An Algorithm to Calculate the p-value of the Monge-Elkan Distance

Petr Ryšavý[1*], Filip Železný[1]

[1]Department of Computer Science,

Faculty of Electrical Engineering,

Czech Technical University in Prague,

Technická 2, Prague, 166 27, Czech Republic

[*]To whom correspondence should be addressed;

E-mail: petr.rysavy@fel.cvut.cz.

June 9, 2025

Keywords: Monge-Elkan distance, p-value, null distribution

Abstract: The Monge-Elkan distance is a straightforward yet popular distance measure used to estimate the mutual similarity of two sets of objects. It was initially proposed in the field of databases, and it found broad usage in other fields. Nowadays, it is especially relevant to the analysis of new-generation sequencing data as it represents a measure of dissimilarity between genomes of two distinct organisms, particularly when applied to unassembled reads. This paper provides an algorithm to calculate the p-value associated with the Monge-Elkan distance. Given the object-level null distribution,

i.e., the distribution of distances between independently and identically sampled objects such as reads, the method yields the null distribution of the Monge-Elkan distance, which in turn allows for calculating the p-value. We also demonstrate an application on sequencing data, where individual reads are compared by the Levenshtein distance.

Final publication is available from Mary Ann Liebert, Inc.: `https://doi.org/10.1089/cmb.2024.0854`.

1 Introduction

The Monge-Elkan similarity was proposed by (Monge and Elkan, 1996). The paper used the concept to solve the *field matching problem*, i.e., the problem of deciding whether two different fields (say, strings) represent the same entity. For example, *John Doe* and *Doe, John* are likely to represent the same person despite different textual representations. The field matching problem often arises in databases when multiple data sources are combined into a single one.

The paper proposed a simple yet effective recursive algorithm to compare the fields. Denote the fields R_A and R_B (see Supplementary Material Table 1 for a summary of notation). Each of the fields is broken into subfields (say, at positions of spaces), and for each subfield a of R_A, the most similar subfield in R_B is assigned. The similarities between such established pairs of subfields are then averaged to produce the value of similarity between the fields.

Due to the simplicity of the idea, it has found its way into many applications, especially in the approximate string matching field and related tasks. The Monge-Elkan distance can be found in a review paper (Cohen et al., 2003) focusing on name matching. The work (Kaplar et al., 2019) evaluates the Monge-Elkan distance in the context information extraction from electronic health

records in the Serbian language. Another review paper, (Gali et al., 2016) finds usage in title matching, where users input text into fields. Further usages include business process model matching in (Abdelkader, 2018), where the goal is to find correspondence between activities, and the Monge-Elkan distance is used together with WordNet. Toponym matching in geography is another field where the Monge-Elkan distance has been applied (Santos et al., 2018). In this work, the goal is to identify whether two names refer to the same place. There are also generalizations of the original paper as in (Jimenez et al., 2009), which uses generalized arithmetic mean instead of the average.

The Monge-Elkan similarity found its way into bioinformatics and related fields as well. The Monge-Elkan distance was used as a baseline in the identification of duplicate biological entities in bioinformatics databases in (Song and Rudniy, 2010), where the authors used an edit distance based on Markov random fields. A similar set of authors followed with work (Rudniy et al., 2014), where the Monge-Elkan distance is evaluated among the set of alternative approaches. The work (Yamaguchi et al., 2012) used the Monge-Elkan distance in the context of biomedical abbreviation clustering. From the biological databases perspective, the ontology alignment problem is an important task as well, as many biological data are used in the form of ontologies. (Stoilos et al., 2005; Cheatham and Hitzler, 2013) use the Monge-Elkan distance as one of the considered approaches.

(Ryšavý and Železný, 2016, 2019, 2023) used the Monge-Elkan similarity to develop a distance measure applied in analyzing raw read data. Instead of fields in database entries, R_A and R_B represented *bags of reads* (i.e., *multisets*) produced by a sequencing machine for two sequences A and B, so the bag elements are individual reads. The Monge-Elkan distance then approximates the Levenshtein distance by (Levenshtein, 1966) between the sequences A and

B without the need for read data assembly.

A main challenge in the outlined biological use case is the interpretation of the computed distance values. Relative comparisons are not problematic: knowing that the read bag distance is, say, 10 means that the original sequences are more similar than if the distance was 20. However, we also want to be able to establish whether a particular computed distance between two read bags indicates similar input sequences rather than random unrelated sequences. As customary in life sciences, we make the former call if the probability of the latter case, i.e., the *p-value*, is smaller than a threshold (e.g., 0.05). Determining the p-value of the Monge-Elkan distance is, however, not trivial.

Being able to calculate the p-value of the result has a broader impact as its knowledge is necessary to assess the statistical significance of the result. By definition, a low p-value means that the result measured is not a random fluke and is more likely to be reproduced under similar conditions. When the null distribution of the used statistic is not known, the user can resort to Barnard's Monte Carlo sampling (Marriott, 1979), where the statistic values are calculated from randomly sampled inputs. While this procedure gives an estimate of the p-value, in many cases, it is computationally very demanding. In such cases, it may not be feasible to calculate the p-value of rare events, which calls for more efficient approaches.

The main contribution of this paper is a method for calculating the p-value for the Monge-Elkan distance. First, we will provide an algorithm for the general version, then apply it to the approach presented by (Ryšavý and Železný, 2016, 2019), and then we will discuss drawbacks, limitations, and necessary approximations involved in the calculation. A minor limitation of the presented algorithm is that the similarity between the subfields (reads) must have only a finite set of outcomes, which is true under commonly used measures with

limited subfield length. This includes the Levenshtein distance (Levenshtein, 1966), the longest-common-subsequence similarity (Wagner and Fischer, 1974), Jaro-Winkler similarity (Winkler, 1990), and others. The restriction to finite domains, together with independence assumptions allows to exploit *probability generating functions* well-known in statistics. As a result, the algorithm is able to calculate the p-value without the need for the standard Barnard's Monte-Carlo sampling in (Marriott, 1979) potentially infeasible for large bags R_A, R_B.

2 Definition of the Problem

Let U be a universe of elements. We assume a distance function dst defined on pairs of elements of U. We also assume that this distance function has a finite range, as is the case for distances such as the Levenshtein distance. The Monge-Elkan distance between two read bags is the average distance from each element of the first read bag to its closest counterpart in the second read bag.

Definition 1 (Monge-Elkan distance, (Monge and Elkan, 1996))**.** *Let R_A, R_B be two bags sampled with replacement from universe U. Then, the Monge-Elkan distance between R_A and R_B, denoted $\mathsf{Dst}(R_A, R_B)$, is defined as*

$$\mathsf{Dst}(R_A, R_B) = \frac{1}{|R_A|} \sum_{a \in R_A} \min_{b \in R_B} \mathsf{dst}(a, b), \tag{1}$$

where $\mathsf{dst} : U \times U \mapsto Y$ is a distance function on U with finite range $Y \subset \mathbb{R}$.

For fixed bag sizes $|R_A|$ and $|R_B|$, the above quantity has the *null distribution* under the *null hypothesis* that all elements in $R_A \cup R_B$ are sampled i.i.d. with replacement from U. The *alternative hypothesis* states that the bags are similar. Given a value of $\mathsf{Dst}(R_A, R_B)$, we reject the null hypothesis if the probability of obtaining that value under the null hypothesis is smaller than a threshold.

Being able to quantify the null distribution of (1) is crucial in order to calculate the p-value, which is the probability that a more extreme distance value is observed. Formally, let d be the Monge-Elkan for two observed read bags. Then the p-value is the probability that $\mathsf{Dst}(R_A, R_B) \leq d$ for two bags R_A, R_B sampled i.i.d. with replacement from U. Therefore, to get the p-value from the null distribution, we need only to sum over all smaller distance values.

