# The Metagenomic Binning Problem: Clustering Markov Sequences

Grant Greenberg
*University of Illinois at Urbana-Champaign*
gcgreen2@illinois.edu

Ilan Shomorony
*University of Illinois at Urbana-Champaign*
ilans@illinois.edu

***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 want 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 Information between the two most similar species. Our result also implies that contigs should be binned using the conditional relative entropy as a measure of distance, as opposed to the Euclidean distance often used in practice.**

## 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.

Significant attention has recently been given to the analysis of the human microbiome through *metagenomics* [1]. 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. Recent research, including the Human Microbiome Project [2], has shown that the composition of the microbiome is a "snapshot" into 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 [3]. 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 [4]. 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. Specfically, it was empirically verified that the distribution of four-letter strings (e.g. $AGCG$) remains relatively constant across an entire bacterial genome [5]. 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[1], many different algorithms and software packages have been proposed to perform metagenomic binning [3], [6].

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 $p_1, \dots, p_M$ with equal probability, where each $p_k$ corresponds to a certain species. In order 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 $L = \bar{L} \log N$). Our goal is to characterize how large $\bar{L}$ needs to be in order to allow perfect binning with high probability.

To obtain our main result, we establish the equivalent of the Chernoff Information [7, Chapter 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

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

where $C(p_k, p_\ell)$ is the Chernoff Information between $p_k$ and $p_\ell$. 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 $\tilde{p}_k$, $k = 1, ..., M$, of the Markov distributions. Each contig $\mathbf{x}$ is then then placed in bin $k$ given by

$$\arg\min_{k \in \{1,\dots,M\}} D_c\left(\hat{p}_{\mathbf{x}} \| \tilde{p}_k\right)$$

[1]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 paper, we focus on using the TNF alone.

where $\hat{p}_{\mathbf{x}}$ is the empirical 4-symbol distribution of $\mathbf{x}$ and $D_c(\cdot\|\cdot)$ is the conditional relative entropy [7, Definition 2.65].

Our main result suggests that the optimal way to bin metagenomic contigs is to estimate the underlying TNF distribution and then bin contigs using the conditional relative entropy 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 VII.

## II. Problem Statement

As shown in [5], the distribution of tetranucleotides (four-letter strings), tends to be stationary across a fixed bacterial genome. Hence, it is natural to assume that each of the $M$ species in our sample was generated by a distribution $p_i$, $i = 1, ..., M$, over all possible tetranucleotides[2] $\{AAAA, AAAC, ..., TTTT\}$.

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 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) guarantees that any $p \in \widetilde{\mathcal{P}}$ corresponds to the tetranucleotide distribution of a stationary third-order Markov chain. More precisely, for a given $p \in \widetilde{\mathcal{P}}$, we can let $p^{(3)}$ be the induced distribution over 3-letter strings; i.e.,

$$p^{(3)}(\mathbf{a}) = \sum_{b \in \mathcal{X}} p(\mathbf{a}b). \tag{2}$$

This uniquely determines a stationary Markov process with initial state distributed as $p^{(3)}$ and transition probabilities

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

Hence we will model each species in the sample using a distribution $p_i \in \widetilde{\mathcal{P}}$. Notice that the constraint that $p_i(\mathbf{c}) > 0$ for any $\mathbf{c} \in \mathcal{X}^4$ guarantees that the resulting Markov processes are irreducible.

### A. Metagenomic Binning Problem

We assume that we have $M$ species in our sample (for a known $M$). Each species is modeled by a stationary third-order Markov process defined by $p_k \in \widetilde{\mathcal{P}}$, for $k = 1, ..., M$. From this genomic mixture, we observe a set of $N$ realizations $\mathcal{C} = \{\mathbf{x}_i\}_{i=1}^N$, which we call *contigs*. Each $\mathbf{x} \in \mathcal{C}$ is generated independently by first choosing a species $k \in \{1, ..., M\}$ uniformly at random, and then generating a length-$L$ sequence according to $p_k$. We wish to determine which contigs originated from the same genome.

We point out that in real metagenomic experiments, the *coverage depth*, i.e. the expected number of contigs containing a specific nucleotide from one of the $M$ genomes, is low. 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 $\delta_L : \mathcal{X}^L \to \{1, \ldots, M\}$ which correctly maps each contig to its respective genome bin.

