# A Comparison Between Similarity Measures Based on Minimal Absent Words: An Experimental Approach*

Giuseppa Castiglione[1], Sabrina Mantaci[1,*], Salvatore L. Pizzuto[1] and Antonio Restivo[1]

[1]*Dipartimento di Matematica e Informatica, Università degli studi di Palermo, Palermo, Italy.*

### Abstract

In this paper we make some experimental considerations on the sets $\mathcal{D}(x,y)$, $M(x)\triangle M(y)$, $M(x)\cup M(y)$ involving minimal absent words of two words $x$ and $y$. This study is motivated by the computation of distances based on these sets.

### Keywords

Minimal absent words, alignment free distances

## 1. Introduction

It is well-known that sequence comparison finds many applications in comparative genomics for the study of evolutions, for building phylogenies, for comparing virus genomes. Besides the traditional methods based on alignment, that consider only local mutations in biological sequences, recently many alignment-free methods have been introduced, in order to consider also global mutations (see [1] for a survey). Some of them compare two sequences by counting their factors frequencies since, intuitively, the more similar two sequences are, the greater it is the number of the factors they share. Other methods use data compression considerations, based on the intuition that the more similar two sequences are, the more effective their joint compression is than their independent compression.

A third class of method generalizes the definition of sequences alignment, where the basic edit operation on characters are integrated with edit operations on blocks of characters. In the context of alignment free methods, in recent years a new class of methods consider the concept of minimal absent word, based on the idea that the negative information well represents the sequence itself, hence two sequences can be compared by comparing the relative sets of minimal absent words. The advantages of this approach are that the set of minimal absent words uniquely characterizes the sequence (cf. [2]), the number of minimal absent words of

*ICTCS 2024: Italian Conference on Theoretical Computer Science, September 11-13, 2024, Torino, Italy*

[*] Partially supported by INdAM-GNCS Project 2024- CUP E53C23001670001, by the FFR fund University of Palermo, and by the project "ACoMPA – Algorithmic and Combinatorial Methods for Pangenome Analysis" (CUP B73C24001050001) funded by the NextGeneration EU programme PNRR ECS00000017 Tuscany Health Ecosystem (Spoke 6).

[*] Corresponding author.

[†] These authors contributed equally.

giuseppa.castiglione@unipa.it (G. Castiglione); sabrina.mantaci@unipa.it (S. Mantaci); salvatoreleonardo.pizzuto@community.unipa.it (S. L. Pizzuto); antonio.restivo@unipa.it (A. Restivo)

0000-0002-1838-9785 (G. Castiglione); 0000-0002-9200-0520 (S. Mantaci); 0000-0002-1972-6931 (A. Restivo)

a sequence of length $n$ is linear in $n$ (cf. [3]), they can be computed in linear time [4]. As a consequence, it is possible to compare two sequences in time proportional to their lengths.

An experimental study of different distance measures based on minimal absent words to analyze similarity/dissimilarity of sequences has been carried out in [5].

In [6] Chairungsee and Crochemore introduced a measure of similarity between two sequences $x$ and $y$ making use of a *length-weighted index* on the symmetric difference $M(x) \triangle M(y)$ of the sets of minimal absent words $M(x)$ and $M(y)$ of $x$ and $y$, respectively. In the same paper, the authors propose to evaluate the length-weighted index on a *sample set*, i.e. the subset of $M(x) \triangle M(y)$ of words of limited length $\ell$. Further developments and an extension of the ideas of [6] can be found in [4].

In [7] a new similarity measure between sequences, based on minimal absent words, has been introduced with the aim to deepen a theoretical comparison with the measures in [6] and [8]. The flaw of the distance in [6] is that the set $M(x) \triangle M(y)$ could contain words that are absent both in $x$ and in $y$, although they are minimal only for one of them. In our opinion, if the aim is to distinguish $x$ and $y$ it is not appropriate to consider such words. Hence, we propose to evaluate the length-weighted index on the sample set $\mathscr{D}(x, y) = (F(x) \cap M(y)) \cup (F(y) \cap M(x))$, where $F(x)$ (resp. $F(y)$) denotes the set of factors of $x$ (resp. $y$). The set $\mathscr{D}(x, y)$ contains words that are minimal absent in one of the two words ($x$ or $y$), but that are factors of the other one. In our proposal, only the words of $\mathscr{D}(x, y)$ really contribute to distinguish $x$ and $y$.