As mentioned in the introduction, the original paper (Monge and Elkan, 1996) defined the Monge-Elkan similarity rather than the distance. We will provide the algorithm for the distance version; however, the ideas may be straightforwardly translated into a similarity version by swapping the min and max operators and the sides of the appropriate inequalities.

3 p-value Calculation

First, we can notice that the possible values calculated in (1) are dependent on the range of dst. Therefore, the null distribution of the Monge-Elkan distance is dependent on the chosen universe U and distance function dst. As a result, we cannot precompute a generally applicable null distribution of the Monge-Elkan distance. A standard approach in such a case would consist in a Monte-Carlo random sampling (Marriott, 1979) of many pairs of bags R_A, and R_B and a consequent evaluation of $\mathsf{Dst}(R_A, R_B)$ allowing the approximation of the null distribution. In this paper, we will use *probability generating functions* to evaluate the null distribution. We will break Formula (1) into smaller pieces for which we will calculate the null distribution and then combine these null distributions into the null distribution of Dst.

3.1 Null Distribution of the min Operation

The innermost part of the calculation in (1) is the dst distance. The null distribution of dst is assumed to be given. We also assume that the distance has a finite range as defined in Section 2. The following theorem allows calculating the null distribution of the minimum operation.

Theorem 1. *Let $(\Omega, \leq)$ be a finite, totally ordered set. Let $p : \Omega \mapsto [0, 1]$ be a probability distribution of a discrete random variable. Suppose that $S \subseteq \Omega$ is a bag of i.i.d. samples drawn with replacement from probability distribution p. Then for any $\omega \in \Omega$, function $q : \Omega \mapsto [0, 1]$ defined as*

$$q(\omega) = \sum_{i=1}^{|S|} \binom{|S|}{i} \cdot p(\omega)^i \cdot \left(\sum_{\{\omega' \in \Omega \,:\, \omega' > \omega\}} p(\omega') \right)^{|S|-i} \tag{2}$$

is the probability of $\min S = \omega$. Should the last term of the multiplication be in the form of 0^0, then this value is considered 1.

Proof. The proof can be found in Supplementary Material, Sect. 2. The proof is done by counting for how many elements in S holds that $\min S = \omega$. □

Corollary 1. *Suppose that the values of $\mathsf{dst}(a, b)$ are random i.i.d. for fixed a. Then,*

$$\begin{aligned} P\left(\min_{b \in R_B} \mathsf{dst}(a, b) = d \right) &= \sum_{i=1}^{|R_B|} \binom{|R_B|}{i} \cdot P(\mathsf{dst}(a, b) = d)^i \cdot P(\mathsf{dst}(a, b) > d)^{|R_B|-i} \\ &= P(\mathsf{dst}(a, b) \geq d)^{|R_B|} - P(\mathsf{dst}(a, b) > d)^{|R_B|}. \end{aligned} \tag{3}$$

We have to be careful when applying Corollary 1. The distribution of the distance is calculated for a fixed a. Consider the situation when the universe $U = \{0, 0.1, 0.2, \ldots, 1.0\}$ and the distance metric is the absolute difference between two numbers. Then, the null distribution looks different in situations when $a =$

0.5 and $a = 0$. In the first case, the maximum distance is 0.5, and in the second case, it is 1.0. However, if we align the points on a circle so that the coordinates go from 0.0 to 1.0, i.e., the distance is defined as $\min\{|a-b|, 1-|a-b|\}$, then the null distribution of the distance looks the same for any a, and all we need is to have i.i.d. uniformly selected elements in R_B. Another example of a distance that has the same null distribution for any selection of a is the Hamming distance (Hamming, 1950).

The Levenshtein distance (Levenshtein, 1966) is another example where the null distribution depends on a. For string AA, there is only one string of distance 1 if A is inserted, but for string CG, there are three strings of distance 1 if A is inserted. The overall effect of this repetition-based discrepancy needs to be assessed experimentally; nevertheless, with larger alphabet sizes and longer sequences, the effect of repeated symbols will get smaller.

3.2 Null Distribution of the Sum

Once we have the null distribution for the minimum operation, we need to evaluate the null distribution of the sum. To this end, we will exploit a *probability generating function*, which is a well-established concept in probability theory. The reader is referred to (Feller, 1966) for more details.

Definition 2 (Probability generating function)**.** *Let Ω be a finite subset of $\mathbb{R}$. Let $p : \Omega \mapsto [0,1]$ be a probability distribution of a discrete random variable. Then, the* probability generating function *is*

$$\text{genpoly}_p(x) = \sum_{\omega \in \Omega} p(\omega) \cdot x^{\omega}. \tag{4}$$

Usually, the probability generating function is defined only when Ω is a subset of non-negative integers. However, for our application, any finite subset

of real numbers is admissible.

Let us illustrate the usage of the probability-generating functions through a simple example. Consider tossing a biased dice that is able to produce three outcomes – 1, 2, and 3 with probabilities of $\frac{1}{6}$, $\frac{1}{3}$, and $\frac{1}{2}$, respectively. Then, the generating polynomial is

$$\frac{1}{6}x^1 + \frac{1}{3}x^2 + \frac{1}{2}x^3. \tag{5}$$

We immediately see the probability of an outcome as a coefficient of the respective power of x. When two dice are tossed, the sum of 4 can be reached by having 1 and 3 with the probability of $\frac{1}{6} \cdot \frac{1}{2} = \frac{1}{12}$, both dice showing 2 with probability $\frac{1}{9}$, or, finally, 3 and 1 with probability $\frac{1}{12}$. The probability of seeing 4 is, therefore, $2 \cdot \frac{1}{12} + \frac{1}{9} = \frac{10}{36}$. The previous calculation is exactly what happens to the coefficients if we multiply two generating polynomials. Consider the second power of the polynomial in (5)

$$\frac{1}{36}x^2 + \frac{4}{36}x^3 + \frac{10}{36}x^4 + \frac{12}{36}x^5 + \frac{9}{36}x^6. \tag{6}$$

We can notice that the coefficient of x^4 is $\frac{10}{36}$. In summary, the probability generating function allows an easy calculation of the null distribution for a sum of independent variables.

Having this intuition at hand, we can define the generating polynomial for $a \in R_A$ in the calculation of $\min_{b \in R_B} \mathsf{dst}(a, b)$ as

$$\text{genpoly}_a(x) = \sum_{d \in Y} P\left(\min_{b \in R_B} \mathsf{dst}(a, b) = d\right) \cdot x^d, \tag{7}$$

where Y is the range of distance $\mathsf{dst} : U \times U \mapsto Y$. The polynomial for the

whole distance in (1) is then the respective power of $\text{genpoly}_a(x)$

$$\text{genpoly}_{R_A}(x) = (\text{genpoly}_a(x))^{|R_A|}. \tag{8}$$

This polynomial can then be used to calculate the null distribution of the Monge-Elkan distance under an independence assumption discussed later.

