Inter-Chromosomal k-mer Distances

Alon Kafri¹, Benny Chor¹ and David Horn²*

¹School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel
²School of Physics and Astronomy, Tel Aviv University, Tel Aviv 69978, Israel
*corresponding author

Email addresses:
AK: akafri@gmail.com
BC: benny@cs.tau.ac.il
DH: horn@tau.ac.il

Abstract

Background
Inversion Symmetry is a generalization of the second Chargaff rule, stating that the count of a string of k nucleotides on a single chromosomal strand equals the count of its inverse (reverse-complement) k-mer. It holds for many species, both eukaryotes and prokaryotes, for ranges of k which may vary from 7 to 10 as chromosomal lengths vary from 2Mbp to 200 Mbp. Building on this formalism we introduce the concept of k-mer distances between chromosomes. We formulate two distance measures, D1 and D2, where the first takes into account k-mers appearing on single strands of the two chromosomes, whereas the second takes into account both strands.

Results
We first define the various distance measures and summarize their properties. We also define distances that rely on existence of synteny blocks between chromosomes of different strains. Studying E Coli and Salmonella strains, we evaluate the different distance measures, and find correlations between synteny distances and k-mer distances, thus establishing the usefulness of the latter as measures of evolitional proximity of chromosomes. Applying our measures to human genomes, we find that chromosomes 5 and 6 are the closest ones on the k-mer distance evolitional scale.

Conclusions
The novel distances carry information about evolutonal proximity and provide useful tools for future studies. The finding of proximity between human chromosomes 5 and 6 is an examples of a novel insight provided by these tools.

**Keywords:** inversion symmetry, k-mer distances, synteny

**Background**

The phenomenon of Inversion Symmetry (IS) has recently been reevaluated and established in [1]. This generalization of the 2nd Chargaff rule [2] implies that the number of occurrences of any sequence $n(S)$ of length $k$ on a chromosomal strand $S$ is equal to the number of occurrences of its inverse (reverse-complement) sequence $n(S^{inv})$ on the same strand. Another way of stating the same fact is that the number of occurrences $n(S_1)$ on one chromosomal strand is equal to the number of occurrences $n(S_2)$ on the other strand provided both are being read along their own 5' to 3' directions. It has been shown [1] that this rule holds for all $k$ up to some limit KL that was defined as the $k$ value when discrepancies of inversion symmetry reach 10%. KL grows logarithmically with chromosomal length $L$. KL values for mammals are of order 9 or 10, while for bacteria they are of order 7 or 8.

Here we define measures of k-mer distances between chromosomes, within the same organism or between different species. This is carried out by comparing frequencies of all strings of same length $k$ on different chromosomes, summing over one or over both strands of each chromosome. When applying such measures to eukaryotes, it is helpful to use both strands, because there exists an ambiguity as to which strand should be compared between different chromosomes. However, when applying it to bacteria, where the positive strand is uniquely defined, one can use the single strand measure.

We apply these methods to families of *E. Coli* and *Salmonella* strains, comparing k-mer distance measures to synteny distance measures between different strains, finding correlations between the two evolutional proximity measures. Applying such measures to human chromosomes we find that chromosomes 5 and 6 are very close to one another.

**Results**

**Definitions and properties of k-mer distances between chromosomes**

The term k-mer refers to all the possible substrings of length $k$ that are contained in a given chromosomal string of length $L$. The total number of their occurrences is $N=L-k+1$. We define the empirical frequency of a specific k-mer, e.g., $a_1$, in the string $S$ as the number of occurrences of this k-mer in $S$ divided by $N$.
Let us define the k-mer distance $D_1$ as the L1-norm of the difference between the k-dim vectors containing frequencies of all k-mers, when comparing two chromosomal strings (e.g. positive strands of two chromosomes) $S_1$ and $S_2$:

$$D_1^k(S_1, S_2) = \sum_{i=1}^{4^k} |f_i(S_1) - f_i(S_2)|$$

(2)

The index 1 in $D_1$ refers to the fact that we use only one strand in this comparison of two chromosomes.

Similarly, we may define a distance measure $D_2$ by taking into account both strands of the two chromosomes, reading them along their own 5’ to 3’ directions. Since each specific k-mer on the negative strand, is accompanied by its inverse (reverse-complement) on the positive strand, we define $D_2$ as

$$D_2^k(S_1, S_2) = \sum_{i=1}^{4^k} |f_i(S_1) + f_i(S_1) - f_i(S_2) - f_i(S_2)| / 2$$

(3)

where

$$a_i = a_i^{inv}$$

and the summation is once again carried out over single strands of the two chromosomes. Division by 2 is introduced in the definition of $D_2$ because the effective number of counts on each chromosome becomes 2N.

