Tel Aviv University
Raymond and Beverly Sackler Faculty of Exact Sciences

On k-mer Distances and Synteny Blocks

A thesis presented for the degree of
Master of Science

The Blavatnik School of Computer Science

Alon Kafri

Prepared under the supervision of
Prof. Benny Chor and Prof. David Horn

2019

Abstract

Chromosomes and whole genomes go through a number of changes as they evolve over time. These include large scale changes, such as inversions, translocations, deletions and many more. There are also very low level changes e.g point mutations. Our understanding of genome and chromsome level evolution is still rather limited. In this paper, we study two approaches for understanding the relations between genomes and chromosomes and compare the two. One is the ratio of synteny blocks sizes among two genomes; the other is the so called *k-mers* distance.

We apply these methods to families of E. Coli species and Salmonella species to understand the correlation between syntenic regions and *k-mer* distances and discuss the advantage of using such methods over traditional ones - mostly time complexity. In addition, we present two interesting methods that can be used with *k-mer* distances. One is identifying chromosomes that are evolutionary close to each other by calculating pairs of chromosomes that have very low distance between them. The second is using *k-mer* distances to reconstruct phylogenetic trees.

Acknowledgements

I would first like to thank my supervisors, Prof. Benny Chor and Prof. David Horn for their support, optimistic spirit and endless patience. They both inspired me with their ambition and knowledge for this research, together with their non compromising attitude for precision. Without them I could never have finished this research.

I want to thank Prof. Uri Gophna for his contribution to this research.

Lastly, to my family and friends who kept believing and supporting me through out this period.

Contents

Chapter 1

Introduction

1.1 Motivation and Background

The phenomenon of Inversion Symmetry (IS) has recently been reevaluated and established in [20]. This generalization of the 2nd Chargaff rule [19] implies that the number of occurrences of any sequence S of length k on a chromosomal strand is equal to the number of occurrences of its inverse (reverse-complement) sequence S_{inv} on the same strand. Another way of stating the same fact is that the number of occurrences of S on one chromosomal strand is equal to the number of occurrences of S on the other strand provided both are being read along their own 5' to 3' directions. It has been shown in [20] that this rule holds, within statistical significance, for large k up to the k-limit KL which grows logarithmically with chromosomal length L. Furthermore, KL values differ for different organisms due to the difference in their length and the difference in their strand structure. Eukaryotes genome is linear, meaning it has a beginning and an end unlike Bacteria genome strand which is circular.

Here we create a measure of distance between chromosomes, within the same organism or between different species. This is carried out by comparing frequencies of all strings of the same length k on different chromosomes, summing over both strands of each chromosome.

1.2 Existing Approaches

There are several approaches to finding similarities between two given strings that consist of genomic data.

1.2.1 String Similarity and Blast Distance

This can be done by measuring the edit distance between two strings, or finding the longest common substring between the sequences. The problem with such a method is that the time complexity of these algorithms is very bad, especially if we are trying to match long sequences such as genomic sequences.

1.2.2 Statistical Similarity

This can be done by applying *NLP* techniques such as *n-grams* or *word2vec*. These will provide statistical difference between the given sequences. This can be done by learning the similarity between well known sequences. However, there are a lot of possible sequences that can occur in real life sequences and training such data can be a difficult task.

1.2.3 Biological Approach

This can be done by identifying similar biological sequences and applying this knowledge on the given strings. The main problem with such an approach is that sometimes biological knowledge is missing, making it impossible to take it into account in the computations.

1.3 Genomic Data

Genomic sequences may contain regions that are considered to have no significant information. There are few examples of such regions. One example is a repeat of CG or AT. Another example is low-complexity regions which are sub sequences of biased composition. Ensembl reference genomes contain three types: masked, unmasked and soft-masked. The unmasked version contains all repeats and low complexity regions; Soft-masked replaces all repeats and low-complexity with lowercase nucleutides to help identify these regions; masked replaces all repeats and low-complexity with N.

Usage of each of these types can yield different results. The masked version may ignore relevant data of the genome whereas the unmasked version may add noise to the data set. It is recommended to use unmasked data when working with alignments and therefore in our research we used unmasked versions (except in parts where we explicitly mention otherwise).

Chapter 2

Our Approach

We will focus on measuring k-mer distance between two species and using it to asses the existence of reverse complement synteny blocks. Our approach consists of three steps

- Calculating k-mer distance
- Calculating Synteny blocks
- Comparing k-mer distances with Synteny blocks

2.1 k-mers Distance

The term *k-mer* refers to all the possible substrings of length k that are contained in a given string. Given a string of length l over Σ, the number of k-mers in it is $l-k+1$. We define the empirical frequency of a specific k-mer in the string S as the number of occurrences of the k-mer in S divided by $l-k+1$. We will use F_S to denote this frequency. Let us order the 4^k k-mers in some canonical lexicographic order. For a given S, we consider the 4^k dimensional vector, where the i-th entry equals the empirical frequency $f_S(a_i)$, where a_i is the i-th k-mer in the canonical ordering. Such vectors of frequencies enable us to define a distance between k-mers of different strings over Σ, even if their lengths are not the same. Let us distinguish between distances defined on probabilities measured on a single strand of both sequences and distances obtained from both strands.

Definition 2.1.1. Given a sequence S, let $F_k(S)$ define the k-mer frequency over the positive strand as

$$F_k(S) = (\frac{f_S(a_1)}{l-k+1}, \frac{f_S(a_2)}{l-k+1}, \cdots, \frac{f_S(a_{4^k})}{l-k+1}) \tag{2.1}$$

Similarly, the frequency vector of the inverse sequence is $F_k(\overline{S})$.

Definition 2.1.2. k-mer Distance
Given two genomic sequences S_1, S_2. The k-mer distance between S_1 and S_2 is defined as the l_1 norm between the frequency vectors.

$$D_1^k(S_1, S_2) = \mathbb{E}\,|F_k(S_1) - F_k(S_2)| = \Sigma_{i=1}^{4^k} |\frac{f_{S_1}(a_i)}{l_1-k+1} - \frac{f_{S_2}(a_i)}{l_2-k+1}| \tag{2.2}$$

In the definition above we do not treat reverse complement k-mers - that is each k-mer is on its own. Similar to the k-mer distance definition we will define the reverse complement k-mer distance where each k-mer is counted with its reverse complement k-mer. That is, concatenating the reversed complement sequence to the original sequence and divide the counts by 2.

Definition 2.1.3. Reverse Complement k-mer Distance
Given two genomic sequences S_1, S_2. The k-mer reversed complement distance between S_1 and S_2 is defined as the l_1 norm between the probability vectors.

$$D_2^k(S_1, S_2) = \frac{1}{2} D_1^k(S_1 \cdot \overline{S_1}, S_2 \cdot \overline{S_2}) = \frac{1}{2} \mathbb{E} \left| F_k(S_1 \cdot \overline{S_1}) - F_k(S_2 \cdot \overline{S_2}) \right| \tag{2.3}$$

Where $S_1 \cdot \overline{S_1}$ is S_1 concatenated by its inverse sequence $\overline{S_1}$.

Let us define the *k-mer distance ratio* as:

$$\Delta_{k-mer}^k(S_1, S_2) = \frac{D_2^k(S_1, S_2)}{D_1^k(S_1, S_2)}$$

Note that the definition above is valid only if $D_1^k(S_1, S_2) \neq 0$.

2.1.1 The validity of the k-mer Distribution

The paper [20] suggests that given two sequences S_1, S_2 of length l_1, l_2 respectively, the K-Limit of these sequences depends on the $\log_k(\min\{l_1, l_2\})$, that is the distances are well defined if $4^k << \min\{l_1, l_2\}$. Otherwise, $\frac{4^k}{\min\{l_1,l_2\}} \to 0$ and therefore changing even half of one of the sequences value will not yield a sufficient difference in the distance.

2.1.2 Calculating k-mer Distance

There are few approaches when it comes to calculating k-mer distances, the benefits of each depends on the value of k. The number of k-mers grows exponential with the increase of k; therefore if the value of k is small its better to use traditional string algorithms. However, if k is big, its better to use distributed map-reduce approaches. In our k-mer counting calculations we used the Jellyfish k-mer counter presented by [13].