Theorem 2. *Assume that the probability* $P(\min_{b \in R_B} \mathsf{dst}(a, b) = d)$ *is independent of the choice* R_B*. Then the coefficients of polynomial* $\text{genpoly}_{R_A}(x)$ *represent the null distribution of the Monge-Elkan distance up to a multiplicative term of* $\frac{1}{|R_A|}$*. In other words, if* a_d *is the coefficient of* x^d *in the polynomial* genpoly_{R_A}*, then* $P\left(\mathsf{Dst}(R_A, R_B) = \frac{1}{|R_A|}d\right) = a_d$*.*

Proof. The proof can be found in Supplementary Material, Sect. 2. The main idea of the proof is to calculate the coefficients of $\text{genpoly}_{R_A}(x)$ and to compare them with the null distribution of the Monge-Elkan distance. □

Corollary 2. *Suppose that*

$$\text{genpoly}_{R_A}(x) = a_0 x^{d_0} + a_1 x^{d_1} + a_2 x^{d_2} + \cdots + a_n x^{d_n}. \tag{9}$$

Then the p-value of $\mathsf{Dst}(R_A, R_B) = d$ *is equal to*

$$P(\mathsf{Dst} \leq d) = \sum_{\{d_i \,:\, d_i \leq |R_A| d\}} P\left(\mathsf{Dst}(R_A, R_B) = \frac{1}{|R_A|} d_i\right) = \sum_{\{d_i \,:\, d_i \leq |R_A| d\}} a_i. \tag{10}$$

3.3 A Numerical Example

Here, we present an example calculation of the p-value. Let distance dst be the Hamming distance over universe of elements $U = \{\mathsf{AA}, \mathsf{AT}, \mathsf{AC}, \mathsf{CC}, \mathsf{CG}\}$. Let $R_A = \{\mathsf{AA}, \mathsf{AC}, \mathsf{AC}\}$ and $R_B = \{\mathsf{AC}, \mathsf{CG}\}$. The Monge-Elkan distance is

$\mathsf{Dst}(R_A, R_B) = \frac{1}{3}(1+0+0)$.

Hamming distance simply calculates the number of mutations; its null distribution on the universe is $(0.2, 0.4, 0.4)$, i.e., distance of 0 can be encountered with 20 % probability, distance of 1 with 40 % probability, and distance of 2 with 40 % probability. By applying Corollary 1, we get null distribution of the minimum as $\frac{1}{25}(9, 12, 4)$. For example, the minimum of 0 when selecting two elements from $\{0, 1, 2\}$ (i.e., the possible dst values) can be obtained from combinations $\{0, 0\}$ (0.04 probability), $\{0, 1\}$ (0.16 probability), and $\{0, 2\}$ (0.16 probability), which is overall with $0.36 = \frac{9}{25}$ probability.

By (8), we get

$$\begin{aligned}\mathrm{genpoly}_{R_A}(x) &= \left(\frac{1}{25}(9x^0 + 12x^1 + 4x^2)\right)^3 \\ &= \frac{1}{15625}(729x^0 + 2916x^1 + 4860x^2 + 4320x^3 + 2160x^4 + 576x^5 + 64x^6). \quad (11)\end{aligned}$$

By Corollary 2, we see that the p-value of this particular choice of R_A and R_B is equal to

$$P\left(\mathsf{Dst} \leq \frac{1}{3}\right) = \frac{1}{15625}(729 + 2916). \quad (12)$$

3.4 The Independence Assumption

We have to look in more detail at the independence assumption in Theorem 2. The fact that the minimum selection should be independent of the set over which we select the minimum is very restrictive and can be satisfied exactly only for very trivial distances dst. The calculation in Theorem 2 corresponds to the situation when for the first element $a \in R_A$, we generate bag R_B and calculate the summand according to the Monge-Elkan distance (1). Then, for the second element $a \in R_A$, a *new* set R_B is generated independently, and the summand is calculated. The calculation then follows with the new set R_B for each $a \in R_A$.

However, in the Monge-Elkan distance, the set R_B is kept throughout the calculation. Hence, we are making some error in the p-value calculation. Consider a space defined by distance function dst. The space looks different in the case where elements in R_B are selected randomly uniformly and in the case where all elements in bag R_B are the same. The first case is, however, common, while the second one is unlikely. Therefore, there are still many situations when Theorem 2 will be applicable as a reasonable approximation.

3.5 Asymptotic Complexity

To evaluate the asymptotic time complexity, the starting point is the size of range Y of distance dst. Theorem 1 is applied to calculate the null distribution of the minimum operation. To evaluate the combinatorial coefficients, $\mathcal{O}(|R_B|^2)$ operations are needed. As both probabilities in (3) can be evaluated in constant time (one by a constant time lookup, the second by using a prefix-sum array), the distribution with the same range can be calculated in $\mathcal{O}(|Y|+|R_B|^2)$ operations.

Calculation of the null distribution using Theorem 2 requires raising a polynomial of degree $|Y|$ to the power of $|R_A|$. Using a naive implementation, this requires at most $|Y|^{|R_A|}$ multiplications as this is the maximum theoretical possible number of terms in the polynomial multiplication. However, we are calculating the power of a polynomial that requires much less work - there will be $\binom{|R_A|+|Y|-1}{|R_A|} = N$ terms in the polynomial.

We can use the algorithm for fast power calculation of large numbers (Knuth, 1981). To calculate the n-th power of a number x, we can compute powers of $x^1, x^2, x^4, \ldots, x^{\lfloor \log_2 n \rfloor}$ and multiply them together, requiring only $2 \cdot \log n$ multiplications. In our case, this requires only $\mathcal{O}(\log |R_A|)$ polynomial multiplications. Using the Fast Fourier transform algorithm for polynomial multiplication (Cantor and Kaltofen, 1991), one operation will be at most $\mathcal{O}\left(N \log(N)\right)$. Over-

all, the runtime complexity is in

$$\mathcal{O}\left(\log(|R_A|)N\log(N)+|Y|+|R_B|^2\right) = \mathcal{O}\left(\log|R_A|\cdot\binom{|R_A|+|Y|-1}{|R_A|}\cdot\log\binom{|R_A|+|Y|-1}{|R_A|}+|Y|+|R_B|^2\right). \tag{13}$$

This bound is general as it assumes that no two pairs of numbers from Y^2 have the same sum unless the pairs are only permuted. This, however, is usually not true. For example, for the Levenshtein distance (Levenshtein, 1966), $Y = \{0, 1, \ldots, |Y| + 1\}$. In this situation, $4 + 5$ yield the same sum as $1 + 8$, $2 + 7$, and $3 + 6$. In such a case, the number of polynomial terms grows linearly instead of exponentially. The runtime complexity is, then,

$$\mathcal{O}\left(\log|R_A| \cdot |R_A||Y| \cdot \log(|R_A||Y|) + |Y| + |R_B|^2\right). \tag{14}$$

4 Application to Read Data

(Ryšavý and Železný, 2016) proposed using the Monge-Elkan distance to estimate the similarity between sequencing data without the need for sequence assembly. The idea was further extended in (Ryšavý and Železný, 2019), applied to contig data in (Ryšavý and Železný, 2017), or a combination of the latter in (Ryšavý and Železný, 2023). The papers have shown that the Monge-Elkan distance with some modifications is a good approximation of the Levenshtein distance between the original genomic sequences, being a good compromise between the traditional sequence alignment and the alignment-free methods (Zielezinski et al., 2017). Here, we will briefly provide the formula for the symmetric version of the Monge-Elkan distance, and modify the p-value algorithm to be applicable to the symmetric version.

In the case of read data, the universe becomes the set of all sequences over

alphabet $\Sigma = \{A, C, G, T\}$ of a selected length l. Length l acts as a *read length* - the length of short fragments that are sampled from the DNA sequence. Due to the nature of the sequencing process, those fragments, called reads, have lengths of tens to hundreds of symbols. Unfortunately, the location of the reads within the original sequence is not available, and *in-silico* reconstruction of the original sequences is usually needed. This reconstruction requires assembly guided by prefix-suffix overlaps between the reads.