The triangular inequality implies that

$$|f_i(S_1) + f_i(S_1) - f_i(S_2) - f_i(S_2)| \leq |f_i(S_1) - f_i(S_2)| + |f_i(S_1) - f_i(S_2)|$$

(4)

for every i. Since summation over all i is tantamount to summation over all I, because it may be regarded as a change in the order of summation over all k-mers, it follows that

$$D_2^k(S_1, S_2) \leq D_1^k(S_1, S_2)$$

(5)

Using the above definitions we summarize the properties of k-mer distances:

1. Positivity. By definition all distances are non-negative.

2. $D_1^k(S_1, S_2) = 0$ implies equivalence between $S_1$ and $S_2$, in the sense that both strings have the same frequencies. This does not necessarily imply that the two string are equal to each other, because they may differ in length.

3. Symmetry. By definition, $D_1^k(S_1, S_2) = D_1^k(S_2, S_1)$.

4. Inequality (5): $D_2^k(S_1, S_2) \leq D_1^k(S_1, S_2)$, as proved above.

5. Triangular inequalities of distances:
\[ D_{1,2}^k(S_1, S_3) \leq D_{1,2}^k(S_1, S_2) + D_{1,2}^k(S_2, S_3). \] (6)

This can be proved in an analogous fashion to property 4.

6. Inversion symmetry [1] implies that \( D_{1,2}^k(S_1, S_2) = 0 \) if \( S_2 \) is the inverse of \( S_1 \) (or equivalent to it in the sense of property 2). Otherwise this distance will be positive. Such a definition of inversion symmetry has been introduced by [3]. \( D_{1,2}^k(S_1, S_2) = 0 \) is a trivial statement for two strings which are inverses of each other.

7. Monotonic increase with \( k \):

\[ D_{1,2}^{k-1}(S_1, S_2) \leq D_{1,2}^k(S_1, S_2) \] (7)

To prove this property note that a \( k \)-mer \( a_i^k \) can be generated from a corresponding \( a_j^{k-1} \), which coincides with all first \( k-1 \) entries of \( a_i^k \), by adding to it one of the four nucleotides \{A, C, G, T\}. Let us define this set as \{j,i\} for a given \( a_j^{k-1} \) and four corresponding \( a_i^k \). It follows then that

\[ D_{1,2}^{k-1}(S_1, S_2) = \sum_{j=1}^{k-1} |f_j(S_1) - f_j(S_2)| \leq \sum_{i=1}^{k} |f_i(S_1) - f_i(S_2)| = D_{1,2}^k(S_1, S_2) \]

by summing over the indices using the \{j,i\} association, and applying the extended triangular inequality to each set of four \( f_i \) whose \( k \)-mers \( a_i^k \) begin with the same \( (k-1) \)-mer \( a_j^{k-1} \) with index \( j \).

This proof can be trivially extended to \( D_2 \).

One condition for these inequalities to hold is that all \( k \)-mers are realized on the chromosomal strings which are being investigated, i.e. all \( n(a_i^k) > 0 \).

Finally we touch upon the question of the range of \( k \)-values for which the distance measures can be applied.

Shporer et al [1] have introduced the notion of the KL limit. This is the \( k \)-value for which Inversion Symmetry fails at the rate of 10%. They demonstrated that chromosomes of different species, as well as different human chromosomal sections, follow a universal logarithmic slope of \( KL \sim 0.7 \ln(L) \), where \( L \) is the length of the chromosome. This limit can also be derived from the assumption that \( L > 4^k \) allowing for all \( k \)-mers to be expressed on the chromosomal string.

As an example of relevant statistics we display in Fig. 1 the percentage of "zero \( k \)-mers", i.e. those which do not appear on the string, and the distance between two close strains of \( E. Coli \) as function of \( k \), demonstrating that good results are obtained for \( k<KL \).
When evaluating distances between two chromosomal strings with different lengths, $L_1$ and $L_2$, one should limit oneself to $KL$ where $L=\min(L_1, L_2)$, guaranteeing that the same $k$ is valid for both chromosomal strings which are being compared.

**Synteny Blocks**

Synteny blocks are genetic sequences on genomes of two species which consist of homologous genes aligned along the same direction. An example of their importance was demonstrated by [4, 5]. In order to validate the meaning of our k-mer distances, we will compare them with distances based on synteny blocks to be defined here.