2.1.3 Time and Space Complexity of k-mer distance calculation

Let S_1, S_2 be sequences of length l_1, l_2 respectively and analyze the complexity of calculating $D_1^k(S_1, S_2)$ and $D_2^k(S_1, S_2)$ distances. In order to calculate the k-mer distribution for each of the sequences we must go through all $l - k + 1$ k-mers. Using a hash map to save the counts we can assume that each increment is an $\mathcal{O}(1)$. Hence, calculating the k-mer distribution of S_1 will cost us $\mathcal{O}(l_1 - k + 1)$ and $\mathcal{O}(l_2 - k + 1)$ for S_2. In order to calculate the k-mer distance between the two sequences, we must compare their distribution vectors of size 4^k each. Therefore the total time complexity is

$$\mathcal{O}(l_1 + l_2 - 2k + 2 + 2 \cdot 4^k) \tag{2.4}$$

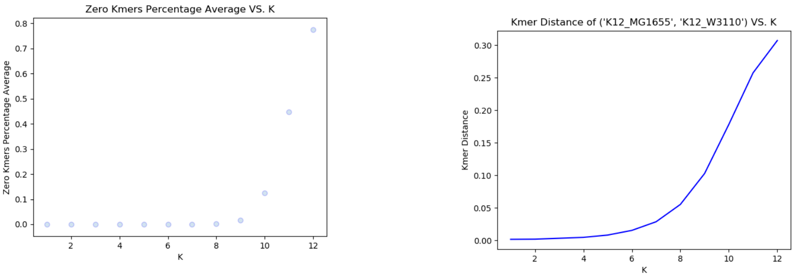
The space complexity of this algorithm consists of saving a count map for each of the k-mers for both sequences. This yields the space complexity of

$$\mathcal{O}(2 \cdot 4^k) \tag{2.5}$$

Note that the space complexity does not depend on l_1 or l_2 which is a strong property of this algorithm. However, the time complexity depends both on l_1, l_2 and on k. In 2.1.1, we require that $4^k << \min\{l_1, l_2\}$ and therefore we can assume that the time complexity is $\mathcal{O}(l_1 + l_2)$ and the space complexity is negligible for lower k values. Note that this algorithm can be easily converted to a distributed version and reduce the time complexity even more.

2.1.4 k-mer Distance Values

Figure 2.1: K Value Implications



(a) Unrealized k-mers Percentage on E. Coli

(b) k-mer Distance VS. K

KL value for E. Coli is 7

First, note that the number of unrealized k-mers and the k-mer distance are monotonically increasing with the value of k as seen in 2.1 and 5. The increase of the number of zero k-mers may imply that the k-mer distance between two sequences will decrease, however as shown in 2.1, it is the opposite. This can be explained by the fact that the actual k-mers that do appear in each of the sequences are different, making the counts vectors to differ. Furthermore, the k value controls the complexity of the calculation. Larger k values tend to increase the calculation time exponentially. The choice of k is therefore an important part of evaluating the correlation. In [20] the authors state that the k-limit (the highest k for which the inversion symmetry holds) is $0.7\ln(L)$ where L is the length of the sequence. Specifically the k-limit of bacteria is around 7. This indicates that the choice of k should be the k-limit. That is for species pair (S_1, S_2), the value of k should be $k = KL(\min\{L(S_1), L(S_2)\})$. In our case, $k = 7$.

2.1.5 k-mer Distance Properties

We will focus on the important properties of both the k-mer distance and the reverse complement k-mer distance.

Theorem 1. *For any two genomic sequences S_1, S_2*

- $D_1^k(S_1, S_2) > 0$

- $D_1^k(S_1, S_2) = 0$ *if and only if* $F_k(S_1) \sim F_k(S_2)$

Proof. By the definition of $D_1^k(S_1, S_2)$ it is easy to conclude that it is greater or equal to 0. In order to prove the 0 case, assume by contradiction that $D_1^k(S_1, S_2) = 0$ but $F_k(S_1) \nsim F_k(S_2)$. Since the two sequences are not equal, there must be at least one k-mer count that is different between the two count vectors. Assume without the loss of generality that it is the i-th k-mer. Therefore, $|\frac{f_{S_1}(a_i)}{l_1-k+1} - \frac{f_{S_2}(a_i)}{l_2-k+1}| > 0$ which follows that

$$D_1^k(S_1, S_2) = \Sigma_{i=1}^{4^k} |\frac{f_{S_1}(a_i)}{l_1 - k + 1} - \frac{f_{S_2}(a_i)}{l_2 - k + 1}| > 0$$

which is a contradiction. The other direction can be shown with the same method. □

Theorem 2. *For any two genomic sequences S_1, S_2*

- $D_2^k(S_1, S_2) > 0$
- $D_2^k(S_1, S_2) = 0$ *if and only if* $F_k(S_1 \cdot \overline{S_1}) \sim F_k(S_2 \cdot \overline{S_2})$

Proof. The proof is based on theorem (1). Mark $S_1 \cdot \overline{S_1}$ as T_1 and $S_2 \cdot \overline{S_2}$ as T_2. Therefore,

$$D_2^k(S_1, S_2) = \frac{1}{2} D_1^k(S_1 \cdot \overline{S_1}, S_2 \cdot \overline{S_2}) = \frac{1}{2} D_1^k(T_1, T_2)$$

which proves the theorem. □

Another interesting fact about k-mer distances is that D_2 is always smaller or equal than D_1.

Theorem 3. *For any two genomic sequences S_1, S_2*

$$D_2^k(S_1, S_2) \leq D_1^k(S_1, S_2) \tag{2.6}$$

Proof. Let S_1, S_2 be two sequences of length l_1, l_2 respectively. Now looking into the frequency vectors of the k-mers we can see that

$$f_{S_1 \cdot \overline{S_1}}(a_i) = f_{S_1}(a_i) + f_{\overline{S_1}}(a_i) \tag{2.7}$$

By the definition of D_2 we can deduce that

$$D_2^k(S_1, S_2) = \frac{1}{2} \mathbb{E} \left| F_k(S_1 \cdot \overline{S_1}) - F_k(S_2 \cdot \overline{S_2}) \right| = \frac{1}{2} \Sigma_{i=1}^{4^k} |\frac{f_{S_1 \cdot \overline{S_1}}(a_i)}{l_1 - k + 1} - \frac{f_{S_2 \cdot \overline{S_2}}(a_i)}{l_2 - k + 1}| \tag{2.8}$$

Now using both (1) and (2) we can deduce that

$$\begin{aligned} D_2^k(S_1, S_2) = \frac{1}{2} \Sigma_{i=1}^{4^k} |\frac{f_{S_1}(a_i)}{l_1 - k + 1} + \frac{f_{\overline{S_1}(a_i)}}{l_1 - k + 1} - \frac{f_{S_2}(a_i)}{l_2 - k + 1} - \frac{f_{\overline{S_2}(a_i)}}{l_2 - k + 1}| \leq \\ \frac{1}{2} \Sigma_{i=1}^{4^k} |\frac{f_{S_1}(a_i)}{l_1 - k + 1} - \frac{f_{S_2}(a_i)}{l_2 - k + 1}| + |\frac{f_{\overline{S_1}(a_i)}}{l_1 - k + 1} - \frac{f_{\overline{S_2}(a_i)}}{l_2 - k + 1}| = \\ \frac{1}{2} \Sigma_{i=1}^{4^k} |\frac{f_{S_1}(a_i)}{l_1 - k + 1} - \frac{f_{S_2}(a_i)}{l_2 - k + 1}| + \frac{1}{2} \Sigma_{i=1}^{4^k} |\frac{f_{\overline{S_1}(a_i)}}{l_1 - k + 1} - \frac{f_{\overline{S_2}(a_i)}}{l_2 - k + 1}| = \\ \frac{1}{2} D_1^k(S_1, S_2) + \frac{1}{2} D_1^k(\overline{S_1}, \overline{S_2}) = \\ D_1^k(S_1, S_2) \end{aligned} \tag{2.9}$$

Where the first step is due to the triangle inequality and the last step is due to the fact that summing all k-mers leads to the same result as summing over all inverse k-mers, amounting to a different order of counting.

□

Remark 1. *From the theorem above, we can deduce that* $0 \leq \Delta_{k-mer}(S_1, S_2) \leq 1$

Theorem 4. *For any three genomic sequences* S_1, S_2, S_3

$$D_1^k(S_1, S_3) \leq D_1^k(S_1, S_2) + D_1^k(S_2, S_3) \tag{2.10}$$

Proof. Let S_1, S_2, S_3 be three genomic sequences of length l_1, l_2, l_3 respectively.

$$\begin{aligned} D_1^k(S_1, S_3) = \Sigma_{i=1}^{4^k} \left| \frac{f_{S_1}(a_i)}{l_1 - k + 1} - \frac{f_{S_3}(a_i)}{l_3 - k + 1} \right| = \\ \Sigma_{i=1}^{4^k} \left| \frac{f_{S_1}(a_i)}{l_1 - k + 1} + \frac{f_{S_2}(a_i)}{l_2 - k + 1} - \frac{f_{S_2}(a_i)}{l_2 - k + 1} - \frac{f_{S_3}(a_i)}{l_3 - k + 1} \right| \leq \\ \Sigma_{i=1}^{4^k} \left| \frac{f_{S_1}(a_i)}{l_1 - k + 1} - \frac{f_{S_2}(a_i)}{l_2 - k + 1} \right| + \Sigma_{i=1}^{4^k} \left| \frac{f_{S_2}(a_i)}{l_2 - k + 1} - \frac{f_{S_3}(a_i)}{l_3 - k + 1} \right| = \\ D_1^k(S_1, S_2) + D_1^k(S_2, S_3) \end{aligned} \tag{2.11}$$

□

Theorem 5. *For any two complete genomic sequences* S_1, S_2 *(containing only nucleutides)*

$$D_1^{k-1}(S_1, S_2) \leq D_1^k(S_1, S_2) \tag{2.12}$$

Proof. Let S_1, S_2 be two genomic sequences of length l_1, l_2 respectively. Let us focus on a specific $(k-1)$-mer and without loss of generality define it as a_i^{k-1}. Now, due to the completeness of the genomic sequences, to expand this $(k-1)$-mer to a (k)-mer we have only 4 options with 4 kinds of nucleotides. That is

$$f_{S_1}(a_i^{k-1}) = f_{S_1}(a_i^{k-1} \cup A) + f_{S_1}(a_i^{k-1} \cup C) + f_{S_1}(a_i^{k-1} \cup T) + f_{S_1}(a_i^{k-1} \cup G) \tag{2.13}$$