Independently in [9, 10] a similar idea has been used for comparing a set of words $T$, called a *target*, against a set of words $R$, called a *reference* by defining a *$T$-specific word* as a factor $f$ of a word in $T$ that is not a factor of any word of $R$ and such that any proper factor of $f$ is a factor of some word of $R$. An algorithm for computing target specific words, whose construction is based on a generalization of suffix automata, is also proposed. Finally, in [11] a generalization of $M(x) \triangle M(y)$ for multiple strings is given.

From the algebraic point of view, the set $\mathscr{D}(x, y)$ is the base of the ideal generated by $M(x) \triangle M(y)$, hence $\mathscr{D}(x, y)$ contains only those words of $M(x) \triangle M(y)$ that do not have a proper factor in the same set. For this reason, in general, $\mathscr{D}(x, y)$ has far fewer elements than $M(x) \triangle M(y)$ and $\mathscr{D}(x, y)$ contains words among the shortest of $M(x) \triangle M(y)$. This choice, from a practical point of view, has a potential advantage in terms of computation time. Although we do not yet have an algorithm for generating the set $D(x, y)$ without considering all the words in $M(x) \cup M(y)$, we are confident that a more direct approach for this calculation can be introduced.

The experiments shown in this paper aim to provide measurements on how smaller the set $\mathscr{D}(x, y)$ is, compared to $M(x) \triangle M(y)$, and how shorter the words in $\mathscr{D}(x, y)$ are, compared to the ones in $M(x) \triangle M(y)$.

The paper is organized as follows: in Section 2 we give some notations and recall the definition of minimal absent word. In Section 3 we recall the similarity measures based on absent words. In Section 4 we comment on some experiments that aim to evaluate the amount of data needed to compute the two distances, that are highlighted in some graphs and tables.

## 2. Definitions and notations

Let $\Sigma$ be a finite alphabet and $\Sigma^*$ the set of the words over $\Sigma$. If $u \in \Sigma^*$, $|u|$ denotes its length. If $X \subset \Sigma^*$, $|X|$ denote its cardinality, i.e. the number of its elements, whereas $s(X) = \sum_{u \in X} |u|$ is the *total length* of $X$. A set $I \subseteq \Sigma^*$ is said to be a (*two-sided*) *ideal* of $\Sigma^*$ if for $u \in I$ and $v \in \Sigma^*$, then $uv, vu \in I$, i.e. $I = \Sigma^* I \Sigma^*$. The *base* of the ideal $I$ is the minimal set $B$ (with respect to the set inclusion) such that $I = \Sigma^* B \Sigma^*$. Let $v$ be a word of $\Sigma^*$, we say that $u$ is a *factor* of $v$ if there exist $z, w \in \Sigma^*$ such that $v = zuw$. In what follows we denote by $F(v)$ the set of factors of $v$. A word $u$ *occurs* in $v$ if it is a factor of $v$.

A word $u$ is an *absent word* for $v$ if it does not occur in $v$. An absent word is a *minimal absent word* (or *MAW*) for a word $v$ if all its proper factors occur in $v$. We denote by $M(v)$ the set of minimal absent words of $v$. For instance if $v = abaabab$, then $M(v) = \{aaa, aabaa, baba, bb\}$.

A language $L \subseteq \Sigma^*$ is called *factorial* if it contains all the factors of its own words, whereas it is called *antifactorial* if no word in the language is a proper factor of another word in the language. In particular, for any word $v \in \Sigma^*$, $F(v)$ is a factorial language and $M(v)$ is antifactorial.

Remark that the complement of $F(v)$ (i.e. the set of the words that are not factors of $v$) is an ideal of $\Sigma^*$ and $M(v)$ is its base. This allows to establish a duality between the sets $F(v)$ and $M(v)$ given by the relations (cf. [3]):

$$F(v) = \Sigma^* \setminus \Sigma^* M(v) \Sigma^*, \qquad M(v) = \Sigma F(v) \cap F(v)\Sigma \cap (\Sigma^* \setminus F(v)).$$

This last relation comes from the fact that if $v \in \Sigma^*$, the word $u = a_1 \cdots a_n$, with $a_i \in \Sigma$ is a MAW for $v$ iff $u \notin F(v)$ and $a_1 \cdots a_{n-1}, a_2 \cdots a_n \in F(v)$.