Further, bags R_A, and R_B represent the bags of reads. Distance dst in our case is the Levenshtein distance (Levenshtein, 1966), which counts the number of insertions, deletions, and substitutions, i.e., the basic evolutionary events. Further, the Levenshtein distance is replaced by a slightly modified version in (Ryšavý and Železný, 2016) that accounts for random locations of the reads by different gap penalty at margins. The usage of the Monge-Elkan distance directly on the read data then avoids the NP-hard assembly problem. The distance is then made symmetric and rescaled, resulting in

$$\underset{\mathsf{MESG}}{\mathsf{Dst}}(R_A, R_B) \;=\; \frac{1}{2}\max\{|R_A|, |R_B|\} \cdot (\mathsf{Dst}(R_A, R_B) + \mathsf{Dst}(R_B, R_A)) \quad (15)$$

The reasoning behind Formula (15) is the following. The alignment between sequences A and B maps similar subsequences together. Therefore, read a should be aligned with the most similar read in R_B. This is done by the Monge-Elkan distance. The results are averaged for all reads in R_A. The distance is then made symmetric in (15). The multiplicative terms at the beginning bring the distance to the proper scale as the Monge-Elkan distance is an average of values no more than read length l while the maximum distance between the original sequences is proportional to $\max\{|R_A|, |R_B|\}$.

4.1 A Modification of the p-value Algorithm

From the p-value calculation perspective, the multiplicative terms in (15) are irrelevant as they only apply a linear scale on the null distribution. Therefore, calculating the p-value for the distance in Formula (15) is equivalent to calculating the p-value for

$$|R_B| \sum_{a \in R_A} \min_{b \in R_B} \mathsf{dst}(a,b) + |R_A| \sum_{b \in R_B} \min_{a \in R_A} \mathsf{dst}(a,b). \tag{16}$$

To derive the formula above, consider multiplying the average of $\mathsf{Dst}(R_A, R_B)$ and $\mathsf{Dst}(R_B, R_A)$ by $2|R_A||R_B|$.

Again, we will assume the independence assumption as discussed above. Instead of a single sum, we will have an addition of two sums. Therefore, we are able to multiply the generating polynomials (under the independence assumption) to obtain the generating polynomial for the new Formula (16).

The last thing to solve is the different weights of both sums in Formula (16). If we multiply each element in range Y of distance dst, then the null distribution of the product is scaled as well. Following the definition of the probability generating function in Definition 2, the multiplication is equivalent to multiplying each exponent in the power of x. Therefore, the resulting generating polynomial is

$$\text{genpoly}_{\mathsf{MESG}}(x) \;=\; \left(\text{genpoly}_a(x^{|R_B|})\right)^{|R_A|} \cdot \left(\text{genpoly}_b(x^{|R_A|})\right)^{|R_B|}. \tag{17}$$

In Formula (17), polynomial genpoly_b is defined similarly to genpoly_a, except the roles of R_A and R_B are swapped. The coefficients of polynomial $\text{genpoly}_{\mathsf{MESG}}$ can then be used to estimate the p-value in the same manner as in Theorem 2.

4.2 A Numerical Example

In this section, we follow the example from Sect. 3.3.[1] A shortest-superstring assembly of the reads would be AAC, and ACG with the Hamming distance of 2.

If we evaluate the symmetric version of the Monge-Elkan distance according to (15), we see that the distance is equal to $\frac{3}{2}\left(\frac{1+0+0}{3}+\frac{1+0}{2}\right)=1.67$. However, in the p-value calculation, we will evaluate (16), which is 5 and does not contain terms that do not influence the p-value calculation. The null distribution of the minimum when selecting 3 reads from U is equal to $\frac{1}{125}(61,56,8)$. We have already seen genpoly_a in Sect. 3.3, genpoly_b is constructed symmetrically from the null distribution $\frac{1}{125}(61,56,8)$. Therefore,

$$\text{genpoly}_{\mathsf{MESG}}(x)=\left(\frac{1}{25}(9x^0+12x^2+4x^4)\right)^3\cdot\left(\frac{1}{125}(61x^0+56x^3+8x^6)\right)^2. \tag{18}$$

By the expansion, we get

$$\begin{aligned}\text{genpoly}_{\mathsf{MESG}}(x)=\frac{1}{25^3\cdot 125^2}(&9^3\cdot 61^2x^0+3\cdot 9\cdot 9\cdot 12\cdot 61^2x^2+\\&+9^3\cdot 2\cdot 61\cdot 56x^3+18084060x^4+19922112x^5+19072368x^6+\cdots).\end{aligned} \tag{19}$$

Since the distance was equal to 5, the p-value is equal to the sum of the coefficients up to term x^5, which is $\frac{1}{25^3\cdot 125^2}(2712609+10850436+4980528+18084060+19922112)=0.23$.

[1] In a more realistic scenario, universe U would contain all 16 possible reads. However, to make the example continuation of the one from Sect. 3.3, we keep only the universe of 5 elements. Also, a more likely comparison of genomic sequences would use the Levenshtein distance, which is identical to the Hamming distance in the presented example.

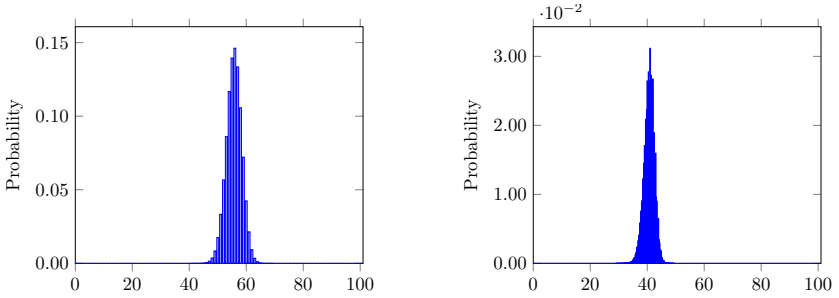
4.3 Runtime Complexity, Independence Assumption, and Null Distribution of dst

In the previous section, we multiplied the generating polynomials for two sums, assuming that the sums were independent. However, this is not true. In the first sum, we are averaging the row minima of a matrix; in the second sum, we are averaging the row maxima of the same matrix. Therefore, the generating polynomial provides only an approximation.

Regarding the computational complexity, the runtime is similar to (13), only we need not calculate the whole null distribution; only values smaller than d are required, where d is the distance calculated by (16). Therefore the p-value calculation is in

$$\mathcal{O}\left(\log(|R_A||R_B|) \cdot d \cdot \log d + l + |R_A|^2 + |R_B|^2\right). \tag{20}$$

The last component is missing, which is the null distribution of the read-read distance dst. In the case where dst is the Levenshtein distance (Levenshtein, 1966), the null distribution is unknown. We were only able to find previous research that has shown that the null distribution for a related problem of the Longest Common Subsequence (Chvátal and Sankoff, 1975) problem follows the Tracy-Widom distribution in (Majumdar and Nechaev, 2005). Therefore, the empirical evaluation of the null distribution for the Levenshtein distance and its modifications that were proposed in (Ryšavý and Železný, 2016) is needed. Figure 1 shows null distributions for the Levenshtein distance, together with the effect of min operation in Theorem 1.



(a) Levenshtein distance dst. (b) Minimum of the Levenshtein distance $\min_{b \in R_B} \mathsf{dst}(\cdot, b)$

Figure 1: Here, we illustrate the empirical null distribution of the underlying distance function dst, which is not known exactly in many cases (including the Levenshtein distance). The figure on the left shows the null distribution of distance dst, and the figure on the right shows the effect of Theorem 1 on the distribution. We see that after the calculation of the minimum, the distribution is asymmetric, unlike the original one - the smaller distances are preferred, and the minimum being higher than 50 is unlikely but possible. The distribution was calculated empirically for 10^8 trials assuming random sequences of length $l = 100$. Nevertheless, distances smaller than 38 and larger than 70 were never registered as their probabilities are very low. The zero probabilities in the calculation could be dealt with using the Laplace smoothing. The minimum distributions were calculated using Theorem 1 for read bag size of 1,000.