Where $a_i^{k-1} \cup C$ is a_i^{k-1} concatenated with C. Therefore, the contribution of a specific k-mer to the distance calculation is

$$f_{S_1}(a_i^k) - f_{S_2}(a_i^k) = \sum_{n}^{\{A,C,T,G\}} f_{S_1}(a_i^{k-1} \cup n) - f_{S_2}(a_i^{k-1} \cup n) \tag{2.14}$$

Using (2.13) and (2.13)

$$\begin{aligned} D_1^{k-1}(S_1, S_2) = \sum_{i=1}^{4^{k-1}} \left| \frac{f_{S_1}(a_i^{k-1})}{l_1 - (k-1) + 1} - \frac{f_{S_2}(a_i^{k-1})}{l_2 - (k-1) + 1} \right| \leq \\ \sum_{i=1}^{4^{k-1}} \sum_{n}^{\{A,C,T,G\}} \left| \frac{f_{S_1}(a_i^{k-1} \cup n)}{l_1 - k + 1} - \frac{f_{S_2}(a_i^{k-1} \cup n)}{l_2 - k + 1} \right| = \\ \sum_{i=1}^{4^k} \left| \frac{f_{S_1}(a_i^k)}{l_1 - k + 1} - \frac{f_{S_2}(a_i^k)}{l_2 - k + 1} \right| = \\ D_1^k(S_1, S_2) \end{aligned} \tag{2.15}$$

where the second step is due to the general triangle inequality and the fact that we decreased the denominator. □

Remark 2. *Note that the completeness of the genomic sequences is necessary to complete the proof. Here is a counter example. Assume the following genomic sequences $S_1 = AAANNN$ and $S_2 = CCCNNN$. For $k = 3$ the distance between the two will be $\frac{1}{4} + \frac{1}{4} = \frac{1}{2}$ since $l - k + 1 = 4$. However, for $k = 4$ there are no valid k-mers in both strings and therefore the distance between the two will be 0.*

2.2 Synteny Blocks

Synteny blocks are genetic sequences on two species which consist of similar genes with the same direction. Extraction of synteny blocks between two species can be a complicated calculation. First, we need to extract the genes that appear on both species and map their coordinates and only then try to find blocks that contain the corresponding genes. In this paper, we will present a new method for assessing the probability that two species have reverse complement synteny blocks.

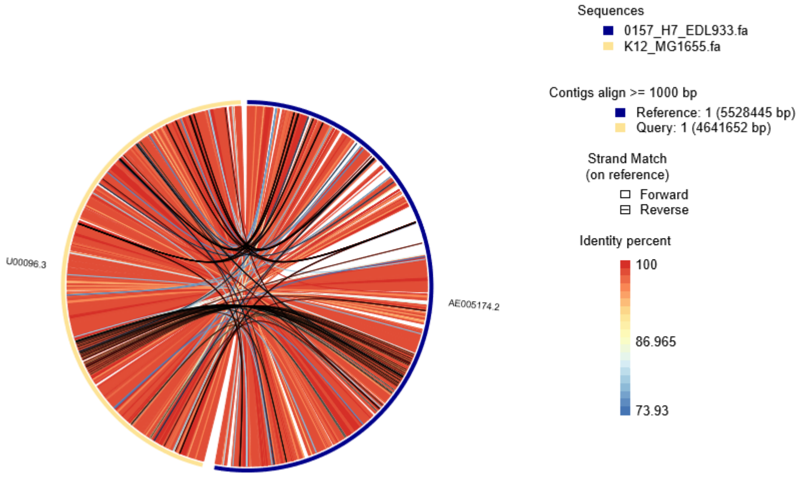


Figure 2.2: Synteny Blocks Between E. Coli 0157-H7-EDL933 and E. Coli K12-MG1655. The colors represent the Identity Percentage where red indicates high percentage and blue indicates low percentage. The black colors indicate an inverted identity, i.e. a reversed complement block

Definition 2.2.1. Synteny Blocks
Let (B_1, B_2) be blocks of bacteria genomes $G = (G_1, G_2)$ respectively. (B_1, B_2) are considered Synteny Blocks if there exists a group of genes $(g_1, g_2, \ldots, g_m)$ that exists on each of the blocks within the same order. We will consider the number m to be the synteny number.

Note that synteny 0 indicates that there are no matching genes between any of the species, and synteny number 1 indicates that there are no groups of genes that appear consecutively in all of the species. Therefore, we will refer to the synteny number to be greater than 1.

Definition 2.2.2. Inverse Synteny Blocks
Let (B_1, B_2) be blocks of bacteria genomes $G = (G_1, G_2)$ respectively. (B_1, B_2) are considered Reverse Complement Synteny Blocks if there exists a group of genes

$(g_1, g_2, \ldots, g_m)$ that exists on one of the blocks and the group $(\overline{g_m}, \overline{g_{m-1}}, \ldots, \overline{g_1})$ exists on the the other block. $\overline{g_i}$ indicates the reverse complement of the gene g_i.

2.2.1 Finding Synteny Blocks

In order to find syntenic blocks we used the BLAST algorithm to identify local alignments of sequences. The output of the BLAST algorithm are blocks of the two input sequences that have good alignment, reflected through the BLAST score. We used the *R* package *OmicCircus* by [8] in order to visualize synteny blocks between species. From the BLAST output, we extracted the synteny blocks that had identity percentage higher than 90% and calculated the sum of sizes of the synteny and reversed complement synteny blocks, we marked these values as $L_{synteny}$ and $L_{synteny-rc}$, respectively.

2.2.2 Synteny Ratio

Similar to the k-mer ratio definition, we looked at the ratio between the synteny blocks and the reverse complement synteny blocks. The synteny ratio represents the fraction of synteny blocks that reversed their alignment from the synteny blocks which kept their alignment. For example, if there would be no mutations between two species, we should expect the synteny ratio to be 0 since the number of reverse synteny blocks would be 0. Otherwise, if all the synteny blocks were reversed, we expect the synteny ratio to be 1.

Definition 2.2.3. Synteny Length
Let $B = (B_1, B_2, \cdots, B_n)$ be the synteny blocks of species (S_1, S_2), and $\overline{B} = (\overline{B_1}, \overline{B_2}, \cdots, \overline{B_m})$ the reverse complement synteny blocks of species (S_1, S_2). We define the synteny length and reversed complement synteny length of (S_1, S_2) as:

$$L_{synteny}(S_1, S_2) = \sum_{b \in B} \sum_{g \in b} |g|, \qquad L_{synteny-rc}(S_1, S_2) = \sum_{b \in \overline{B}} \sum_{g \in b} |g|$$

We define *synteny ratio* to be

$$\Delta_{synteny}(S_1, S_2) = \frac{L_{synteny-rc}(S_1, S_2)}{L_{synteny}(S_1, S_2) + L_{synteny-rc}(S_1, S_2)}$$

We will define synteny percentages for later analysis.

Definition 2.2.4. Synteny Percentage
Let (S_1, S_2) be two sequences.

$$P_{synteny}(S_1, S_2) = \frac{L_{synteny}(S_1, S_2)}{\max\{l_1, l_2\}}$$

Similarly

$$P_{synteny-rc}(S_1, S_2) = \frac{L_{synteny}(S_1, S_2) + L_{synteny-rc}(S_1, S_2)}{\max\{l_1, l_2\}}$$

We will focus on analyzing the correlation between k-mer distances and synteny sizes and the correlation between $\Delta_{k-mer}(S_1, S_2)$ and $\Delta_{synteny}(S_1, S_2)$.

Chapter 3

Results

In our research we analyzed the correlation between the k-mer distance to synteny blocks. Our data-set (3.1, 3.2) consists of 24 species of *ecoli* referenced from [12] and 15 species of *Salmonella Enterica*. Based on this data set, we computed pairwise k-mer distance and reversed complement k-mer distance. Furthermore, we computed the synteny blocks and reversed synteny blocks between each of the species pairs and between the species.