For each $k \in \{1, \ldots, M\}$, let $\mathcal{C}_k$ be the set of contigs generated according to $p_k$. 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 $\delta_L$ is

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

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 [8], in which $N \to \infty$ and

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

where $\bar{L}$ is the "normalized contig length". 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. Intuitively, a larger value of $\bar{L}$ should allow one to bin a contig with higher accuracy. 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 $\{\delta_L\}$ such that $Pr(\mathcal{E}_{\delta_L}) \to 0$ as $N \to \infty$.*

### C. Main Result

Interestingly, the fundamental limit of resolvability relies on the Chernoff Information, which we define next.

**Definition 2.** *For two Markov models $p_k$ and $p_\ell$ and some $\lambda \in (0, 1)$, let*

$$p_\lambda(\mathbf{c}) = \frac{p_k^{1-\lambda}(\mathbf{c}) p_\ell^{\lambda}(\mathbf{c})}{\sum_{\mathbf{d} \in \mathcal{X}} p_k^{1-\lambda}(\mathbf{d}) p_\ell^{\lambda}(\mathbf{d})}, \tag{6}$$

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

*for every* $\mathbf{c} \in \mathcal{X}^4$. *The* ***Chernoff Information*** *for* $p_k$ *and* $p_\ell$ *is given by*

$$C(p_k, p_\ell) = D_c\left(p_{\lambda^*} \| p_k\right) = D_c\left(p_{\lambda^*} \| p_\ell\right) \tag{7}$$

*where* $D_c$ *is the conditional relative entropy and* $\lambda = \lambda^*$ *is chosen such that* $D_c\left(p_{\lambda^*} \| p_k\right) = D_c\left(p_{\lambda^*} \| p_\ell\right)$.

The distribution $p_\lambda$ traces a curve between $p_k$ and $p_\ell$ and the Chernoff Information can be thought of as the distance to the midpoint of this curve from either distribution. Our main result establishes that the minimum normalized contig length, $\bar{L}$, depends exlusively on the minimum Chernoff Information between species distributions.

**Theorem 1.** *Let* $C_{\min} \triangleq \min_{k,\ell} C(p_k, p_\ell)$. *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 lengths must be large enough to distinguish between the two closest distributions.

## III. Achievability

The achievability proof of Theorem 1 is first shown in the form of an algorithm so as to highlight the algorithmic nature of metagenomic binning. Given a contig $\mathbf{x} \in \mathcal{C}$, we define the empirical fourth-order distribution of $\mathbf{x}$ as $\hat{p}_{\mathbf{x}}$ and we use $d$ as the $\ell_1$ distance between distributions, i.e. $d(p, q) = \sum_{\mathbf{a} \in \mathcal{X}^4} |p(\mathbf{a}) - q(\mathbf{a})|$.

**Algorithm 1:** Binning Contigs

**Result:** Decision Rule $\delta(\mathbf{x})$
**Input:** Contigs $\mathcal{C}$, Parameter $\alpha \in [0, 1]$
**begin**
  $\mathcal{D} \longleftarrow$ sort ascending $\{d(\hat{p}_{\mathbf{x}}, \hat{p}_{\mathbf{y}}) : \mathbf{x}, \mathbf{y} \in \mathcal{C}\}$
  **for** $\epsilon$ *in* $\mathcal{D}$
    $\mathcal{G}_\epsilon \longleftarrow (V = \mathcal{C}, E_\epsilon = \{(\mathbf{x}, \mathbf{y}) : d(\hat{p}_{\mathbf{x}}, \hat{p}_{\mathbf{y}}) \leq \epsilon\})$
    **if** $\mathcal{G}_\epsilon$ has cliques $\{\mathcal{K}_k\}_{k=1}^M, |\mathcal{K}_k| \geq (1-\alpha)\frac{N}{M}$
      **for** $k \longleftarrow 1$ to $M$
        $\widetilde{p}_k \longleftarrow \frac{1}{|\mathcal{K}_k|} \sum_{\mathbf{x} \in \mathcal{K}_k} \hat{p}_{\mathbf{x}}$
      **break**
  **for** $\mathbf{x} \in \mathcal{C}$
    $\delta(\mathbf{x}) \longleftarrow \arg\min_{k \in \{1,\ldots,M\}} D_c\left(\hat{p}_{\mathbf{x}} \| \tilde{p}_k\right)$

The algorithm first estimates the species distributions by averaging the empirical distributions of the contigs in each clique, then 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 Theorem 1. The proof of Theorem 2 is provided in the longer version of this paper [12].