## 3. Similarity measures based on sets of minimal absent words

The idea to measure similarity by minimal absent words is based on the intuition that two words, $x$ and $y$, are as more distant as bigger is the set of the non common absent words and as shorter are the words in it. This idea was first formalized in a paper by Chairungsee and Crochemore [6] where the notion of length weighted index of a set is used in order to define a dissimilarity measure of two sequences. The *length weighted index* is defined as the measure that associates to a set $X \subseteq \Sigma^*$ the quantity $\mu(X) = \sum_{w \in X} \frac{1}{|w|^2}$.
This measure is used in [6] in order to define the distance function dist between two words $x$ and $y$, by taking the set $X = M(x) \triangle M(y)$, where $\triangle$ denotes the symmetric difference operator between two sets. Therefore the distance is defined as:

$$\mathrm{dist}(x, y) = \mu(M(x) \triangle M(y)) = \sum_{w \in M(x) \triangle M(y)} \frac{1}{|w|^2}$$

We remark that $\mathrm{dist}(x, y)$ is not substantially affected by long minimal absent words. This is why in [6] the authors propose to ignore from $M(x) \triangle M(y)$ those words with length longer than a fixed threshold $\ell$, and define a distance $\mathrm{dist}_\ell$ as the length weighted index over $M_\ell(x) \triangle M_\ell(y)$, where $M_\ell(x)$ ($M_\ell(y)$, resp.) denotes the set of MAWs of $x$ ($y$, resp.) with length smaller than or equal to $\ell$.

In [7] a different distance also based on the measure $\mu$ is considered, but applied to a subset of $M(x) \triangle M(y)$ that better captures the difference between two words. Moreover, by considering this subset, the requirement of having words with limited length is undirectely satisfied. This subset of $M(x) \triangle M(y)$ is in fact made of those factors of $x$ that are minimal absent words for $y$ and viceversa. In other terms, we want the comparison of the two sequences $x$ and $y$ not to be influenced by those minimal absent words of $y$ that are absent (but not minimal) also for $x$. This idea is formally described as follows. For all $x, y \in \Sigma^*$ we define

$$\mathscr{D}(x, y) = (F(x) \cap M(y)) \cup (F(y) \cap M(x)).$$

The following theorem summarizes some algebraic properties of $\mathscr{D}(x, y)$ also in relation with $M(x) \triangle M(y)$ proved in [7] (Lemma 4.1 and Theorem 4.3). Note that, in general, $M(x) \triangle M(y)$ is not antifactorial and $\Sigma^*(M(x) \triangle M(y))\Sigma^*$ is an ideal.

**Theorem 1.** *For all $x, y \in \Sigma^*$*

1. *$\mathscr{D}(x, y) = \emptyset$ if and only if $x = y$.*
2. *$\mathscr{D}(x, y) \subseteq M(x) \triangle M(y)$.*
3. *$\mathscr{D}(x, y)$ is antifactorial.*
4. *$\mathscr{D}(x, y)$ is the base of the ideal $\Sigma^*(M(x) \triangle M(y))\Sigma^*$.*

Point 4 of Theorem 1 states that considering $\mathscr{D}(x, y)$ is equivalent to ignore, in $M(x) \triangle M(y)$, those words that have a proper factor in the same set. Therefore one can define a distance based on the length weighted index applied to $\mathscr{D}(x, y)$:

$$\delta(x, y) = \mu(\mathscr{D}(x, y)) = \sum_{w \in \mathscr{D}(x,y)} \frac{1}{|w|^2}$$

We remark that as in the case of $\mathrm{dist}_\ell$, the distance $\delta$ takes into consideration elements among the shortest of $M(x) \triangle M(y)$ because they are elements of the base of the ideal $\Sigma^*(M(x) \triangle M(y))\Sigma^*$.