Figure 3.1: E. Coli Species Data Set

E. Coli Species					
Id	Species	Size(bp)	No. genes	Accession Number	Reference
1	E. coli 0157:H7 EDL933	5,620,522	5,312	AE005174	[17]
2	E. coli 0157:H7 Sakai	5,594,477	5,230	BA000007	[7]
3	E. coli 0111:H- 11128	5,766,081	5,407	AP010960	[14]
4	E. coli O26:H11 11368	5,851,458	5,516	AP010958	[14]
5	E. coli 536	4,938,920	4,620	CP000247	[3]
6	E. coli 55989	5,154,862	4,763	CU928145	[4]
7	E. coli APECO1	5,497,653	4,428	CP000468	[11]
9	E. coli CFT073	5,231,428	5,339	AE014075	[24]
10	E. coli 0127:H6 E2348/69	5,069,678	4,554	FM180568	[9]
11	E. coli E24377A	5,249,288	4,749	CP000800	[18]
12	E. coli 0157:H7 EC4115	5,704,171	5,315	CP001164	
13	E. coli ED1a	5,209,548	4,915	CU928162	[23]
14	E. coli HS	4,643,538	4,378	CP000802	[18]
15	E. coli IAI1	4,700,560	4,353	CU928160	[23]
16	E. coli K12 MG1655	4,639,675	4,149	U00096	[1]
17	E. coli K12 W3110	4,646,332	4,226	AP009048	[6]
18	E. coli B str. REL606	4,629,812	4,205	CP000819	[10]
19	E. coli S88	5,032,268	4,696	CU928161	[23]
20	E. coli SE11	5,155,626	4,679	AP009240	[15]
21	E. coli SE15	4,839,683	4,488	AP009378	[22]
22	E. coli SMS-3-5	5,215,377	4,743	AP009378	[5]
23	E. coli UMN026	5,324,391	4,826	CU928163	[23]
24	E. coli UTI89	5,179,971	5,021	CP000243	[2]

Figure 3.2: Salmonella Enterica Species Data Set

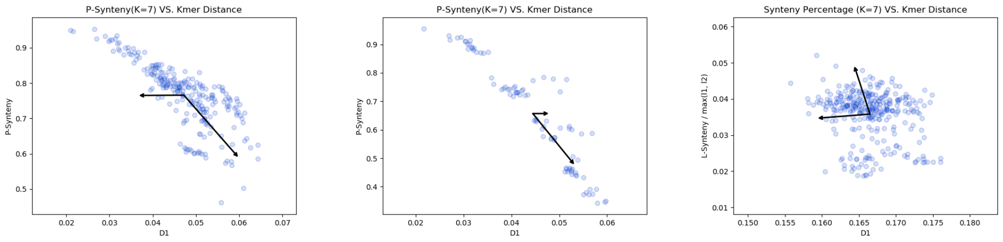
Salmonella Enterica Species			
Id	Species	Size(bp)	Accession Number
1	S. Enterica serovar Typhimurium	4,951,383	ASM694v2
2	S. Enterica serovar Typhi	5,133,713	ASM19599v1
3	S. Enterica serovar Choleraesuis	4,944,000	ASM810v1
4	S. Enterica serovar Enteritidis	4,685,848	ASM950v1
5	S. Enterica serovar Gallinarum	4,658,697	ASM952v1
6	S. Enterica serovar Paratyphi A	4,585,229	ASM1188v1
7	S. Enterica serovar Newport	5,007,719	ASM1604v1
8	S. Enterica serovar Paratyphi C	4,888,494	ASM1838v1
9	S. Enterica serovar Paratyphi B	4,858,887	ASM1870v1
10	S. Enterica serovar Heidelberg	4,983,515	ASM2070v1
11	S. Enterica serovar Schwarzengrund	4,823,887	ASM2074v1
12	S. Enterica serovar Agona	4,836,638	ASM2088v1
13	S. Enterica serovar Dublin	4,917,459	ASM2092v1
14	S. Enterica serovar Montevideo	4,694,375	ASM18895v5

The sequences were taken from NCBI

3.1 Correlation Between $P_{synteny}$ and D_1

We evaluated D_1 and synteny sizes of all pairs from each species on k values from 1 to 10. On each k value, we ran *PCA* ([16]) and the Pearson correlation coefficient in order to identify the correlation between the two measures. We expect that $P_{synteny}$ will have a strong correlation to D_1 since both measures rely on the positive strand.

Figure 3.3: $P_{synteny}$ VS. D_1 (k=7)



(a) E. Coli (b) Salmonella (c) E. Coli and Salmonella

The arrows represent the principal components of PCA

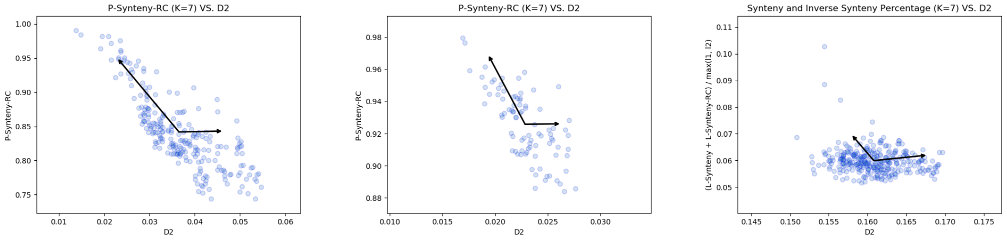
As seen in 3.3, there is a strong correlation between D_1 and $P_{synteny}$ both on E. Coli (a) and Salmonella (b) species. On (c) we can identify that there is no significant correlation between E. Coli and Salmonella, as expected.

However, since D_1 only addresses regular and not inversed synteny blocks, species that are evolutionary close to each other that contain only mutual inversed syntenies will still have a relatively large D_1 value (further explained in 3.6).

3.2 Correlation Between $P_{synteny-rc}$ and D_2

We evaluated D_2 and synteny sizes (both regular and inverse) of all pairs from each species on k values from 1 to 10. On each k value, we ran *PCA* ([16]) and the Pearson correlation coefficient in order to identify the correlation between the two measures. We expect that $P_{synteny-rc}$ will have a strong correlation to D_2 since both measures rely on both strands.

Figure 3.4: $P_{synteny-rc}$ VS. D_2 (k=7)



(a) E. Coli (b) Salmonella (c) E. Coli and Salmonella

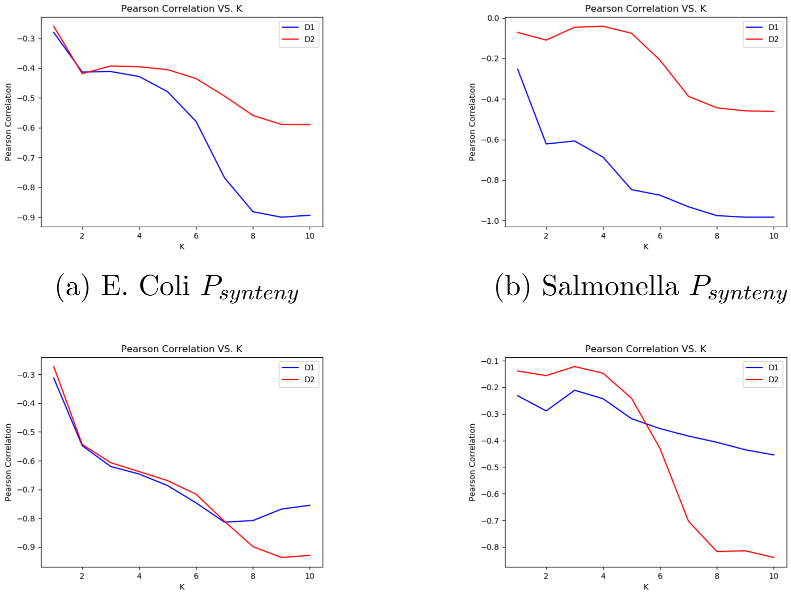
The arrows represents the principal components of PCA

As seen in 3.4, there is a strong correlation between D_2 and $P_{synteny-rc}$ both across E. Coli (a) and Salmonella (b) species. On (c) we can identify that there is no significant correlation between E. Coli and Salmonella, as expected.

3.3 Comparison between D_1 and D_2

As seen in 3.5, D_1 has a better correlation with the $P_{synteny}$ both on E. Coli and Salmonella (see (a) and (b)) than D_2. In addition, D_2 distance has a better correlation with $P_{synteny-rc}$ than D_1. This is expected since species that have larger numbers of mutual inverted sequences will still lead to a high D_1 as opposed to D_2 which takes into account these sequences. This leads to the conclusion that D_2 distance is a better measure for evolutionary proximity than D_1.

Figure 3.5: Comparison between D_1 and D_2 with Pearson Correlation



(a) E. Coli $P_{synteny}$

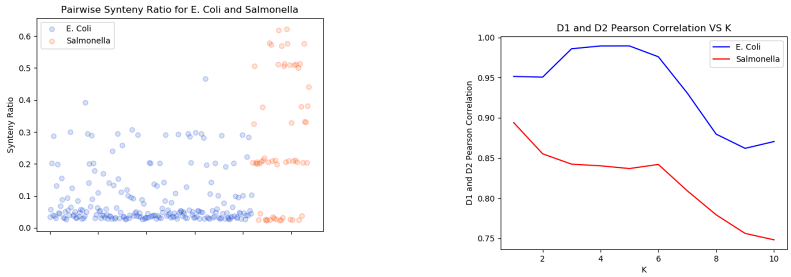
(b) Salmonella $P_{synteny}$

(c) E. Coli $P_{synteny-rc}$

(d) Salmonella $P_{synteny-rc}$

An interesting remark is the difference in the Pearson correlation between the two species. On E. Coli, the Pearson correlation both on D_1 and D_2 is relatively similar as opposed to the Salmonella Pearson correlation in which D_2 Pearson correlation gets stronger and D_1 Pearson correlation stays almost the same. This can lead to the conclusion that Salmonella species contain more inverted syntenies than E. Coli that fails to be identified by D_1.

