1^* is found by minimizing $D(p_1^*\|p_1)$, subject to the decision boundary constraint,

$$D(p_1^*\|p_1) - D(p_1^*\|p_2) \geq 0, \tag{51}$$

the consistency constraints for all $a \in \mathcal{X}^3$,

$$\sum_{b\in\mathcal{X}} p_1^*(\mathbf{a}b) = \sum_{b\in\mathcal{X}} p_1^*(b\mathbf{a}), \tag{52}$$

and the sum-to-one constraint,

$$\sum_{\mathbf{c}\in\mathcal{X}^4} p_1^*(\mathbf{c}) = 1 \tag{53}$$

This will yield the distribution $p_1^* \in \mathcal{P}_2$ that is closest to p_1. Moreover, we claim that p^* must lie on the boundary, i.e., (51) holds with equality. This can be proven by contradiction: suppose p' is the optimal solution to the minimization problem and suppose

$$D_c\big(p'\|p_1\big) - D_c\big(p'\|p_2\big) > 0.$$

For $0 \leq \lambda \leq 1$, let $p_\lambda = \lambda\, p_1 + (1-\lambda)\, p'$ be a convex combination of p' and p_1. We know from Lemma 4 that $p_\lambda \in \widetilde{\mathcal{P}}$ for any value of λ and furthermore, there exists a $\lambda = \lambda^*$ such that

$$D_c(p_{\lambda^*}\|p_1) - D_c(p_{\lambda^*}\|p_2) = 0$$

since the boundary linearly divides $\widetilde{\mathcal{P}}$. Now, to show by contradiction that

$$D_c(p_{\lambda^*}\|p_1) < D_c\big(p'\|p_1\big),$$

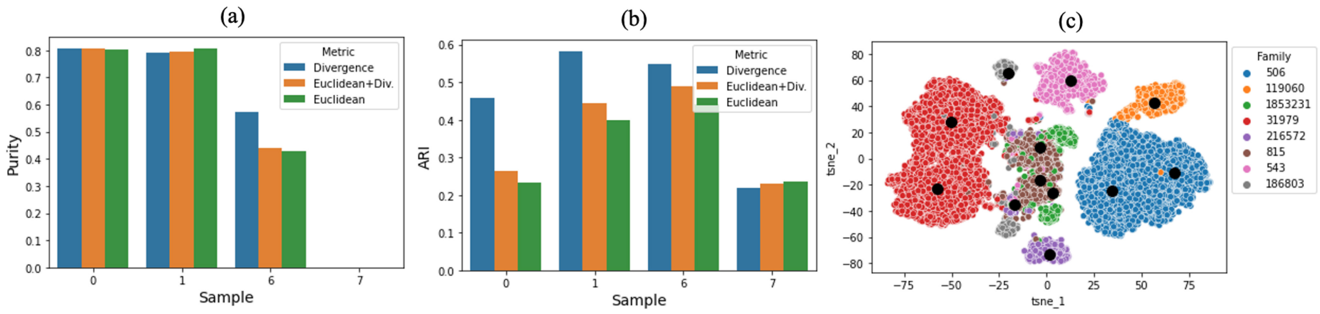

Fig. 3. **(a)** Bar graphs of purity for k-means clustering on several CAMI samples using three different measures for the assignment step, with k set to the number of ground-truth clusters. **(b,c)** Adjusted Rand index and t-SNE plots for $k = 1.5\times$ the number of ground-truth clusters. The taxonomic IDs are given in the legend in (c).