5 Experimental Baselines for Comparison

In the experimental evaluation in Section 6, we will need to compare the algorithm from Section 3 with unrelated but alternative ways to evaluate the p-value. Here, we provide two competing approaches stemming from well-known mathematical theorems - the Central Limit Theorem (CLT) and the Bernstein's inequality (BI) (Bernstein, 1924). Both baselines allow us to estimate the null distribution of the Monge-Elkan distance under the common assumption with our method that the null distribution of distance dst is known. As a result, we will be able to compare the null distributions of the Monge-Elkan distance calculated by the three alternatives in terms of the Kullback-Leibner divergence.

The CLT approximation will use the property that the sum of numbers sampled from the same distribution will converge to a normal distribution with

known parameters. In our case, we have $|R_A|$ distances, the sum of which can be approximated by CLT. In the case of BI, we can bound the p-value by an upper bound.

5.1 The Central Limit Theorem Approximation

Formula (13) provides the complexity for the general case when the time grows exponentially. We will provide an alternative approximation of the p-value in the case when $|R_A|$ is too large to evaluate Formula (10), and range Y does not allow the complexity of (14).

The generating polynomial approach in Sect. 3.2 offers an exact method to calculate the null distribution, assuming the independence of the min calculations on the choice of R_B. This assumption might be applied in the Lindeberg–Lévy Central Limit Theorem (CLT), which can be reformulated as follows.

Theorem 3 (CLT)**.** *Assume that the probability* $P\left(\min_{b\in R_B} \mathsf{dst}(a,b) = d\right)$ *is independent of the choice* R_B*. For* $\min_{b\in R_B} \mathsf{dst}(a,b)$*, denote by* μ *its expected value, and* σ^2 *its variance. Then, as* $|R_A| \to \infty$*,*

$$\left[\mathsf{Dst}(R_A,R_B) = \frac{1}{|R_A|}\sum_{a\in R_A}\min_{b\in R_B}\mathsf{dst}(a,b)\right] \to \frac{1}{\sqrt{|R_A|}}\mathcal{N}(0,\sigma^2) + \mu, \qquad (21)$$

where $\mathcal{N}(0,\sigma^2)$ *is the normal distribution with mean* 0*, and variance* σ^2*.*

Corollary 3 (CLT approximation)**.** *For sufficiently large* $|R_A|$*, the p-value of* $\mathsf{Dst}(R_A,R_B) = d$ *is approximately equal to*

$$P(\mathsf{Dst}(R_A,R_B) \le d) \approx P\left(\mathcal{N}(0,\sigma^2) \le \sqrt{|R_A|}(d-\mu)\right). \qquad (22)$$

5.2 Bernstein Inequality as an Upper Bound

Here, we will apply a version of Bernstein's inequality (BI) (Bernstein, 1924), to calculate an upper bound on the p-value. The BI provides a bound on the probability that the sum of random variables deviates from its mean, similar to Hoeffding's inequality, (Hoeffding, 1963). In our setting, the inequality stands as follows. Note that the bound is usually formalized for the case when the sum is higher than the mean, but a symmetric version for smaller sums than expected also exists.

Theorem 4 (Bernstein Inequality)**.** *Let probability* $P\left(\min_{b\in R_B} \mathsf{dst}(a,b) = d\right)$ *be independent of the choice* R_B*. For* $\min_{b\in R_B} \mathsf{dst}(a,b)$*, denote by* μ *its expected value, and* σ^2 *its variance. Let* $M = \max_{y\in Y} |y|$*. Then, for positive* t*,*

$$P\left(\frac{1}{|R_A|}\sum_{a\in R_A} \min_{b\in R_B} \mathsf{dst}(a,b) \leq \mu - t\right) \leq \exp\left(-\frac{|R_A|t^2}{2\sigma^2 + \frac{2}{3}Mt}\right). \tag{23}$$

By substituting $d = \mu - t$, we obtain an upper bound on the p-value.

Corollary 4 (BI upper bound)**.** *Let* $\mu > d$*. Then, the p-value of* $\mathsf{Dst}(R_A, R_B) = d$ *is*

$$P(\mathsf{Dst}(R_A,R_B) \leq d) \leq \exp\left(-\frac{|R_A|(\mu-d)^2}{2\sigma^2 + \frac{2}{3}M(\mu-d)}\right). \tag{24}$$

6 Experimental Evaluation and Discussion

Given that the presented approach relies on an approximation assuming independence, which is not necessarily true, it is important to verify the validity of the null distribution obtained through Theorem 2 by comparing it with the actual data. To do so, we selected three simple examples of distance functions and evaluated them for various choices of bag sizes as well as different universes. The distances include:

- the Levenshtein distance (Levenshtein, 1966) on binary strings of length l;
- the Hamming distance (Hamming, 1950) on binary strings of length l (see Supplementary Material, Sect. 3);
- the absolute difference between two numbers from $\{0, 1, 2, \ldots, n-1\}$ aligned on a circle. Formally, $\min\{|a-b|, n-|a-b|\}$ where a and b are numbers in the respective set.

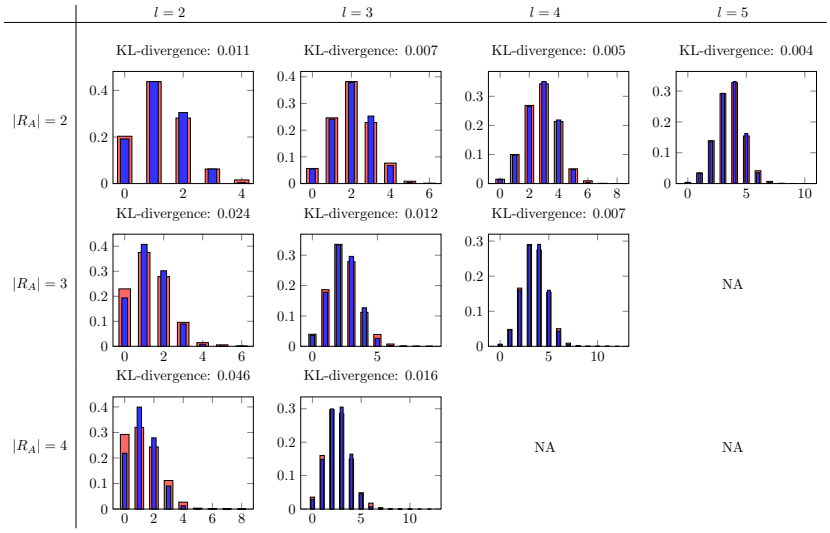
6.1 Comparison with the Exact Null Distribution

In the first experiment, the parameters l and n were selected so that the null distribution of the Monge-Elkan distance could be calculated by mere enumeration of all possible bags R_A and R_B. For simplicity of presentation, $|R_A|$ was set the same as $|R_B|$. The null distribution was then calculated by enumeration of all possible bags and using Theorem 2. Those two distributions were then compared visually as well as using the Kullback-Leibner divergence (Kullback and Leibler, 1951) (KL-divergence, sometimes called relative entropy). In the KL-divergence, the natural logarithm was used.