Figure 3.6: D_1 and D_2 Correlation



(a) Average Synteny Ratio

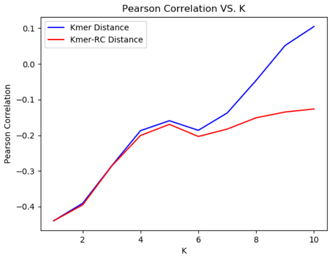
(b) Pearson Correlation Between D_1 and D_2

Figure 3.6a does prove that the average synteny ratio (defined by 2.2.3) is higher in Salmonella than in E. Coli. This may be due to the actual species that are part of the dataset or to the fact that E. Coli species evolutionary age is higher than Salmonella. In addition, we evaluated the pearson correlation between D_1 and D_2, as seen in 3.6b. The figure displays a very high correlation between the distances on E. Coli and smaller correlation on Salmonella. Note that both correlations starts a significant decrease when k reaches 6 which matches 3.5. Furthermore, note that the correlation between D_1 and D_2 on Salmonella reaches 0.75 which indicates that there are differences between the two measures. This led us to research the ratio between the two that will be further discussed in section 3.5.

3.4 Correlation Between E. Coli and Salmonella

In order to verify our method we evaluated the correlation between the two species. Since there is no actual biological correlation between the two species, we expected to see no actual correlation within our method. As expected, we can see from 3.7 that there is no strong indication of correlation between the two.

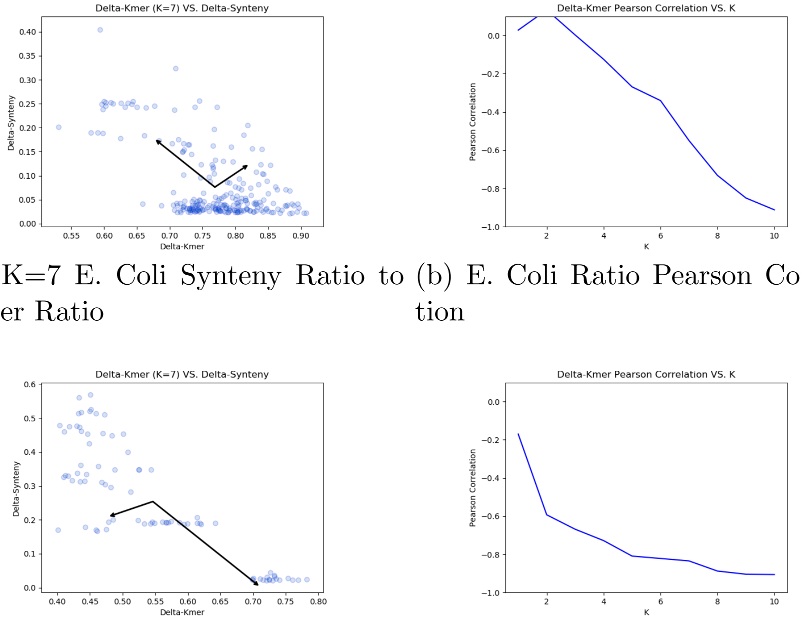
Figure 3.7: E. Coli and Salmonella Pearson Correlation



3.5 Correlation Between $\Delta_{synteny}$ and Δ_{kmer}

We evaluated the k-mer ratio and the synteny ratio of all the pairs from the 2 species on k values from 1 to 10. On each k value, we ran a *PCA* in order to identify the correlation between the two measures and calculated the Pearson correlation coefficient. The result identifies a strong correlation between the two which indicates that the k-mer ratio increases as the percentage of reversed complement synteny blocks decreases. This can be explained by the fact that when the k-mer ratio increases, it means that the size of synteny blocks is similar to the size of the inversed synteny blocks (otherwise there won't be a difference between D_1 and D_2). As shown in 3.8, the E. Coli correlation is much stronger than the Salmonella correlation. However, note that the synteny ratios in (a) are much smaller than in (c).

Figure 3.8: Synteny Ratio VS. k-mer Ratio



(a) K=7 E. Coli Synteny Ratio to k-mer Ratio

(b) E. Coli Ratio Pearson Correlation

(c) K=7 Salmonella Synteny Ratio to k-mer Ratio

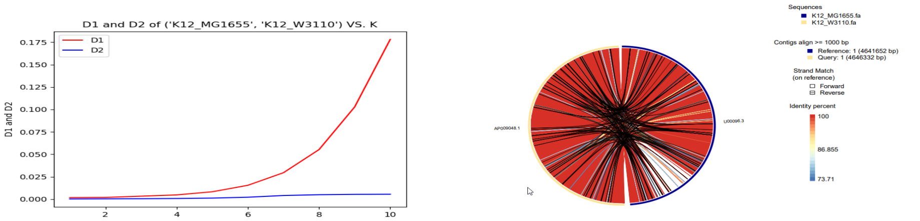
(d) Salmonella Ratio Pearson Correlation

The arrows represents the principal components of PCA

3.6 Outliers

In our experiments, one pair of the species *(K12-MG1655, K12-W3110)* had a significant difference in its k-mer ratio. As seen in 3.9 (a), D_1 increases while the D_2 distance stays the same and very close to 0. This leads the ratio to increase with the increase of k. However, the synteny size does not change with k and is close to 0.6 percent of the whole sequence as opposed to the reverse complement synteny size which is close to 0.1. The synteny blocks can be seen in figure 3.9 (b). Therefore, the k-mer ratio is reaching 0 with the increase of k while the synteny ratio stays the same. The fact that more than half of the sequences are synteny blocks is not surprising since the species are genetically close to each other. This pair is an outlier example, the k-mer ratio reaches 0 with the increase of k which may indicate that there are not a lot of inversed syntenies. However, this is not true as most of the syntenies between these species are inversed.

Figure 3.9: K12-MG1655 and K12-W3110 pair k-mer Analysis



(a) k-mer Distances Analysis

(b) K12-MG1655 and K12-W3110 synteny blocks

3.7 Implications for Human Species

The results we presented in this paper anchor on the fact that E. Coli and Salmonella species are bacteria in which most of their genome represents actual genes. Therefore, the correlation between the k-mers distance ratio and the reversed complement synteny blocks percentage can be explained by the fact that most of the k-mers represent genes; since genes determine the synteny blocks, the correlation is implied. For other species like Eukaryotes, the evolutionary distance is larger and therefore there are less reversed complement synteny blocks. This will cause the synteny ratio to reach high numbers and will fail to detect reverse complement sequences. However, as part of analyzing the human data, we found some interesting facts that are worth mentioning. We used the human chromosomes of hg38 and hg19 taken from the UCSC genome browser and compared the *k-mer* and *k-mer-rc* distances between the pairwise chromosomes.

3.7.1 What is considered a low k-mer Distance?

In order to achieve a measure of a distance which can be considered low we computed D_2 distances between hg38 and hg19. The distances between each chromosome to itself within the other version can be thought of the error measure between the two versions.

Figure 3.10: K=10 hg38 VS hg19 Masked D2 Heatmap

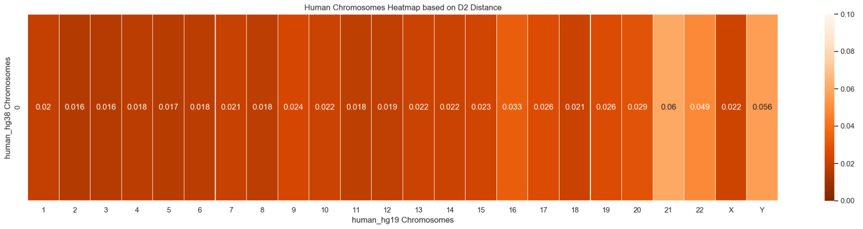
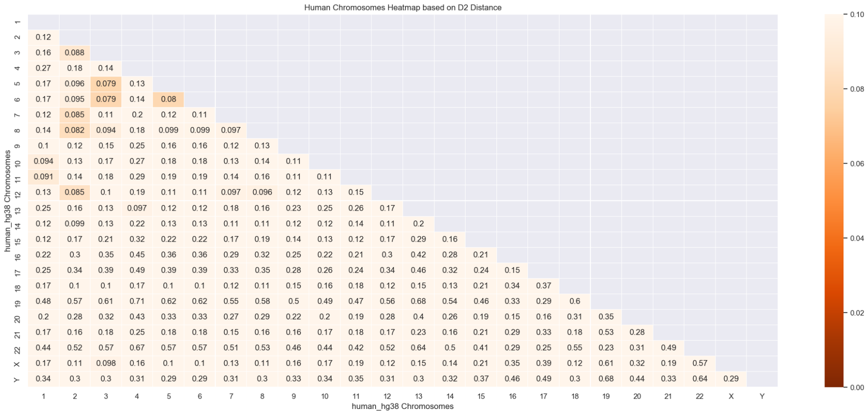


Figure 3.10 shows the distances between both versions on each of the chromosomes to itself with k=10 (the full heatmap is shown in 5.3). We can identify that the maximum distance between relatively long chromosomes (1-10) is 0.024. Therefore, we will consider a D_2 distance lower than 0.024 to be a low k-mer distance for k=10.