### 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

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

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 will first show that once the algorithm identifies cliques $\mathcal{K}_1, \ldots, \mathcal{K}_M$, each clique is "pure" in the sense that it contains "good" contigs from only one species. We will let $d_{\min} \triangleq \min_{k \neq \ell} d(p_k, p_\ell)$ be the minimum $\ell_1$ distance between any pair of the $M$ species distributions.

**Lemma 1.** *If* $\mathcal{K}_k$ *is a clique in* $\mathcal{G}_\epsilon$ *for* $\epsilon < \frac{d_{\min}}{2}$, *then*

$$\sum_{\ell=1}^{M} \mathbb{1}\{\mathcal{K}_k \cap \mathcal{C}_{\ell,\epsilon/2} \neq \emptyset\} \leq 1, \tag{9}$$

Lemma 1 establishes that, if Algorithm 1 finds $M$ cliques of size $(1 - \alpha)\frac{N}{M}$ in $\mathcal{G}_\epsilon$ for $\epsilon < \frac{d_{\min}}{2}$, then each clique contains contigs from at most one $\mathcal{C}_{\ell,\epsilon/2},\ \ell = 1, \ldots, M$. In order to establish that these $M$ cliques will exist in $\mathcal{G}_\epsilon$, we use the following lemma, which essentially says that, for $N$ large enough, a large fraction of the contigs will be close to their respective generating distributions.

**Lemma 2.** *For* $\epsilon > 0$, $k \in \{1, \ldots, M\}$, *and* $N$ *large enough,*

$$\Pr\left(|\mathcal{C}_{k,\epsilon/2}| < (1-\alpha)\frac{N}{M}\right) \leq \frac{1}{\alpha} e^{-\gamma \alpha^2 L}, \tag{10}$$

*where* $\gamma$ *is a positive constant.*

Fixing $\epsilon < \frac{d_{\min}}{2}$, Lemma 2 guarantees that for a reasonably chosen $\alpha$, as long as $N$ is large enough, we will have $|\mathcal{C}_{k,\epsilon/2}| \geq (1-\alpha)\frac{N}{M}$ for all $k = 1, \ldots, M$. Moreover, by the triangle inequality, any two contigs $\mathbf{x},\ \mathbf{y} \in \mathcal{C}_{k,\epsilon/2}$ will be at a distance $\epsilon$ or less of each other and will thus have an edge between them. Hence, $\mathcal{C}_{k,\epsilon/2}$ forms a clique in $\mathcal{G}_\epsilon$.

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 $\epsilon$ in increasing order, for some $\epsilon < \frac{d_{\min}}{2}$, $M$ cliques of size $(1-\alpha)\frac{N}{M}$ will exist with probability $1 - o(1)$. Lemma 1 will then guarantee that any cliques $\mathcal{K}_1, \ldots, \mathcal{K}_M$ that are found will be pure.

Consider a clique $\mathcal{K}_k$ and let $\ell$ be such that $\mathcal{K}_k \cap \mathcal{C}_{\ell,\epsilon/2} \neq \emptyset$. By Lemma 2, the fraction of "good" contigs in $\mathcal{K}_k$ will be

$$\frac{|\mathcal{K}_k \cap \mathcal{C}_{\ell,\epsilon/2}|}{|\mathcal{K}_k|} \geq 1 - \frac{N - M \cdot (1-\alpha)\frac{N}{M}}{(1-\alpha)\frac{N}{M}} = 1 - \frac{\alpha M}{1-\alpha} \tag{11}$$

with probability $1 - o(1)$. If we set $\alpha = \frac{1}{\log N}$, (11) converges to 1 and (10) 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 $\mathcal{C}_{\ell,\epsilon/2}$. Since distribution vectors are bounded, their impact on $\tilde{p}_k$ also vanishes, and we conclude that the distribution estimate $\tilde{p}_k = \frac{1}{|\mathcal{K}_k|} \sum_{\mathbf{x} \in \mathcal{K}_k} \hat{p}_{\mathbf{x}} \to p_\ell$ 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 $N \to \infty$. Thus, in the remainder of the achievability proof, we will assume that we have $\{p_k\}_{k=1}^M$. A continuity argument can then be used to establish the result when instead we have arbitrarily accurate estimates $\{\tilde{p}_k\}_{k=1}^M$.