we will show that KL divergence rate is convex in its first argument. For some distribution $q \in \widetilde{\mathcal{P}}$,

$$\begin{aligned}
&D_c(p_\lambda \| q)\\
&\quad= \sum_{\mathbf{a}b\in\mathcal{X}^4} p_\lambda(\mathbf{a}b)\log\frac{p_\lambda(b|\mathbf{a})}{q(b|\mathbf{a})}\\
&\quad= \lambda \sum_{\mathbf{a}b\in\mathcal{X}^4} p_1(\mathbf{a}b)\log\frac{p_\lambda(b|\mathbf{a})}{q(b|\mathbf{a})}\\
&\qquad+ (1-\lambda)\sum_{\mathbf{a}b\in\mathcal{X}^4} p'(\mathbf{a}b)\log\frac{p_\lambda(b|\mathbf{a})}{q(b|\mathbf{a})}\\
&\quad= \lambda \sum_{\mathbf{a}b\in\mathcal{X}^4} p_1(\mathbf{a}b)\left(\log\frac{p_1(b|\mathbf{a})}{q(b|\mathbf{a})} - \log\frac{p_1(b|\mathbf{a})}{p_\lambda(b|\mathbf{a})}\right)\\
&\qquad+ (1-\lambda)\sum_{\mathbf{a}b\in\mathcal{X}^4} p'(\mathbf{a}b)\left(\log\frac{p'(b|\mathbf{a})}{q(b|\mathbf{a})} - \log\frac{p'(b|\mathbf{a})}{p_\lambda(b|\mathbf{a})}\right)\\
&\quad= \lambda D_c(p_1\|q) + (1-\lambda)D_c(p'\|q)\\
&\qquad- \lambda D_c(p_1\|p_\lambda) - (1-\lambda)D_c(p'\|p_\lambda)\\
&\quad< \lambda D_c(p_1\|q) + (1-\lambda)D_c(p'\|q),
\end{aligned}$$

which proves convexity and concludes the proof of Theorem 2.

E. Extended Results

In Section V, we present the results of two sets of experiments. In the second experiment, we ran k-means clustering on several simulated CAMI metagenomic samples. In Figure 3a, the results are shown for the purity metric. We can see that the fraction of pure bins is similar across samples, with KL divergence rate outperforming the other methods in only Sample 6.

Thus far, the results presented using k-means clustering have only set k equal to the number of ground truth clusters (families) in the given sample. As we point out in Section V, in practice, we cannot know the true number of clusters in a sample; thus, evaluating the methods for different numbers of clusters is pertinent. To that end, in Figures 3b,c, we present brief results with k set to 1.5× the number of families. We see that the ARI is generally improved when using KL divergence rate as the binning measure.

Acknowledgment

The idea of modeling the species in a metagenomic sample as distinct Markov processes and utilizing the large deviation principle to characterize their separability came from discussions with Xingyu Jin, Vasilis Ntranos, and David Tse.

References

[1] M. Land et al., "Insights from 20 years of bacterial genome sequencing," *Funct. Integr. Genom.*, vol. 15, no. 2, pp. 141–161, Mar. 2015, doi: 10.1007/ s10142-015-0433-4.

[2] A. S. Motahari, G. Bresler, and D. N. C. Tse, "Information theory of DNA shotgun sequencing," *IEEE Trans. Inf. Theory*, vol. 59, no. 10, pp. 6273–6289, Oct. 2013.

[3] S. Mohajer, A. Motahari, and D. Tse, "Reference-based DNA shotgun sequencing: Information theoretic limits," in *Proc. IEEE Int. Symp. Inf. Theory*, 2013, pp. 1635–1639.

[4] B. Tahmasebi, M. A. Maddah-Ali, and A. S. Motahari, "Genome-wide association studies: Information theoretic limits of reliable learning," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, 2018, pp. 2231–2235.

[5] B. Tahmasebi, M. A. Maddah-Ali, and S. A. Motahari, "Information theory of mixed population genome-wide association studies," in *Proc. IEEE Inf. Theory Workshop (ITW)*, 2018, pp. 1–5.

[6] S. Kannan, J. Hui, K. Mazooji, L. Pachter, and D. Tse, "Shannon: An information-optimal de novo RNA-seq assembler," bioRxiv, 2016. [Online]. Available: https://www.biorxiv.org/content/early/2016/02/09/039230

[7] S. Mao, L. Pachter, D. Tse, and S. Kannan, "RefShannon: A genome-guided transcriptome assembler using sparse flow decomposition," *PLoS ONE*, vol. 15, Jun. 2020, Art. no. e0232946.