3.7.2 Low Distance Human Chromosomes

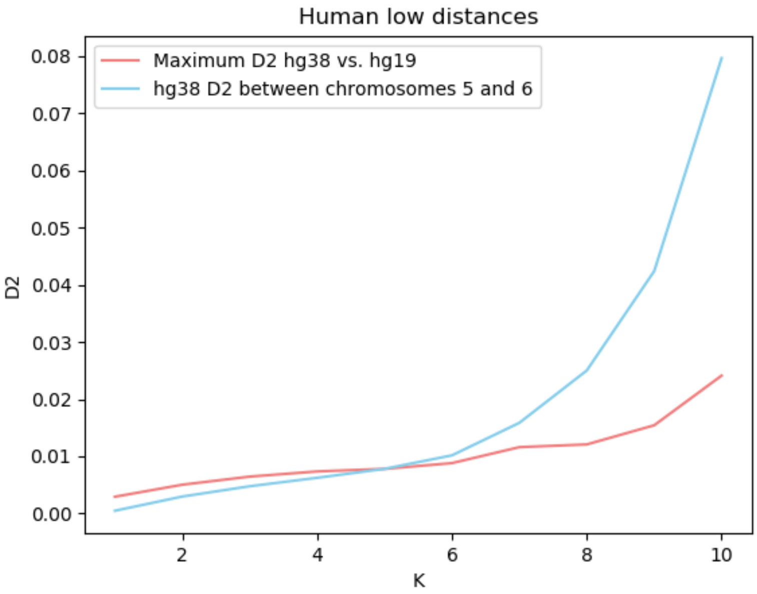
We compared both the masked version of hg38 chromosomes to identify what chromosomes are considered to have low distance between them. From figure 3.11 there are several clusters of chromosomes that have low distance between them.

Figure 3.11: K=10 Human (hg38) Masked D_2 Distance Heatmap



We specifically identified chromosomes 5 and 6 to have very small distance with most of the k values. As seen in 3.12, the distance between the two chromosomes is smaller than the low k-mer distance between hg38 and hg19 until k reaches 5. This may be a good indication for the evolutionary similarity between two chromosomes.

Figure 3.12: Human (hg38) Masked Chromosomes 5-6 Distances



3.8 Phylogenetic Tree Reconstruction based on k-mer distance

By using k-mer distance we can reconstruct phylogenetic trees to identify phylogeny between species. This can help in classifying species into their sub groups or understanding a different phylogeny based on k-mer distances. The benefit of using k-mer distance similarity is time complexity. Other similarity algorithms are using string similarity matching which can be expensive when using long genomic sequences.

In order to reconstruct a phylogenetic tree, we calculated the pairwise E. Coli k-mer distance between all the species and based on that distance matrix we used hierarchical clustering to build the tree.

Figure 3.13: K=7 E. Coli Phylogenetic Tree Reconstruction

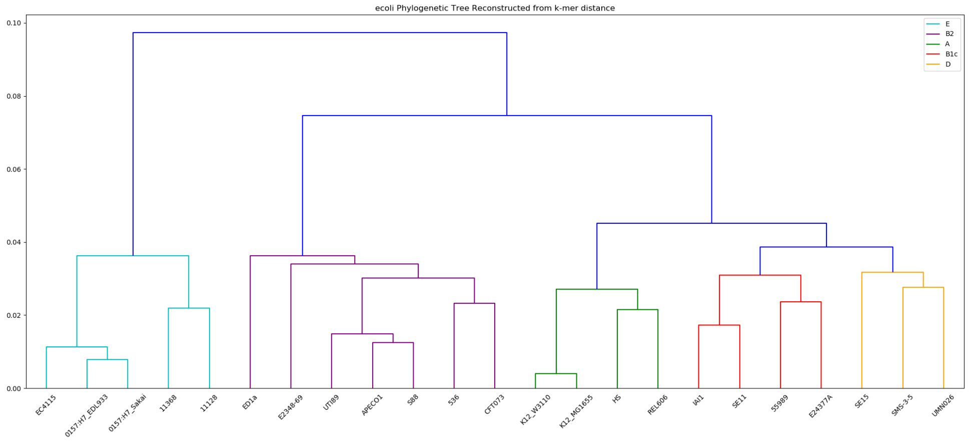
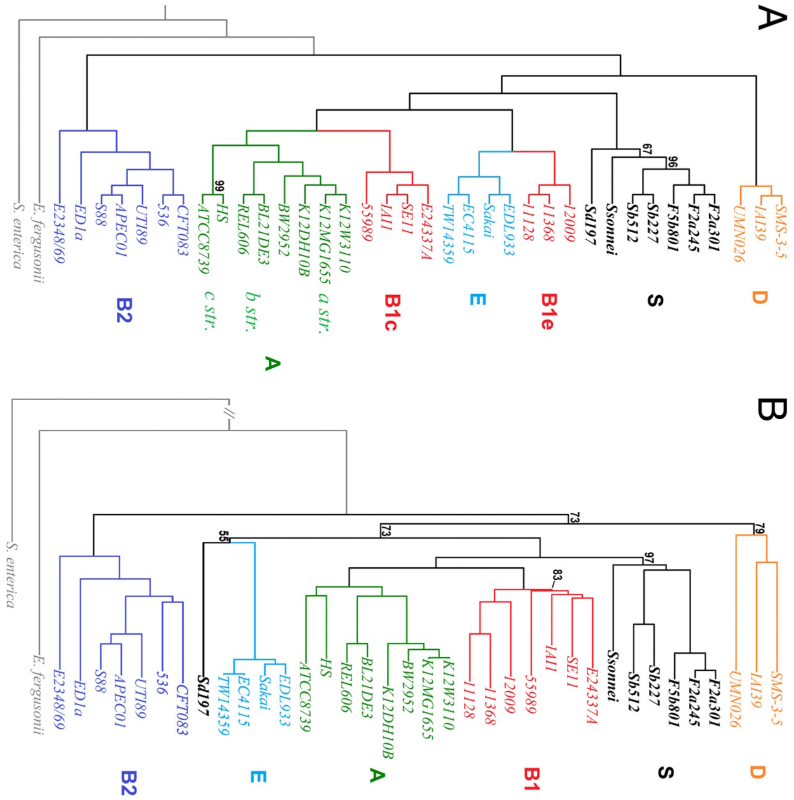


Figure 3.13 shows the reconstructed phylogenetic tree from E. Coli k-mer distances. The cluster colors are selected by distance threshold, in this case 0.037. The cluster groups (as shown in the legend) are manually added to help the assessment of the clustering quality. In 3.14 there are two E. Coli phylogenetic trees presented by [21]. Both trees are constructed using FFPs (feature frequency profile). The first tree (A) is constructed with all features and as the author states, the tree is likely to reflect phenetic relationships. The second tree (B) uses genomic features, not genes, which is thus likely to reflect evolutionary history. The FFP method is similar to our method, aside from including the reversed complement k-mers and using Jensen-Shannon divergence instead of vector norm between the frequency vectors. Looking at the cluster groups, we can identify that there is high similarity between our reconstructed tree to the one suggested in 3.14. However, our tree is reconstructed with $k = 7$ whereas the paper reconstructs their tree using features of length 24, i.e. $k = 24$. Our selection of $k = 7$ is widely discussed in the previous sections and correlates both with the length of the species sequences and with the quality of the reconstructed tree. In 5.4 there is a side by side comparison between our phylogenetic tree to the one in 3.14 (B).

Figure 3.14: E. Coli Phylogenetic Tree from [21]. (A) is using all features without any filtering and (B) uses genomic features, not genes to construct the tree



Chapter 4

Conclusions

4.1 Conclusions

We have introduced measures of k-mer distances, and applied them to bacteria and to human chromosomes. The two measures, D_1 and D_2 were compared to synteny measures in bacteria. We identified a strong correlation between D1 to regular syntenic regions and a strong correlation between D_2 to regular and inversed syntenies which indicates evolutionary similarity between two species. We argued that the k-mer ratio is a good measure for identifying the existence of large inversed syntenies.

Our method provides a good measure for similarity and is different from traditional similarity measures. Two important differences are that: A, the k-mer distances do not take into account prior knowledge such as genes and low-complexity regions. B, the time complexity of using k-mer distances depends mostly on the value of k while other similarity measures are at least quadratic in the length of the genomic sequences. We have shown that for relatively low k values (depending on the K-Limit), our method competes well with other methods.

Applying k-mer distance evaluations to human chromosomes we have argued that chromosomes 5 and 6 display very small evolutionary distances. We suggest using k-mer distances as the first step of evolutionary similarity assessment before applying additional string matching algorithms. This can help pointing out sequences that with high probability are not close to each other or identifying sequences that have a large amount of mutual syntenies.