Next we show that, binning the contigs based on the conditional relative entropy achieves (8). Consider the hypothesis test between two Markov processes $p_k$ and $p_\ell$. Given prior probabilities $\pi_k$ and $\pi_\ell$, 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 $p_\ell$.

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

$$\lim_{L\to\infty} \frac{1}{L} \log \Pr(\mathcal{E}_{k,\ell}) = -C(p_k, p_\ell), \tag{12}$$

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

The proof of Theorem 2 is given in Section VI. We point out that a similar characterization of this error exponent was recently provided by Moulin and Veeravalli [9, Chapter 10] (under the name of Chernoff divergence).

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 by considering the two closest distributions. Then

$$\begin{aligned}
\Pr(\mathcal{E}_{\delta_L}) &\le \sum_{k=1}^{M} \sum_{\mathbf{x}\in\mathcal{C}_k} \sum_{\ell\neq k} \Pr(\mathcal{E}_{k,\ell}) && (13)\\
&\le (M-1)N \max_{k\neq\ell} \Pr(\mathcal{E}_{k,\ell}) && (14)\\
&\le M 2^{L\left(1/\bar{L} + \max_{k\neq\ell}(1/L)\log\Pr(\mathcal{E}_{k,\ell})\right)} && (15)
\end{aligned}$$

where (13) follows from the union bound. By Theorem 2, $\max_{k\neq\ell} \frac{1}{L}\log\Pr(\mathcal{E}_{k,\ell}) \to -\min_{k\neq\ell} C(p_k,p_\ell) = -C_{\min}$ as $N\to\infty$. Hence, if $\bar{L} > \frac{1}{C_{\min}}$, $\Pr(\mathcal{E}_{\delta_L}) \to 0$. 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 $\delta_L$ and a contig $\mathbf{x}\in\mathcal{C}$, let

$$\widetilde{\mathcal{E}}_{1,2,\mathbf{x}} = \{\mathbf{x}\in\mathcal{C}_1, \delta_L(\mathbf{x})\neq 1\} \cup \{\mathbf{x}\in\mathcal{C}_2, \delta_L(\mathbf{x})\neq 2\}$$

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

$$\begin{aligned}
\Pr(\mathcal{E}_{\delta_L}) &\ge \Pr\left(\bigcup_{i=1}^{N} \widetilde{\mathcal{E}}_{1,2,\mathbf{x}_i}\right) = 1 - (1-\Pr(\widetilde{\mathcal{E}}_{1,2,\mathbf{x}_1}))^N\\
&\ge 1 - \left[\left(1-\frac{2}{M}\Pr(\mathcal{E}_{1,2})\right)^{1/\Pr(\mathcal{E}_{1,2})}\right]^{N\Pr(\mathcal{E}_{1,2})}\\
&\ge 1 - e^{-\frac{2}{M}N\Pr(\mathcal{E}_{1,2})} && (16)
\end{aligned}$$

where (16) follows from the bound $(1-ap)^{1/p} \le e^{-a}$ for $p\in[0,1], a\in\mathbb{R}$. We see that, if $\frac{2}{M}N\Pr(\mathcal{E}_{1,2}) \not\to 0$, then $\Pr(\mathcal{E}) \not\to 0$. Since

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

then by Theorem 2, $\frac{2}{M}N\Pr(\mathcal{E}_{1,2}) \not\to 0$ when $\bar{L} \le \frac{1}{C_{\min}}$. 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 conditional relative entropy is a good metric for binning contigs,
2) the Chernoff information 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. To this end, we utilized several previously sequenced and assembled bacterial genomes, available at NCBI [10]. For each bacterial species, we computed its fourth-order distribution $p_k$ (i.e., the overall tetranucleotide frequency vector), and were able to simulate contigs of a desired length $L$ by randomly sampling length-$L$ substrings from the genome.

In order to verify the usefulness of the conditional relative entropy and compare it to the Euclidean distance (used in state-of-the-art tools such as [3], [6]), we considered the following experiment: we generated 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 conditional relative entropy. In Figure 1a, the conditional relative entropy metric[3] outperforms the Euclidean metric in the test between the species Alistipes Obesi and Megamonas Funiformi, as was the case for most pairs of species. We note, however, that in a select few pairs, the Euclidean metric gives the lower error.