**Example 1.** *Let $x = cbaabdcb$ and $y = abcba$ words over $\Sigma = \{a, b, c, d\}$. Then,*
*$M(x) = \{ac, ad, bb, bc, ca, cc, cd, da, db, dd, aaa, aba, bab, cbd, dcba\}$*
*$M(y) = \{aa, ac, bb, ca, cc, aba, bab, cbc, d\}$,*
*$M(x) \cup M(y) = \{aa, aaa, aba, ac, ad, bab, bb, bc, ca, cc, cbc, cbd, cd, d, da, db, dcba, dd\}$,*
*$M(x) \triangle M(y) = \{d, aa, ad, bc, cd, da, db, dd, aaa, cbd, cbc, dcba\}$,*
*$\mathscr{D}(x, y) = \{d, aa, bc\}$.*

*Remark that the word cd, for instance, is absent both in x and in y (although not minimal in y) so, in some way, it represents a common property of the two words, and it should not be considered as a contribution to the distance. The same holds for the words ad, da, db, dd, aaa, cbd, cbc, and dcba. On the other hand, the word d, for instance, is a minimal absent word in y, but occurs in x and therefore discriminates the two words. Viceversa, the word aa is minimal absent in y but occurs in x i.e. it also contributes to their dissimilarity. In Example 1, the cardinality of the set $\mathscr{D}(x, y)$ is much smaller than the one of $M(x) \triangle M(y)$, and the words in $\mathscr{D}(x, y)$ are among the smallest in $M(x) \triangle M(y)$. Finally:*

$$\mathrm{dist}(x, y) = 1 + \frac{7}{4} + \frac{3}{9} + \frac{1}{16} = \frac{453}{144} \approx 3.1 \qquad \delta(x, y) = 1 + \frac{1}{2} = \frac{3}{2} = 1.5$$

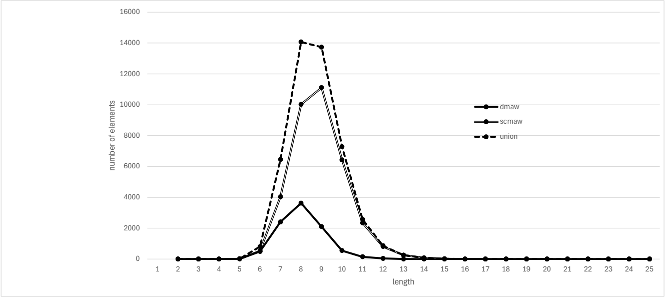

**Figure 1:** Distributions of MAWs lengths in $\mathscr{D}(x, y)$, $M(x)\triangle M(y)$ and $M(x) \cup M(y)$ for $x$ =Human mtDNA and $y$ =Gorilla mtDNA

## 4. Experimental results on the $\mathscr{D}(x, y)$ set

In the previous section we have observed that the set $\mathscr{D}(x, y)$ is the base of the ideal $\Sigma^*(M(x)\triangle M(y))\Sigma^*$ and then it is likely to have a smaller cardinality and that involves the words among the shortest. Actually, in [4], due to computational reasons, the distance $\mathrm{dist}_\ell$ is considered instead of the distance dist, but the authors do not give any motivation on how they choose the goode value of $\ell$. Moreover, some experiments that will appear in [12] show that the $\delta$ and the $\mathrm{dist}_\ell$ distances behave in a similar way on biological datasets with respect to the generated taxonomies.

Having an idea about the quantities involved could be interesting for the computation of $\delta$ and dist, whose computational complexity depends on the computation of the sets $\mathscr{D}(x, y)$ and $M(x)\triangle M(y)$, respectively. Therefore it is worth to see how much smaller $|\mathscr{D}(x, y)|$ is, w.r.t. $|M(x)\triangle M(y)|$ and $|M(x) \cup M(y)|$.

It is also interesting to consider and compare the total lengths $s$ of the three sets.

With these motivations here we present some experimental results. Our first experiments on this topic is performed by exploring sets $\mathscr{D}(x, y)$, $M(x)\triangle M(y)$, $M(x) \cup M(y)$ on a 41 mammals *mitochondrial DNA* (or *mtDNA*) benchmark dataset (https://github.com/NaserAnjum21/CD-MAWS/tree/master/Data). The sequences in this dataset are approximately 17000 bases long.