Chapter 5

Appendix

5.1 D_1 and D_2 Distances are Monotonic

Figure 5.1 demonstrates both the monotonic property of the k-mer and k-mer-RC distances and theorems 3 and 5 which indicates for any two given sequences S_1, S_2 $D_2^k(S_1, S_2) \leq D_1^k(S_1, S_2)$ and $D_1^{k-1}(S_1, S_2) \leq D_1^k(S_1, S_2)$

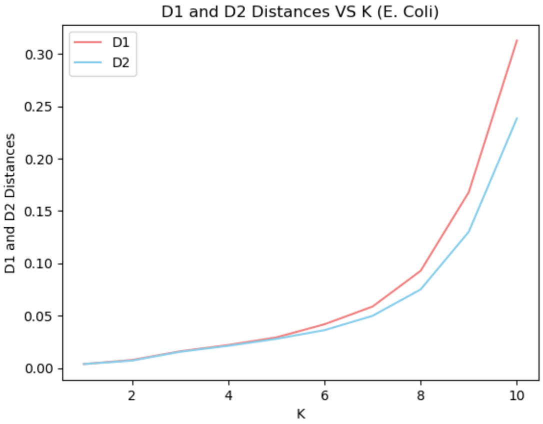
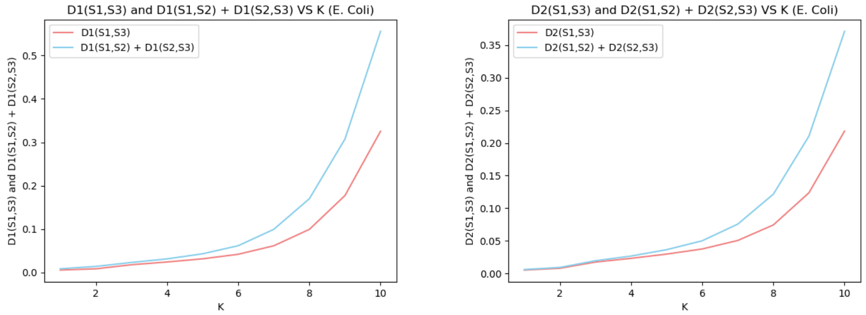


Figure 5.1: k-mer and k-mer-RC Distances VS K

5.2 D_1 and D_2 Distance Triangle Inequality

Figure 5.2 demonstrates triangle inequality property of the k-mer and k-mer-RC distances and theorem 3 which indicates for any three given sequences S_1, S_2, S_3 $D_1^k(S_1, S_3) \leq D_1^k(S_1, S_2) + D_1^k(S_2, S_3)$.

Figure 5.2: D_1 and D_2 Distances Triangle Inequality

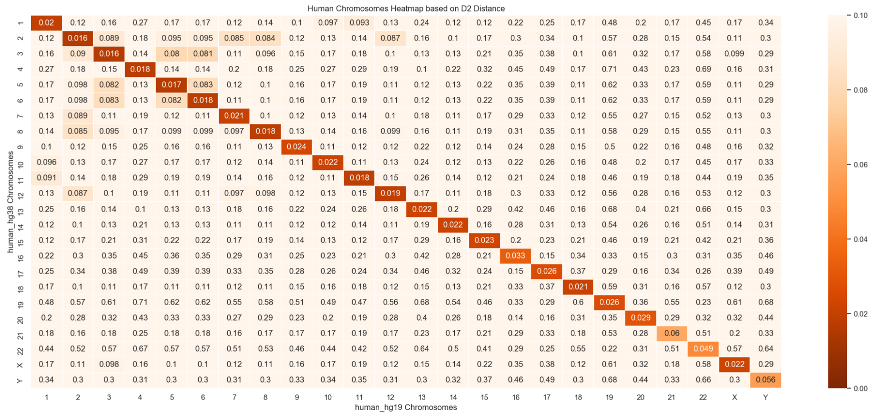


(a) K=8 E. Coli

(b) K=9 E. Coli

5.3 Human hg38 VS. Human hg19 Distance Matrix

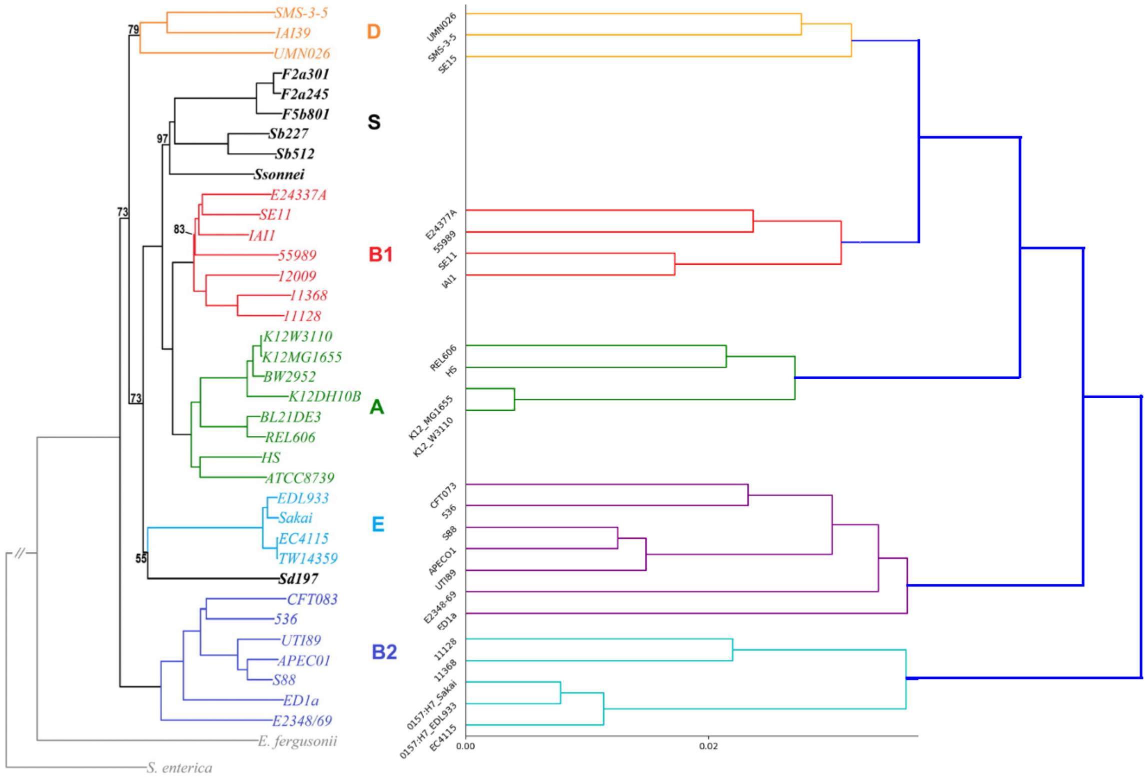
Figure 5.3: Human hg38 VS. Human hg19 Distance Matrix (k=10)



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	X	Y
1	0.02	0.12	0.16	0.27	0.17	0.17	0.12	0.14	0.1	0.097	0.093	0.13	0.24	0.12	0.12	0.22	0.25	0.17	0.48	0.2	0.17	0.45	0.17	0.34
2	0.12	0.016	0.089	0.18	0.095	0.095	0.085	0.084	0.12	0.13	0.14	0.087	0.16	0.1	0.17	0.3	0.34	0.1	0.57	0.28	0.15	0.54	0.11	0.3
3	0.16	0.09	0.016	0.14	0.08	0.081	0.11	0.096	0.15	0.17	0.18	0.1	0.13	0.13	0.21	0.35	0.38	0.1	0.61	0.32	0.17	0.58	0.099	0.29
4	0.27	0.18	0.15	0.018	0.14	0.14	0.2	0.18	0.25	0.27	0.29	0.19	0.1	0.22	0.32	0.45	0.49	0.17	0.71	0.43	0.23	0.69	0.16	0.31
5	0.17	0.098	0.082	0.13	0.017	0.083	0.12	0.1	0.16	0.17	0.19	0.11	0.12	0.13	0.22	0.35	0.39	0.11	0.62	0.33	0.17	0.59	0.11	0.29
6	0.17	0.098	0.083	0.13	0.082	0.018	0.11	0.1	0.16	0.17	0.19	0.11	0.12	0.13	0.22	0.35	0.39	0.11	0.62	0.33	0.17	0.59	0.11	0.29
7	0.13	0.089	0.11	0.19	0.12	0.11	0.021	0.1	0.12	0.13	0.14	0.1	0.18	0.11	0.17	0.29	0.33	0.12	0.55	0.27	0.15	0.52	0.13	0.3
8	0.14	0.085	0.095	0.17	0.099	0.099	0.097	0.018	0.13	0.14	0.16	0.099	0.16	0.11	0.19	0.31	0.35	0.11	0.58	0.29	0.15	0.55	0.11	0.3
9	0.1	0.12	0.15	0.25	0.16	0.16	0.11	0.13	0.024	0.11	0.12	0.12	0.22	0.12	0.14	0.24	0.28	0.15	0.5	0.22	0.16	0.48	0.16	0.32
10	0.096	0.13	0.17	0.27	0.17	0.17	0.12	0.14	0.11	0.022	0.11	0.13	0.24	0.12	0.13	0.22	0.26	0.16	0.48	0.2	0.17	0.45	0.17	0.33
11	0.091	0.14	0.18	0.29	0.19	0.19	0.14	0.16	0.12	0.11	0.018	0.15	0.26	0.14	0.12	0.21	0.24	0.18	0.46	0.19	0.18	0.44	0.19	0.35
12	0.13	0.087	0.1	0.19	0.11	0.11	0.097	0.098	0.12	0.13	0.15	0.019	0.17	0.11	0.18	0.3	0.33	0.12	0.56	0.28	0.16	0.53	0.12	0.3
13	0.25	0.16	0.14	0.1	0.13	0.13	0.18	0.16	0.22	0.24	0.26	0.18	0.022	0.2	0.29	0.42	0.46	0.16	0.68	0.4	0.21	0.66	0.15	0.3
14	0.12	0.1	0.13	0.21	0.13	0.13	0.11	0.11	0.12	0.12	0.14	0.11	0.19	0.022	0.16	0.28	0.31	0.13	0.54	0.26	0.16	0.51	0.14	0.31
15	0.12	0.17	0.21	0.31	0.22	0.22	0.17	0.19	0.14	0.13	0.12	0.17	0.29	0.16	0.023	0.2	0.23	0.21	0.46	0.19	0.21	0.42	0.21	0.36
16	0.22	0.3	0.35	0.45	0.36	0.35	0.29	0.31	0.25	0.23	0.21	0.3	0.42	0.28	0.21	0.033	0.15	0.34	0.33	0.15	0.3	0.31	0.35	0.46
17	0.25	0.34	0.38	0.49	0.39	0.39	0.33	0.35	0.28	0.26	0.24	0.34	0.46	0.32	0.24	0.15	0.026	0.37	0.29	0.16	0.34	0.26	0.39	0.49
18	0.17	0.1	0.11	0.17	0.11	0.11	0.12	0.11	0.15	0.16	0.18	0.12	0.15	0.13	0.21	0.33	0.37	0.021	0.59	0.31	0.16	0.57	0.12	0.3
19	0.48	0.57	0.61	0.71	0.62	0.62	0.55	0.58	0.51	0.49	0.47	0.56	0.68	0.54	0.46	0.33	0.29	0.6	0.026	0.36	0.55	0.23	0.61	0.68
20	0.2	0.28	0.32	0.43	0.33	0.33	0.27	0.29	0.23	0.2	0.19	0.28	0.4	0.26	0.18	0.14	0.16	0.31	0.35	0.029	0.29	0.32	0.32	0.44
21	0.18	0.16	0.18	0.25	0.18	0.18	0.16	0.17	0.17	0.17	0.19	0.17	0.23	0.17	0.21	0.29	0.33	0.18	0.53	0.28	0.06	0.51	0.2	0.33
22	0.44	0.52	0.57	0.67	0.57	0.57	0.51	0.53	0.46	0.44	0.42	0.52	0.64	0.5	0.41	0.29	0.25	0.55	0.22	0.31	0.51	0.049	0.57	0.64
X	0.17	0.11	0.098	0.16	0.1	0.1	0.12	0.11	0.16	0.17	0.19	0.12	0.15	0.14	0.22	0.35	0.38	0.12	0.61	0.32	0.18	0.58	0.022	0.29
Y	0.34	0.3	0.3	0.31	0.3	0.3	0.31	0.3	0.33	0.34	0.35	0.31	0.3	0.32	0.37	0.46	0.49	0.3	0.68	0.44	0.33	0.66	0.3	0.056