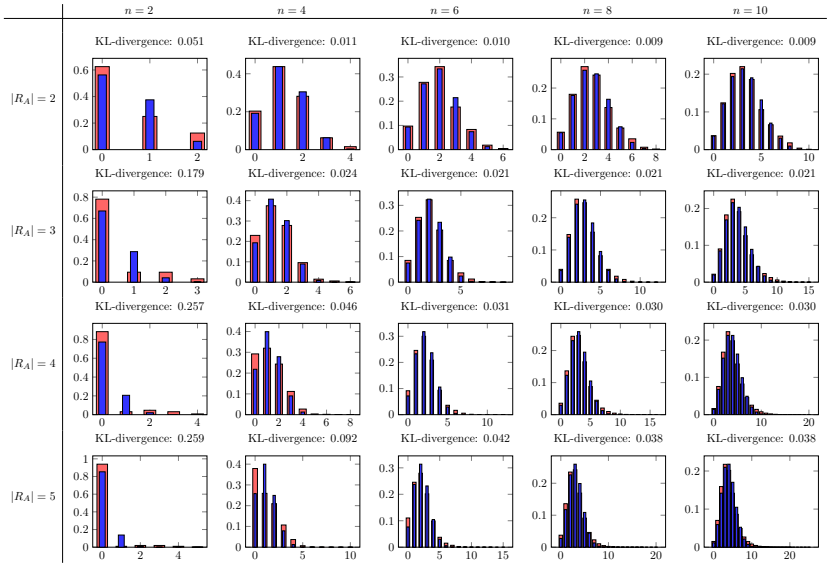
In the case of the string distances, the boundary set for enumeration was $2l|R_A| = 27$, which meant 2^{27} elements in the null distribution at most. In the case of the distance between the numbers, the limit was $n^{2|R_A|} = 10^{10}$, which meant 10^{10} elements in the null distribution at most.

The experimental evaluation is in Fig. 2 and Supplementary Material Fig. 1. From the figures, we can notice that the KL-divergence is growing with larger bag sizes (the independence assumption is more relied on in the multiplication). It might be expected that the KL-divergence would decrease with a universe of more elements (i.e., higher l or n), which is, however, supported only by the data in Fig. 2.

From the plots, we can also notice that the approximation underrepresents



(a) The Levenshtein distance over binary strings of length l.



(b) $\min\{|a-b|, n-|a-b|\}$ for $a, b \in \{0, 1, 2, \ldots, n-1\}$.

Figure 2: Comparison of the approximated (blue, narrow) and ground-truth (red, wide) null distribution. The approximated distribution was calculated using Theorem 2 while the ground-truth distribution was calculated enumerating all possible choices of R_A and R_B of the same size. NA means that with the given settings, it was not feasible to enumerate the distribution.

the low distances for more cardinal universes U. This is not a desired behavior; however, there remains an open window for future work in modifying the approach so that the p-value cannot be underestimated.

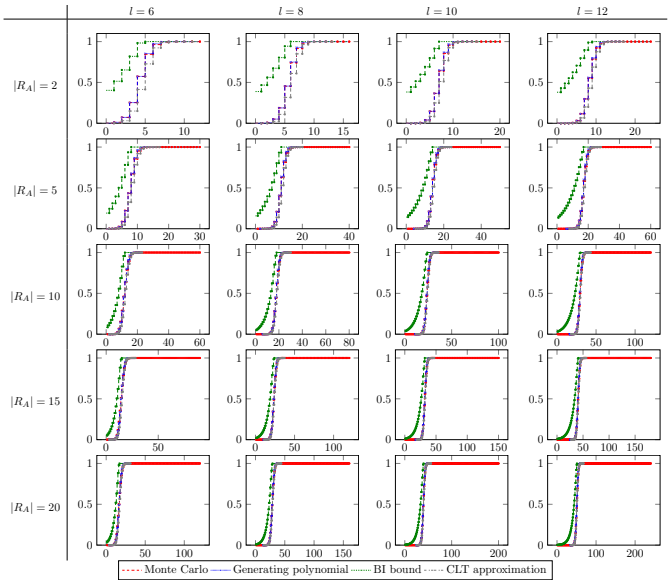
6.2 Comparison with an Empirical Distribution

Next, to test the methods under a wider range of conditions, we used an empirical distribution. We sampled 10^6 (uniformly, i.i.d.) samples together with uniform priors to avoid zero probability values (and thus undefined KL-divergence). This matches the standard Barnard's Monte Carlo sampling approach (Marriott, 1979) for p-value estimation. The comparison in the form of the cumulative distribution functions (CDFs) is in Fig. 3 and Supplementary Material Fig. 2. The KL-divergence between the methods is shown in Table 1 and Supplementary Material Table 2.

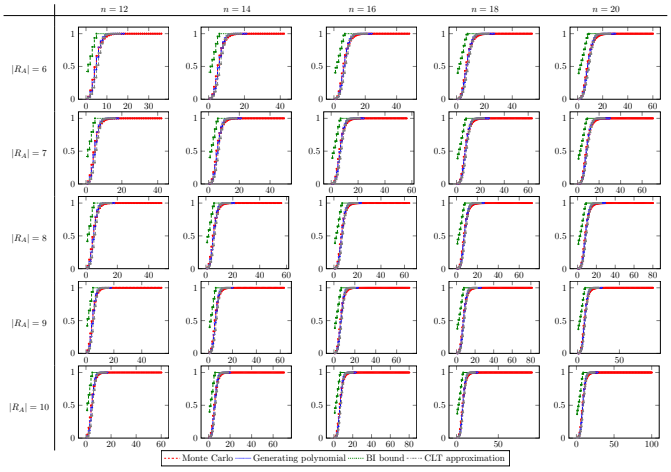
6.3 Comparison with the Baselines

To provide a further comparison, we included baselines presented in Sect. 5. We include those two baselines in Fig. 3 and Table 1. In both approaches described in Sect. 5, Corollary 1 was used to calculate the null distribution of the minimum operator. The reason for that is the fact that $|R_B|$ is too large to allow a full enumeration of the null distribution of the minimum operator.

From the experimental results, we can notice that the Bernstein inequality serves as an upper bound even in the case of approximations. Nevertheless, the BI upper bound is a poor approximation, with the KL divergence always above 0.1. In all cases, the generating polynomial method in Theorem 2 is capable of providing the best results. Nevertheless, we might notice that with larger $|R_A|$, the method provides worse results. On the contrary, the error of the central limit theorem approximation grows slower. Extrapolating this behavior, for



(a) The Levenshtein distance over binary strings of length l.



(b) $\min\{|a-b|, n-|a-b|\}$ for $a, b \in \{0, 1, 2, \ldots, n-1\}$.

Figure 3: Comparison of the approximated (blue) and empirical (red) null distribution CDFs. The approximated distribution was calculated using Theorem 2 while the ground-truth empirical distribution was calculated using Barnard's Monte Carlo sampling for 10^6 samples of R_A and R_B of the same size. The green null distribution represents an upper bound given by the Bernstein inequality (see Corollary 4), and the gray distribution is the one which uses the CLT instead of Theorem 2 (see Corollary 3). In both BI and CLT cases, Corollary 1 was used to calculate the null distribution of the minimum operator.