Theorem 1 implies that the inverse of the Chernoff Information characterizes how long the contigs need to be in order for two species to be reliably distinguishable. In order to verify that, we considered different pairs of species $p_k$, $p_\ell$. For each such pair, we computed the Chernoff Information $C(p_k,p_\ell)$ and $\bar{L}_{5\%}$, the minimum normalized contig length $\bar{L}$ required to guarantee a $5\%$ error rate in a Bayesian hypothesis test between $p_k$ and $p_\ell$ with equal priors. In Figure 1 (right), 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.

## VI. Proofs

### A. Proof of Lemma 1

Suppose by contradiction that $\mathbf{x},\mathbf{y}\in K_j$, $\mathbf{x}\in\mathcal{C}_{k,\epsilon/2}$ and $\mathbf{y}\in\mathcal{C}_{\ell,\epsilon/2}$, for $k\neq\ell$. Then $d(\mathbf{x},\mathbf{y})<\epsilon$, and we have

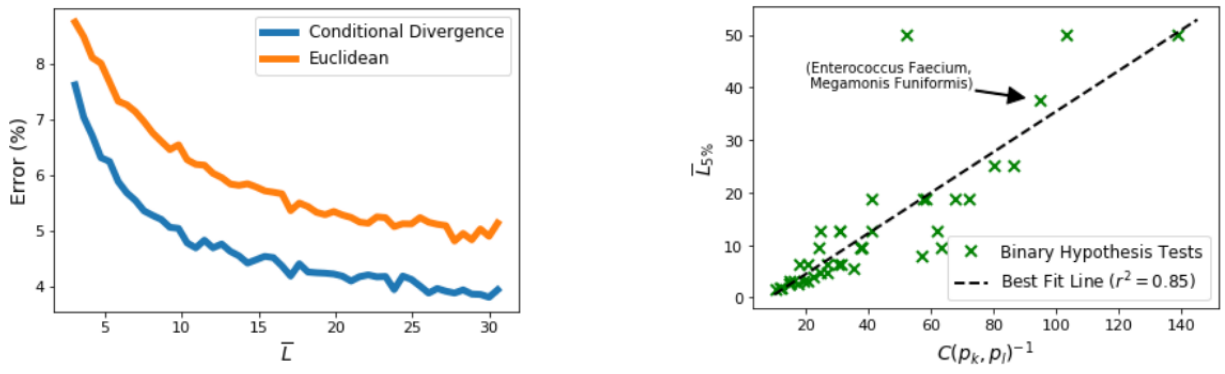
$$\begin{aligned}
d(p_k,p_\ell) &\le d(p_k,\hat{p}_{\mathbf{x}}) + d(\hat{p}_{\mathbf{x}},p_\ell)\\
&\le d(p_k,\hat{p}_{\mathbf{x}}) + d(\hat{p}_{\mathbf{x}},\hat{p}_{\mathbf{y}}) + d(\hat{p}_{\mathbf{y}},p_l)\\
&< \epsilon/2 + \epsilon + \epsilon/2 < d_{\min},
\end{aligned}$$

which is a contradiction to the definition of $d_{\min}$.

### B. Proof of Lemma 2

Let $\mathcal{E}_k$ be the event of interest, $\left\{|\mathcal{C}_{k,\epsilon/2}| < (1-\alpha)\frac{N}{M}\right\}$, and let $\mathcal{A}_k = \left\{|C_k| < (1-\frac{\alpha}{2})\frac{N}{M}\right\}$. Note that we use $\frac{\alpha}{2}$ for $\mathcal{A}_k$ as opposed to $\alpha$ because we need $|C_k|$ to be larger than $|C_{k,\epsilon/2}|$. By Hoeffding's inequality, $\Pr(\mathcal{A}_k) \le e^{-N\frac{\alpha^2}{2M^2}}$. This means that, with high probability, $p_k$ will generate enough contigs.

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




**(a)** **(b)**

**Fig. 1:** (a) Comparison of conditional divergence and Euclidean distance for a hypothesis test between Alistipes Obesi and Megamonas Funiformis; (b) Normalized contig length needed for 5% error ($\overline{L}_{5\%}$) vs the inverse of the Chernoff information for several pairs of species.