For bacteria, where the positive strand is well defined, we differentiate between Direct Synteny Blocks (DSB), appearing along the same strand in both genomes, and Inverse Synteny Blocks (ISB), lying on opposite strands. An example is shown in Fig. 2.

![Fig. 2. Synteny Blocks between E. Coli 0157-H7-EDL933 and E. Coli K12-MG1655. The colors represent the Identity Percentage where red indicates high identity percentage of DSB and blue indicates low identity percentage of DSB. The black colors represent ISBs.](image)
Searching for synteny blocks, we have first used BLAST to identify local alignments of sequences. To visualize results we have used the R package OmicCircus [6]. From the BLAST output, we extracted the synteny blocks that had identity percentage higher than 90%, and calculated the overall sequence lengths of DSB and ISB ($L_{DSB}$ and $L_{ISB}$) respectively.

In our analyses we make use of the percentages of direct synteny

$$P_{DSYN}(S_1, S_2) = \frac{L_{DSB}}{\min(L_1, L_2)} \quad (8)$$

and overall synteny

$$P_{SYN}(S_1, S_2) = \frac{L_{DSB} + L_{ISB}}{\min(L_1, L_2)} \quad (9)$$

where $L_1$ and $L_2$ are the lengths of the chromosomes $S_1$ and $S_2$ which are being compared.

**Distance measures in bacteria**

We study genomes of 23 strains of *E Coli* and 14 strains of *Salmonella Enterica*. They are listed in Tables 1 and 2.

| Id | Species               | Size (bp) | No. genes | Accession Number |
|----|-----------------------|-----------|-----------|------------------|
| 1  | E. coli 0157:H7 EDL933| 5,620,522 | 5,312     | AE005174         |
| 2  | E. coli 0157:H7 Sakai | 5,594,477 | 5,230     | BA000007         |
| 3  | E. coli 0111:H- 11128 | 5,766,081 | 5,407     | AP010960         |
| 4  | E. coli O26:H11 11368 | 5,851,458 | 5,516     | AP010958         |
| 5  | E. coli 536           | 4,938,920 | 4,620     | CP000247         |
| 6  | E. coli 55989         | 5,154,862 | 4,763     | CU928145         |
| 7  | E. coli APECO1        | 5,497,653 | 4,428     | CP000468         |
| 8  | E. coli CFT073        | 5,231,428 | 5,339     | AE014075         |
| 9  | E. coli 0127:H6 E2348/69 | 5,069,678 | 4,554     | FM180568         |
| 10 | E. coli E24377A       | 5,249,288 | 4,749     | CP000800         |
| 11 | E. coli 0157:H7 EC4115| 5,704,171 | 5,315     | CP001164         |
| 12 | E. coli ED1a          | 5,209,548 | 4,915     | CU928162         |
| 13 | E. coli HS            | 4,643,538 | 4,378     | CP000802         |
| 14 | E. coli IAI1          | 4,700,560 | 4,353     | CU928160         |
| 15 | E. coli K12 MG1655    | 4,639,675 | 4,149     | U00096           |
| 16 | E. coli K12 W3110     | 4,646,332 | 4,226     | AP009048         |
| 17 | E. coli B str. REL606 | 4,629,812 | 4,205     | CP000819         |
| 18 | E. coli S88           | 5,032,268 | 4,696     | CU928161         |
Table 1. E. Coli data, taken from [7].

| 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       |

Table 2. Salmonella data. Taken from NCBI [8].

In Fig. 3 we present correlations of $P_{DSYN}$ with $D_1$ for (a) E Coli and for (b) Salmonella strains. In each of the two data sets we have looked into all pairs of strains. The data are presented for $k=7$. No significant correlation was found between strains of the two different species.
Fig. 3. Correlation of $P_{DSYN}$ with $D_1$ (k=7) for pairs of (a) E Coli strains and (b) Salmonella strains. Arrows indicate the two principal components, delineating the variance of the data.

Next we turn to correlations of over-all synteny with $D_2$. This is presented in Fig. 4 for k=7. Once again we note the strong correlations in the data.

![Figure 4](image)

Fig. 4. Correlation of $P_{SYN}$ with $D_2$ (k=7) for pairs of (a) E Coli strains and (b) Salmonella strains.

In Fig. 5 we display the Pearson correlations of $D_1$ and $D_2$ for E Coli pairs of strains, as function of k, for the two classes of synteny measures. We observe that both $D_1$ and $D_2$ are negatively correlated with $P_{SYN}$, as expected, but we find different correlations of the two measures with $P_{DSYN}$. Whereas $D_1$ displays the expected negative correlation, $D_2$ is less sensitive to the direct synteny measure. Thus $D_1$ correlates strongly with both $P_{DSYN}$ and $P_{SYN}$ for $k \leq 7$.