The Levenshtein distance (Levenshtein, 1966) over binary strings of length l.				
		Gen. poly. approximation	CLT approx.	BI upper bound
$\lvert R_A\rvert = 2$	$l = 6$	**0.004**	0.074	0.897
	$l = 8$	**0.004**	0.061	0.579
	$l = 10$	**0.004**	0.052	0.409
	$l = 12$	**0.004**	0.045	1.030
$\lvert R_A\rvert = 5$	$l = 6$	**0.005**	0.042	0.268
	$l = 8$	**0.005**	0.033	0.261
	$l = 10$	**0.006**	0.027	0.273
	$l = 12$	**0.007**	0.022	0.290
$\lvert R_A\rvert = 10$	$l = 6$	**0.004**	0.028	0.169
	$l = 8$	**0.003**	0.021	0.232
	$l = 10$	**0.004**	0.015	0.367
	$l = 12$	**0.005**	0.011	0.562
$\lvert R_A\rvert = 15$	$l = 6$	**0.006**	0.026	0.251
	$l = 8$	**0.003**	0.015	0.317
	$l = 10$	**0.003**	0.011	0.129
	$l = 12$	**0.005**	0.007	0.446
$\lvert R_A\rvert = 20$	$l = 6$	**0.008**	0.026	0.169
	$l = 8$	**0.002**	0.013	0.235
	$l = 10$	**0.002**	0.009	0.262
	$l = 12$	**0.004**	0.005	0.242
$\min\{\lvert a-b\rvert, n-\lvert a-b\rvert\}$ for $a, b \in \{0, 1, 2, \ldots, n-1\}$.				
		Gen. poly. approximation	CLT approx.	BI upper bound
$\lvert R_A\rvert = 6$	$n = 12$	**0.045**	0.132	0.161
	$n = 14$	**0.046**	0.135	0.214
	$n = 16$	**0.046**	0.136	0.265
	$n = 18$	**0.047**	0.139	0.315
	$n = 20$	**0.047**	0.140	0.372
$\lvert R_A\rvert = 7$	$n = 12$	**0.050**	0.136	0.145
	$n = 14$	**0.050**	0.138	0.190
	$n = 16$	**0.053**	0.143	0.233
	$n = 18$	**0.052**	0.142	0.273
	$n = 20$	**0.052**	0.145	0.312
$\lvert R_A\rvert = 8$	$n = 12$	**0.055**	0.139	0.135
	$n = 14$	**0.055**	0.139	0.178
	$n = 16$	**0.056**	0.142	0.215
	$n = 18$	**0.056**	0.144	0.249
	$n = 20$	**0.058**	0.147	0.281
$\lvert R_A\rvert = 9$	$n = 12$	**0.060**	0.139	0.137
	$n = 14$	**0.058**	0.140	0.173
	$n = 16$	**0.060**	0.143	0.205
	$n = 18$	**0.060**	0.146	0.235
	$n = 20$	**0.061**	0.148	0.264
$\lvert R_A\rvert = 10$	$n = 12$	**0.063**	0.141	0.146
	$n = 14$	**0.063**	0.142	0.179
	$n = 16$	**0.062**	0.141	0.206
	$n = 18$	**0.063**	0.145	0.229
	$n = 20$	**0.064**	0.145	0.254

Table 1: The KL-divergence that compares the approximated (see Theorem 2) null distribution, the CLT approximation (see Corollary 3), and the BI upper bound (see Corollary 4). The ground-truth distribution was calculated for 10^6 samples of R_A and R_B of the same size. In both BI and CLT cases, Corollary 1 was used to calculate the null distribution of the minimum operator.

large $|R_A|$, it might be better to use Corollary 3 instead of Theorem 2. Also, with growing $|R_A|$, the CLT is more efficient than the generating-polynomials in terms of runtime. Thus, in studies such as (Song and Rudniy, 2010; Rudniy et al., 2014; Yamaguchi et al., 2012; Stoilos et al., 2005; Cheatham and Hitzler, 2013), the generating polynomials would be the method of choice, while in the case of (Ryšavý and Železný, 2016, 2019, 2023), the CLT approximation would be more efficient.

6.4 Dependence on Key Parameters

The problem of calculating the p-value has two major independent parameters - the size of the bags R_A and R_B, and the size of the universe (i.e., parameters l or n). To further illustrate the behavior of the methods, we include in Supplementary Material Fig. 3 plots that show averages of the results from Table 1 when only one of the parameters varies at a time.

From the plots, we can learn that the KL divergence of the presented method slightly grows with increasing $|R_A|$. This is in contrast with CLT approximation, where the KL divergence decreases in two of the three cases. In the case of the BI bound, the KL divergence decreases with the number of sampled elements. Nevertheless, this result is not as surprising as BI's having a much higher KL-divergence. The results of both the presented method, as well as CLT approximation do not change significantly with growing universe size. Nevertheless, with a larger universe, the error of the BI bound grows.

6.5 Comparison of Runtime Requirements

One of the advantages of the presented approach is its effectiveness compared to Barnard's Monte Carlo sampling approach (Marriott, 1979), where we repeatedly randomly sample bags, calculate distance, and compare it with the result.

The general runtime complexity of the method presented in this paper can be found in (13).

In comparison, when using Barnard's Monte Carlo sampling for a single evaluation of the Monge-Elkan distance, we need, in general, $\mathcal{O}(|R_A||R_B|t_d)$, where t_d is the complexity of single dst calculation. Therefore, to evaluate the p-value $P(\mathsf{Dst} \leq d)$, we need at least $\theta\left(\frac{1}{P(\mathsf{Dst} \leq d)}\right)$ distance calculations (assuming that we require a constant number of decimals points to be known).

To illustrate on a specific example, consider for now the Levenshtein distance on the position of dst. In this case $|Y| = \{0, 1, 2, \ldots, l\}$, where l is the length of the sequences. Then, the runtime of the generating-polynomial p-value calculation is in

$$\mathcal{O}\left(\log|R_A| \cdot |R_A||Y| \cdot \log(|R_A|l) + l + |R_B|^2\right), \tag{25}$$

which is even less than an evaluation of the single Monge-Elkan distance value. In the case of Barnard's Monte Carlo sampling, we need

$$\mathcal{O}\left(\frac{|R_A||R_B|(l+k^2)}{P(\mathsf{Dst} \leq d)}\right), \tag{26}$$

where k is the estimate of the maximum possible distance between any two strings in $R_A \times R_B$. The runtime of $\mathcal{O}(l + k^2)$ per Levenshtein distance calculation stems from the Ukkonen's cutoff heuristic (Ukkonen, 1985).

With the assumption that the null distribution of dst is known, both CLT approximation in Corollary 3, and BI upper bound in Corollary 4 can be evaluated in $\mathcal{O}(1)$.

To support the theoretical runtime requirements with real-world measurements, in Supplementary Material Table 3, we include the time needed to calculate the approximated null distribution, the empirical distribution from 10^6

samples made by Barnard's Monte Carlo sampling, the CLT approximation, and the BI upper bound. The measurements include time obtained as an average from ten consecutive calculations, followed by the standard deviation of the result. The results show that the runtime needed to evaluate CLT and BI is neglectable. Nevertheless, this speed is paid by worse KL-divergence than the one of the presented method. The results show that the speedup obtained by the generating polynomial method is in the order of hundreds to thousands in the case of Hamming and Levenshtein distances. In the third considered case, the speedup is up to tens of thousands.

7 Conclusion and Future Work

We have presented an algorithm to estimate the null distribution of the Monge-Elkan distance that can be used to compare sequence similarity from unassembled read bags. Our paper provides a method formulated using generating polynomials. In an experimental evaluation, we compared the proposed methods with Bernstein's inequality upper bound and an approximation based on the Central Limit theorem.

The methodology contains two simplifying assumptions that represent possible sources of error. However, we have confirmed empirically that their detrimental effect is generally not significant. In particular, the KL-divergence between the calculated distribution and the one obtained by Monte-Carlo sampling tends to be negligible. The contributed method thus represents a feasible tool that may even be a necessity when Monte-Carlo sampling is intractable due to the slow evaluation of the Monge-Elkan distance. The experiments and runtime complexity also show that with the growing size of the bags, the Central Limit theorem approximation will eventually become better in terms of runtime.