[8] F. Breitwieser, J. Lu, and S. Salzberg, "A review of methods and databases for metagenomic classification and assembly," *Brief. Bioinf.*, vol. 20, no. 4, pp. 1125–1136, Jul. 2019.

[9] K. Chen and L. Pachter, "Bioinformatics for whole-genome shotgun sequencing of microbial communities," *PLoS Comput. Biol.*, vol. 1, no. 2, p. e24, Jul. 2005, doi: 10.1371/journal.pcbi.0010024.

[10] E. L. Chatelier et al., "Richness of human gut microbiome correlates with metabolic markers," *Nature*, vol. 500, pp. 541–546, Aug. 2013.

[11] D. D. Kang, J. Froula, R. Egan, and Z. Wang, "MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities," *PeerJ*, vol. 3, p. e1165, Jan. 2015.

[12] J. N. Nissen et al., "Binning microbial genomes using deep learning," bioRxiv, 2018. [Online]. Available: https://www.biorxiv.org/content/early/2018/12/10/490078

[13] P. A. Noble, R. W. Citek, and O. A. Ogunseitan, "Tetranucleotide frequencies in microbial genomes," *Electrophoresis*, vol. 19, no. 4, pp. 528–535, Apr. 1998.

[14] J. Mrázek, "Phylogenetic signals in DNA composition: Limitations and prospects," *Mol. Biol. Evol.*, vol. 26, no. 5, pp. 1163–1169, May 2009, doi: 10.1093/molbev/msp032.

[15] J. Bohlin, E. Skjerve, and D. W. Ussery, "Correction: Investigations of oligonucleotide usage variance within and between prokaryotes," *PLoS Comput. Biol.*, vol. 4, May 2008, Art. no. e1000057.

[16] P. S. Krawczyk, L. Lipinski, and A. Dziembowski, "PlasFlow: Predicting plasmid sequences in metagenomic data using genome signatures," *Nucl. Acids Res.*, vol. 46, no. 6, p. e35, Apr. 2018.

[17] R. A. Edwards, K. McNair, K. Faust, J. Raes, and B. E. Dutilh, "Computational approaches to predict bacteriophage–host relationships," *FEMS Microbiol. Rev.*, vol. 40, no. 2, pp. 258–272, Dec. 2015, doi: 10.1093/femsre/fuv048.

[18] Y. Y. Lu et al., "COCACOLA: Binning metagenomic contigs using sequence composition, read coverage, co-alignment and paired-end read linkage," *Bioinformatics*, vol. 33, no. 6, pp. 791–798, Mar. 2017.

[19] D. E. Wood and S. L. Salzberg, "Kraken: Ultrafast metagenomic sequence classification using exact alignments," *Genome Biol.*, vol. 15, no. 3, p. R46, Mar. 2014, doi: 10.1186/gb-2014-15-3-r46.

[20] Y. Luo, Y. W. Yu, J. Zeng, B. Berger, and J. Peng, "Metagenomic binning through low-density hashing," *Bioinformatics*, vol. 35, no. 2, pp. 219–226, Jul. 2018, doi: 10.1093/bioinformatics/bty611.

[21] R. Ounit, S. Wanamaker, T. J. Close, and S. Lonardi, "CLARK: Fast and accurate classification of metagenomic and genomic sequences using discriminative *k*-mers," *BMC Genom.*, vol. 16, no. 1, p. 236, 2015, doi: 10.1186/s12864-015-1419-2.

[22] L. Schaeffer, H. Pimentel, N. Bray, P. Melsted, and L. Pachter, "Pseudoalignment for metagenomic read assignment," *Bioinformatics*, vol. 33, no. 14, pp. 2082–2088, Feb. 2017, doi: 10.1093/bioinformatics/btx106.