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

$$\mathcal{F}_k = \left\{ p \in \widetilde{\mathcal{P}} : d\left(p_k, p\right) \geq \frac{\epsilon}{2} \right\}. \tag{17}$$

By a version of Sanov's theorem for Markov chains, proven by Vidyasagar [11], [12], for any $\mathbf{x} \in \mathcal{C}_k$,

$$\Pr(\hat{p}_{\mathbf{x}} \in \mathcal{F}_k) \leq (L+1)^4 2^{-L D_c(p^* \| p_k)} \tag{18}$$

where $p^* = \arg\inf_{p \in \mathcal{F}_k} D_c\left(p \| p_k\right)$; i.e., $p^*$ is the distribution in $\mathcal{F}_k$ closest to $p_k$ in conditional relative entropy. Notice that $\mathcal{E}_k$ occurs when more than $|C_k| - (1-\alpha)\frac{N}{M} + 1$ 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 $p_k$,

$$\Pr(\mathcal{E}_k | \mathcal{A}_k^c) \tag{19}$$

$$= \Pr\left( \sum_{\mathbf{x} \in \mathcal{C}_k} \mathbb{1}\{\hat{p}_{\mathbf{x}} \in \mathcal{F}_k\} \geq |\mathcal{C}_k| - (1-\alpha)\frac{N}{M} + 1 \middle| \mathcal{A}_k^c \right) \tag{20}$$

$$\leq \sum_{i=1}^{N} \Pr\left( \mathbb{1}\{\hat{p}_{\mathbf{x}_0} \in \mathcal{F}_k\} \geq \frac{\alpha}{2}\frac{N}{M} \middle| \mathcal{A}_k^c \right) \tag{21}$$

$$\leq \frac{2M}{\alpha} \cdot \Pr(\hat{p}_{\mathbf{x}_0} \in \mathcal{F}_k | \mathcal{A}_k^c) \tag{22}$$

where (21) follows from symmetry between contigs and (22) from Markov's inequality. Combining the probabilities,

$$\Pr(\mathcal{E}_k) = \Pr(\mathcal{E}_k | \mathcal{A}_k^c)\Pr(\mathcal{A}_k^c) + \Pr(\mathcal{E}_k | \mathcal{A}_k)\Pr(\mathcal{A}_k) \tag{23}$$

$$\leq \frac{2M}{\alpha} \cdot \Pr(\hat{p}_{\mathbf{x}_0} \in \mathcal{F}_k | \mathcal{A}_k^c)\Pr(\mathcal{A}_k^c) + \Pr(\mathcal{A}_k) \tag{24}$$

$$\leq \frac{2M}{\alpha}(L+1)^4 2^{-L D_c(p^* \| p_k)} + e^{-N\frac{\alpha^2}{2M^2}} \tag{25}$$

$$\leq \frac{1}{\alpha} e^{-\gamma \alpha^2 L} \tag{26}$$

where $\gamma > 0$ is some constant (guaranteed to exist for $\alpha \in [0,1]$) such that (26) holds for large $N$.

## VII. Concluding Remarks

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 Information 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 conditional relative entropy between a contig and an estimate of a species TNF. Through simple experiments, we provided preliminary evidence that this metric often outperforms the Euclidean metric. However, it is unclear whether for real genomes one can reliably estimate the overall TNF of a genome. Directions for future work include verifying whether it is possible to achieve Theorem 1 without estimating the underlying TNF distributions, and finding a computationally efficient version of Algorithm 1.

## References


[1] F. Breitwieser, et. al. "A review of methods and databases for metagenomic classification and assembly," *Briefings in bioinformatics*, 2017.

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

[3] 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, 2015.

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

[5] Noble, P. A., et al. "Tetranucleotide frequencies in microbial genomes," *Electrophoresis*, vol. 19, no. 4, pp. 528–535, Apr 1998.

[6] Lu, Yang Young, 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, 2017.

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

[8] A. S. Motahari, G. Bresler, and D. Tse, "Information theory of DNA sequencing," http://arxiv.org/abs/1203.6233

[9] P. Moulin and V. V. Veeravalli, *Statistical Inference for Engineers and Data Scientists*. Cambridge University Press, 2018.

[10] National center for biotechnology information. Available: www.ncbi.nlm.nih.gov/

[11] M. Vidyasagar, "An elementary derivation of the large deviation rate function for finite state markov chains," in *Proceedings of the 48th IEEE Conference on Decision and Control (CDC)*, Dec 2009, pp. 1599–1606.

[12] G. Greenberg and I. Shomorony, "The metagenomic binning problem: Clustering markov sequences," *ilanshomorony.com/papers/itw2019.pdf*.