5.4 E. Coli Phylogenetic Tree Reconstruction Comparison

Figure 5.4: E. Coli Phylogenetic Tree Reconstruction Comparison (k=7)



References

[1] Frederick R Blattner et al. “The complete genome sequence of Escherichia coli K-12”. In: *science* 277.5331 (1997), pp. 1453–1462.

[2] Swaine L Chen et al. “Identification of genes subject to positive selection in uropathogenic strains of Escherichia coli: a comparative genomics approach”. In: *Proceedings of the National Academy of Sciences* 103.15 (2006), pp. 5977–5982.

[3] Ulrich Dobrindt et al. “Genetic structure and distribution of four pathogenicity islands (PAI I536 to PAI IV536) of uropathogenic Escherichia coli strain 536”. In: *Infection and immunity* 70.11 (2002), pp. 6365–6372.

[4] Ulrich Dobrindt et al. “Genetic structure and distribution of four pathogenicity islands (PAI I536 to PAI IV536) of uropathogenic Escherichia coli strain 536”. In: *Infection and immunity* 70.11 (2002), pp. 6365–6372.

[5] W Florian Fricke et al. “Insights into the environmental resistance gene pool from the genome sequence of the multidrug-resistant environmental isolate Escherichia coli SMS-3-5”. In: *Journal of bacteriology* 190.20 (2008), pp. 6779–6794.

[6] Koji Hayashi et al. “Highly accurate genome sequences of Escherichia coli K-12 strains MG1655 and W3110”. In: *Molecular systems biology* 2.1 (2006).

[7] Tetsuya Hayashi et al. “Complete genome sequence of enterohemorrhagic Eschelichia coli O157: H7 and genomic comparison with a laboratory strain K-12”. In: *DNA research* 8.1 (2001), pp. 11–22.

[8] Ying Hu et al. “OmicCircos: a simple-to-use R package for the circular visualization of multidimensional omics data”. In: *Cancer informatics* 13 (2014), CIN–S13495.

[9] Atsushi Iguchi et al. “Complete genome sequence and comparative genome analysis of enteropathogenic Escherichia coli O127: H6 strain E2348/69”. In: *Journal of Bacteriology* 191.1 (2009), pp. 347–354.

[10] Haeyoung Jeong et al. “Genome sequences of Escherichia coli B strains REL606 and BL21 (DE3)”. In: *Journal of molecular biology* 394.4 (2009), pp. 644–652.

[11] Timothy J Johnson et al. “The genome sequence of avian pathogenic Escherichia coli strain O1: K1: H7 shares strong similarities with human extraintestinal pathogenic E. coli genomes”. In: *Journal of bacteriology* 189.8 (2007), pp. 3228–3236.

[12] Oksana Lukjancenko, Trudy M Wassenaar, and David W Ussery. “Comparison of 61 sequenced Escherichia coli genomes”. In: *Microbial ecology* 60.4 (2010).

[13] Guillaume Marçais and Carl Kingsford. “A fast, lock-free approach for efficient parallel counting of occurrences of k-mers”. In: *Bioinformatics* 27.6 (2011), pp. 764–770.

[14] Yoshitoshi Ogura et al. “Comparative genomics reveal the mechanism of the parallel evolution of O157 and non-O157 enterohemorrhagic Escherichia coli”. In: *Proceedings of the National Academy of Sciences* 106.42 (2009), pp. 17939–17944.

[15] Kenshiro Oshima et al. “Complete genome sequence and comparative analysis of the wild-type commensal Escherichia coli strain SE11 isolated from a healthy adult”. In: *DNA research* 15.6 (2008), pp. 375–386.

[16] Karl Pearson. “LIII. On lines and planes of closest fit to systems of points in space”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11 (1901), pp. 559–572.

[17] Nicole T Perna et al. “Genome sequence of enterohaemorrhagic Escherichia coli O157: H7”. In: *Nature* 409.6819 (2001), p. 529.

[18] David A Rasko et al. “The pangenome structure of Escherichia coli: comparative genomic analysis of E. coli commensal and pathogenic isolates”. In: *Journal of bacteriology* 190.20 (2008), pp. 6881–6893.

[19] Rivka Rudner and Virgilija Remeza. “Chromatographically fractionated complementary strands of Bacillus subtilis deoxyribonucleic acid: biological properties”. In: *Journal of bacteriology* 113.2 (1973), pp. 739–753.

[20] Sagi Shporer et al. “Inversion symmetry of DNA k-mer counts: validity and deviations”. In: *BMC genomics* 17.1 (2016), p. 696.

[21] Gregory E Sims and Sung-Hou Kim. “Whole-genome phylogeny of Escherichia coli/Shigella group by feature frequency profiles (FFPs)”. In: *Proceedings of the National Academy of Sciences* 108.20 (2011), pp. 8329–8334.

[22] Marie Touchon et al. “Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths”. In: *PLoS genetics* 5.1 (2009), e1000344.

[23] J Wei et al. “Complete genome sequence and comparative genomics of Shigella flexneri serotype 2a strain 2457T”. In: *Infection and immunity* 71.5 (2003), pp. 2775–2786.

[24] PCY Woo et al. “Then and now: use of 16S rDNA gene sequencing for bacterial identification and discovery of novel bacteria in clinical microbiology laboratories”. In: *Clinical Microbiology and Infection* 14.10 (2008), pp. 908–934.

אוניברסיטת תל אביב

הפקולטה למדעים מדוייקים ע"ש ריימונד ובברלי סאקלר

על קשרים בין מרחק קמרים לבין קטעים סינתניים

חיבור זה הוגש כחלק מהדרישות לקבלת התואר
"מוסמך אוניברסיטה" - M.Sc באוניברסיטת תל-אביב

ביה"ס למדעי המחשב ע"ש בלבטניק

אלון כפרי

העבודה הוכנה בהדרכתם של
פרופ' בני שור ופרופ' דוד הורן

2019

תקציר

בעבודה זו חקרנו את הקשר בין מרחקי קמרים (מרחקים המוגדרים ע"י תדירות של תתי סדרות) לבין קטעים סינתניים (קטעים גנומיים המכילים מספר משותף של גנים). הוכחנו שמרחק קמרים קטן הינו מדד טוב לקיומם של קטעים סינתניים. הראינו שהיחס בין מרחק קמרים המכיל את שני גדילי ה-DNA הינו מדד יותר טוב מאשר מרחק קמרים המכיל רק גדיל אחד לקיומם של קטעים סינתניים. בנוסף, הראינו דוגמאות לכרומוזומים אשר מרחק הקמרים ביניהם קטן מאוד ונתנו דוגמא לשימוש במרחק זה לצורך שחזור עצים פילוגנטיים.