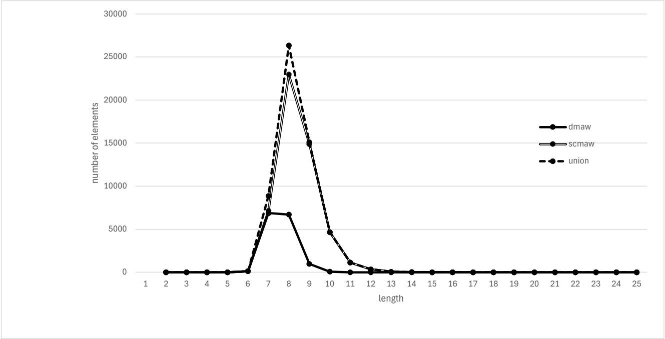

**Figure 2:** Distributions of MAWs lengths in $\mathscr{D}(x, y)$, $M(x)\triangle M(y)$ and $M(x) \cup M(y)$ for 4-letters alphabet and length 17000.

| $\vert\Sigma\vert$ | Lengths | $D(x, y)$ | $M(x)\triangle M(y)$ | $M(x) \cup M(y)$ |
|---|---|---|---|---|
| 2 | 8500 | 13 | 14 | 14 |
| | 17000 | 14 | 15 | 15 |
| | 34000 | 15 | 16 | 16 |
| | 68000 | 16 | 17 | 17 |
| | 136000 | 17 | 18 | 18 |
| 4 | 8500 | 7 | 8 | 8 |
| | 17000 | 7 | 8 | 8 |
| | 34000 | 8 | 9 | 9 |
| | 68000 | 8 | 9 | 9 |
| | 136000 | 9 | 10 | 10 |
| 8 | 8500 | 5 | 6 | 5 |
| | 17000 | 5 | 6 | 6 |
| | 34000 | 6 | 6 | 6 |
| | 68000 | 6 | 7 | 6 |
| | 136000 | 6 | 7 | 7 |

**Table 1**
This table shows, for each dataset $D_{\Sigma,m}$, the most represented length of MAWs in $D(x, y)$, $M(x)\triangle M(y)$, $M(x) \cup M(y)$, respectively, with $x, y \in D_{\Sigma,m}$.

Figure 1 summarizes the results concerning the distribution of MAW lengths in the three sets where $x$ corresponds to human's and $y$ to gorilla's mtDNA. The experiments on other pairs of species give similar curves.

A natural question is to ask what happens if the same experiments are performed on random strings. In fact this kind of experiments would allow us to infer some combinatorial properties of the sets. Then, in order to compare the results with those on biological strings, we produced a 17000-long randomly generated strings on a 4-letters alphabet dataset, whose results are displayed in Figure 2. We are interested to study the sensitivity of the sets $D(x, y)$ and $M(x)\triangle M(y)$ to the values of two parameters: the alphabet size and the dataset words length.

In order to run the experiment, we generated some random datasets to work on. For each alphabet size $\Sigma = 2, 4, 8$ and for each words-length $m = 8500$, 17000, 34000 68000, 136000, a dataset $D_{\Sigma,m}$ of random strings has been produced. (i.e. we iteratively doubled the alphabet size and the sequences lengths).

Then for each pair $x, y \in D_{\Sigma,m}$, we have computed the distribution of the MAWs lengths in $D(x, y)$, $M(x)\triangle M(y)$ and $M(x) \cup M(y)$. The results of some of these experiments are summarized in the histograms in Figures 3, 4, 5 where two random sequences $x, y \in D_{\Sigma,m}$ (with different values of $|\Sigma|$ and $m$) are considered and the corresponding distributions of the MAWs lengths in the different sets are shown. We observe that:

- The values for the three sets are distributed on a bell shaped curve and the values are nonzero in a small interval.
- The maximum for $\mathcal{D}(x, y)$ approximates $\log_{|\Sigma|} |x|$. This observation is coherent to a result in [2], stating that for a randomly generated word $x$ with a memoryless source and identical symbol probability, the maximal length of a minimal absent word is $O(\log_{|\Sigma|} |x|)$.

This value appears to be always one unity less than the maximum for $M(x) \triangle M(y)$ and $M(x) \cup M(y)$ (see Table 1).

- The curve for $\mathscr{D}(x, y)$ is much lower than the curves for $M(x) \triangle M(y)$ and $M(x) \cup M(y)$. This intuitively means that the number of the words in $\mathscr{D}(x, y)$ is much smaller than the ones in $M(x) \triangle M(y)$.
- The curves for $M(x) \triangle M(y)$ and $M(x) \cup M(y)$ are very close, i.e. the $M(x) \triangle M(y)$ involve most of the MAWs of $x$ and $y$.