![Figure 5](image)

Fig. 5. Pearson correlations of the two k-mer distance measures of pairs of E Coli strains, as function of k, with (a) $P_{SYN}$ and (b) $P_{DSYN}$.

**Proximities between human chromosomes**

As another example of the use of k-mer distances, we explore the proximity of human chromosomes to each other. The results, for masked chromosomes of HG38 are displayed in Fig. 6. The lowest results are highlighted.
In order to decide which values should be regarded as low, we compare each chromosome with itself using two different versions of the human chromosomes: HG19 and HG38 [8].

![Image of distance matrix](image.png)

Fig. 6  $D_2$ distances (k=10) between masked human chromosomes of HG38.

|     | 1  | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9  | 10 | 11 | 12 |
|-----|----|----|----|----|----|----|----|----|----|----|----|----|
| 1   | .020 | .016 | .017 | .017 | .018 | .021 | .018 | .024 | .022 | .018 | .019 |
| 2   | .022 | .022 | .023 | .033 | .026 | .021 | .026 | .029 | .060 | .049 | .022 | .056 |

Table 3. $D_2$ distances between masked human chromosomes of HG19 and HG36 (lower rows). Upper rows are chromosome numbers.

The first ten chromosomes are the longest human chromosomes. The largest $D_2$ distance, for k=10, between the two versions is 0.024, as can be seen from Table 3. This may be viewed as a margin of error, describing the level of inaccuracy of the biological analysis. Looking at the $D_2$ distances displayed in Fig. 6, we see that none approaches this low value, but some reach values of the same order of magnitude. The closest pair, chromosomes 5 and 6, display consistent proximity for a large range of k-values, often below the error level read off the different versions of the ten leading chromosomes. The comparison between the two is displayed in Fig. 7. Other close pairs of chromosomes highlighted in Fig. 6 do not display such consistent behavior. Hence we conclude that chromosomes 5 and 6 lead in their strong evolutionary proximity.
Fig. 7. Comparison of the maximal $D_2$ distance of the first ten chromosomes between two versions of the human genome, with the $D_2$ distance between masked chromosomes 5 and 6, as function of $k$.

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 $D_1$ and direct syntenic regions and a strong correlation between $D_2$ and both direct and inverse syntenies, which indicates evolutionary similarity between two species. We argue therefore that k-mer ratios are validated as good measures for evolutionary distances.

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 calculations 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 KL limit), our method competes well with other methods. Applying k-mer distance evaluations to human chromosomes we argue that chromosomes 5 and 6 display very small evolutionary distances between each other.

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 are distant from each other, or identify sequences that have a large amount of mutual syntenies.

References:

1. Shporer S, Chor B, Rosset S and Horn D. Inversion symmetry of DNA k-mer counts: validity and deviations. BMC Genomics 2016, 17:696.
2. Rudner R, Karkas JD, Chargaff E. Separation of B. subtilis DNA into reverseary strands. III. Direct Analysis. Proc Natl Acad Sci USA 1968, 60:921-922.
3. Baisnee P-F, Hampson S, Baldi P. Why are reverseary DNA strands symmetric? Bioinformatics. 2002; 18:1021–33.
4. Pevzner P and Tesler G. Genome Rearrangements in Mammalian Evolution: Lessons from Human and Mouse Genomes. Genome Research (2003), 13:37–45.
5. Pham S K and Pevzner P A. DRMM-Synteny: decomposing genomes into evolutionary conserved segments. Bioinformatics (2010) 26, 2509-2516.
6. Ying Hu et al. OmicCircos: a simple-to-use R package for the circular visualization of multidimensional omics data". Cancer informatics (2014) 13: 13–20.
7. Oksana Lukjancenko, Trudy M Wassenaar, and David W Ussery. Comparison of 61 sequenced Escherichia coli genomes. In: Microbial ecology 60.4 (2010).
8. NCBI taxonomy browser at https://www.ncbi.nlm.nih.gov

**Declarations**

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Availability of data and materials.
Data sharing is not applicable to this article as no datasets were generated or analysed during the current study.

Competing interests
The authors declare they have no competing interests.

Funding
This research was partially supported by the research fund of the Blavatnik School of Computer Science.

Authors’ contributions
BC and DH initiated the study and contributed to its design.
AK carried out the numerical data analysis.
DH prepared the manuscript.
All authors read and approved the final manuscript.

Acknowledgements
We thank Uri Gophna for helpful discussions.