[23] A. Brady and S. Salzberg, "PhymmBL expanded: Confidence scores, custom databases, parallelization and more," *Nat. Methods*, vol. 8, no. 5, pp. 367–367, May 2011. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/21527926

[24] D. H. Huson, S. Mitra, H. J. Ruscheweyh, N. Weber, and S. C. Schuster, "Integrative analysis of environmental sequences using MEGAN4," *Genome Res.*, vol. 21, pp. 1552–1560, Sep. 2011.

[25] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley Series in Telecommunications and Signal Processing). New York, NY, USA: Wiley, 2006.

[26] S. Nurk, D. Meleshko, A. Korobeynikov, and P. A. Pevzner, "MetaSPAdes: A new versatile metagenomic assembler," *Genome Res.*, vol. 27, no. 5, pp. 824–834, May 2017.

[27] P. Moulin and V. V. Veeravalli, *Statistical Inference for Engineers and Data Scientists*. Cambridge, U.K.: Cambridge Univ. Press, 2018.

[28] (Natl. Center Biotechnol. Inf. Co., Bethesda, MA, USA). 1988. [Online]. Available: https://www.ncbi.nlm.nih.gov/

[29] A. Fritz et al., "CAMISIM: Simulating metagenomes and microbial communities," *Microbiome*, vol. 7, no. 1, p. 17, Feb. 2019, doi: 10.1186/s40168-019-0633-6.

[30] F. Meyer et al., "Critical assessment of metagenome interpretation: The second round of challenges," *Nat. Methods*, vol. 19, no. 4, pp. 429–440, Apr. 2022, doi: 10.1038/s41592-022-01431-4.

[31] P. J. Turnbaugh, R. E. Ley, M. Hamady, C. M. Fraser-Liggett, R. Knight, and J. I. Gordon, "The human microbiome project," *Nature*, vol. 449, no. 7164, pp. 804–810, 2007.

[32] H. Akaike, *Information Theory and an Extension of the Maximum Likelihood Principle*. New York, NY, USA: Springer, 1998, pp. 199–213. [Online]. Available: https://doi.org/10.1007/978-1-4612-1694-0_15

[33] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, no. 2, pp. 461–464, 1978, doi: 10.1214/aos/1176344136.

[34] E. Abbe, A. S. Bandeira, and G. Hall, "Exact recovery in the stochastic block model," *IEEE Trans. Inf. Theory*, vol. 62, no. 1, pp. 471–487, Jan. 2016.

[35] M. Vidyasagar, "An elementary derivation of the large deviation rate function for finite state Markov chains," in *Proc. 48th IEEE Conf. Decis. Control (CDC)*, 2009, pp. 1599–1606.

[36] M. Gutman, "Asymptotically optimal classification for multiple tests with empirically observed statistics," *IEEE Trans. Inf. Theory*, vol. 35, no. 2, pp. 401–408, Sep. 2006, doi: 10.1109/18.32134.

Grant Greenberg (Member, IEEE) received the B.S. degree in electrical engineering from the University of Illinois at Urbana–Champaign, where he is currently pursuing the Ph.D. degree in electrical and computer engineering advised by Prof. I. Shomorony. He has published in the areas of genome assembly, sequence alignment, multiomic data integration, and metagenomic binning. His research focuses on the analysis of large-scale omics data using machine learning and information theory.

Ilan Shomorony (Member, IEEE) received the Ph.D. degree in electrical and computer engineering from Cornell University in 2014. He was a Postdoctoral Scholar with UC Berkeley through the NSF Center for Science of Information until 2017. After that, he spent a year working as a Researcher and a Data Scientist with Human Longevity Inc., a personal genomics company. He is currently an Assistant Professor of Electrical and Computer Engineering with the University of Illinois at Urbana–Champaign, where he is a member of the Coordinated Science Laboratory. His research interests include information theory, communications, and computational biology. He received the NSF CAREER Award in 2021.