The method provides several options for future work. Of particular inter-

est is a theoretical assessment of the difference between the approximated null distribution and the exact one. From the practical point of view, the p-value calculation could be modified so that the p-value is not overestimated. Also, the ideas from Sect. 4 call for more experimental insights to evaluate the influence of non-uniformity of reads and compare with other tools and methods that use best-scoring matches.

Acknowledgments

This work was supported by Czech Science Foundation project 24-11664S.

The authors thank the anonymous reviewers for their valuable comments and suggestions. We also acknowledge the opportunity to present preliminary findings of this work at the ISBRA 2022 conference. A preprint version of the manuscript is available at the authors' website: `https://ida.fel.cvut.cz/zelezny/pubs/isbra22.pdf`.

Final publication is available from Mary Ann Liebert, Inc.: `https://doi.org/10.1089/cmb.2024.0854`.

Authorship confirmation and contribution statement

Petr Ryšavý: Conceptualization, Methodology, Software, Validation, Formal analysis, Writing - Original Draft, Writing - Review & Editing, and Visualization.

Filip Železný: Conceptualization, Methodology, Validation, Formal analysis, Writing - Review & Editing, Resources, Supervision, Project administration, and Funding acquisition.

Both authors have read and approved the final version of the manuscript.

Authors' disclosure

The authors declare that they have no competing interests.

Funding statement

This work was supported by Czech Science Foundation project 24-11664S.

References

Abdelkader, M. A method based on WordNet and Monge-Elkan distance for business process model matching. *Int. J. Inf. Syst. Model. Des.*, 9(4):37–48, oct 2018. ISSN 1947-8186. doi: 10.4018/IJISMD.2018100103.

Bernstein, S. On a modification of Chebyshev's inequality and of the error formula of laplace. *Ann. Sci. Inst. Sav. Ukraine, Sect. Math*, 1924.

Cantor, D. G. and Kaltofen, E. On fast multiplication of polynomials over arbitrary algebras. *Acta Informatica*, 28(7):693–701, Jul 1991. ISSN 1432-0525. doi: 10.1007/BF01178683.

Cheatham, M. and Hitzler, P. String similarity metrics for ontology alignment. In Alani, H. et al., editors, *The Semantic Web – ISWC 2013*, pages 294–309, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. ISBN 978-3-642-41338-4. doi: 10.1007/978-3-642-41338-4_19.

Chvátal, V. and Sankoff, D. Longest common subsequences of two random sequences. *Journal of Applied Probability*, 12(2):306–315, 1975. ISSN 00219002.

Cohen, W. W., Ravikumar, P., and Fienberg, S. E. A comparison of string distance metrics for name-matching tasks. In *Proceedings of the 2003 Inter-*

national Conference on Information Integration on the Web, IIWEB'03, page 73–78, USA, 2003. AAAI Press. doi: 10.1007/978-3-031-23101-8_26.

Feller, W. *Introduction to probability theory and its applications*. 1966. ISBN 978-0-471-25709-7.

Gali, N., Mariescu-Istodor, R., and Fränti, P. Similarity measures for title matching. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 1548–1553, 2016. doi: 10.1109/ICPR.2016.7899857.

Hamming, R. W. Error detecting and error correcting codes. *The Bell System Technical Journal*, 29(2):147–160, April 1950. ISSN 0005-8580. doi: 10.1002/j.1538-7305.1950.tb00463.x.

Hoeffding, W. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

Jimenez, S., Becerra, C., Gelbukh, A., et al. Generalized Mongue-Elkan method for approximate text string comparison. In Gelbukh, A., editor, *Computational Linguistics and Intelligent Text Processing*, pages 559–570, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

Kaplar, A., Aleksić, A., Stošović, M., et al. Evaluating string distance metrics for approximate dictionary matching: A case study in serbian electronic health records, 2019.

Knuth, D. *The art of computer programming (Seminumerical Algorithms)*, volume 2. Addison-Wesley, USA, 1981.

Kullback, S. and Leibler, R. A. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79 – 86, 1951. doi: 10.1214/aoms/1177729694.

Levenshtein, V. I. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady*, 10(8):707, 1966.

Majumdar, S. N. and Nechaev, S. Exact asymptotic results for the Bernoulli matching model of sequence alignment. *Phys. Rev. E*, 72:020901, Aug 2005.

Marriott, F. H. C. Barnard's Monte Carlo tests: How many simulations? *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):75–77, 1979. ISSN 00359254, 14679876.

Monge, A. E. and Elkan, C. P. The field matching problem: Algorithms and applications. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, pages 267–270, Portland, Oregon, 1996. AAAI Press.

Rudniy, A., Song, M., and Geller, J. Mapping biological entities using the longest approximately common prefix method. *BMC Bioinformatics*, 15(1): 187, Jun 2014. ISSN 1471-2105. doi: 10.1186/1471-2105-15-187.

Ryšavý, P. and Železný, F. Estimating sequence similarity from read sets for clustering sequencing data. In Boström, H. et al., editors, *Advances in Intelligent Data Analysis XV*, pages 204–214. Springer International Publishing, Cham, 2016. ISBN 978-3-319-46349-0. doi: 10.1007/978-3-319-46349-0_18.

Ryšavý, P. and Železný, F. Estimating sequence similarity from contig sets. In Adams, N. et al., editors, *Advances in Intelligent Data Analysis XVI*, pages 272–283, Cham, 2017. Springer International Publishing. ISBN 978-3-319-68765-0. doi: 10.1007/978-3-319-68765-0_23.

Ryšavý, P. and Železný, F. Estimating sequence similarity from read sets for clustering next-generation sequencing data. *Data Mining and Knowledge Discovery*, 33(1):1–23, January 2019. doi: 10.1007/s10618-018-0584-8.

Ryšavý, P. and Železný, F. Reference-free phylogeny from sequencing data. *BioData Mining*, 16(1):13, Mar 2023. ISSN 1756-0381.

Santos, R., Murrieta-Flores, P., and Martins, B. Learning to combine multiple string similarity metrics for effective toponym matching. *International Journal of Digital Earth*, 11(9):913–938, 2018.

Song, M. and Rudniy, A. Detecting duplicate biological entities using markov random field-based edit distance. *Knowledge and Information Systems*, 25 (2):371–387, Nov 2010. ISSN 0219-3116. doi: 10.1007/s10115-009-0254-7.

Stoilos, G., Stamou, G., and Kollias, S. A string metric for ontology alignment. In Gil, Y. et al., editors, *The Semantic Web – ISWC 2005*, pages 624–637, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg. ISBN 978-3-540-32082-1. doi: 10.1007/s10115-009-0254-7.

Ukkonen, E. Algorithms for approximate string matching. *Information and Control*, 64(1):100 – 118, 1985. ISSN 0019-9958. doi: 10.1016/S0019-9958(85)80046-2. URL `http://www.sciencedirect.com/science/article/pii/S0019995885800462`. International Conference on Foundations of Computation Theory.

Wagner, R. A. and Fischer, M. J. The string-to-string correction problem. *Journal of the Association for Computing Machinery*, 21(1):168–173, January 1974. ISSN 0004-5411. doi: 10.1145/321796.321811.

Winkler, W. E. String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. 1990.

Yamaguchi, A., Yamamoto, Y., Kim, J.-D., et al. Discriminative application of string similarity methods to chemical and non-chemical names for biomedical

abbreviation clustering. *BMC Genomics*, 13(3):S8, Jun 2012. ISSN 1471-2164. doi: 10.1186/1471-2164-13-S3-S8.

Zielezinski, A., Vinga, S., Almeida, J., et al. Alignment-free sequence comparison: benefits, applications, and tools. *Genome Biology*, 18(1):186, Oct 2017. ISSN 1474-760X. doi: 10.1186/s13059-017-1319-7.