Figures 3, 4 and 5 show the distributions of MAWs lengths in $\mathscr{D}(x, y)$, $M(x) \triangle M(y)$ and $M(x) \cup M(y)$ for two random strings on different alphabeth sizes $|\Sigma|$ and different string lengths $n$. It is easy to see, in all the dispalyes cases, how higher are the curves of length distributions of $M(x) \triangle M(y)$ and $M(x) \cap M(y)$ compared to the one of $\mathscr{D}(x, y)$. In particular:

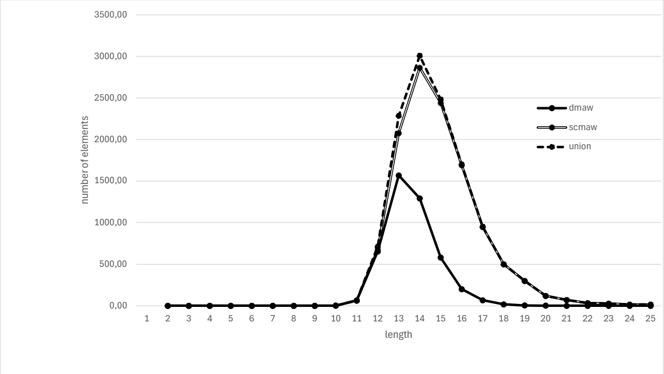

**Figure 3:** Distributions of MAWs lengths in $\mathscr{D}(x, y)$, $M(x) \triangle M(y)$ and $M(x) \cup M(y)$ for 2-letters alphabet and length 8500.

- For $m = 8500$ and $|\Sigma| = 2$ (cf. Figure 3) the curve for $\mathscr{D}(x, y)$ has its maximum in correspondence with length 13 (i.e. MAWs of length 13 are the most frequent in $\mathscr{D}(x, y)$) and the frequence is around 1500, whereas the maximum for both $M(x) \triangle M(y)$ and $M(x) \cup M(y)$ is in correspondence of length 14 with a frequence around 3000 (note that $\log_2 8500 = 13,053$). Nonzero values for $\mathscr{D}(x, y)$ are in the interval $[10, 18]$ whereas for both $M(x) \triangle M(y)$ and $M(x) \cup M(y)$ are in $[10, 24]$.
- For $m = 8500$ and $|\Sigma| = 4$ (cf. Figure 4) the curve for $\mathscr{D}(x, y)$ has its maximum in correspondence with length 7 and the frequence is around 5000, whereas the maximum for both $M(x) \triangle M(y)$ and $M(x) \cup M(y)$ is in correspondence of length 8 with a frequence around 11000 (note that $\log_4 17000 = 7,027$). Nonzero values for $\mathscr{D}(x, y)$ are in the interval $[5, 9]$ whereas for both $M(x) \triangle M(y)$ and $M(x) \cup M(y)$ are in $[5, 12]$.
- For $m = 136000$ and $|\Sigma| = 8$ (Figure 5) the curve for $\mathscr{D}(x, y)$ has its maximum in correspondence with length 6 and the frequence is around 120000, whereas the maximum for both $M(x) \triangle M(y)$ and $M(x) \cup M(y)$ is in correspondence of length 7 with a frequence around 500000 (note that $\log_8 136000 = 5,684$). Nonzero values for $\mathscr{D}(x, y)$ are in the interval $[5, 8]$ whereas for both $M(x) \triangle M(y)$ and $M(x) \cup M(y)$ are in $[5, 10]$.

The experiment, repeated on different sample sequences, gives similar curves and equal maximum frequence. The curves are similar also for sample sequences taken from biological datasets with lengths comparable to the random sequences here considered (see, for instance, Figure 1).

For the investigation about cardinalities, in another experiment, for all of the pairs $x, y \in D_{\Sigma,m}$ we computed the ratios $|\mathscr{D}(x,y)|/|M(x) \triangle M(y)|$, $|\mathscr{D}(x,y)|/|M(x) \cup M(y)|$, $s(\mathscr{D}(x,y))/s(M(x) \triangle M(y))$ and $s(\mathscr{D}(x,y))/s(M(x) \cup M(y))$. Tables 2 and 3 show the average of these values and the corresponding standard deviation. One can note that:

- As the cardinality of the alphabet grows, $|\mathscr{D}(x,y)|/|M(x) \triangle M(y)|$ and $|\mathscr{D}(x,y)|/|M(x) \cup M(y)|$ decrease. This is also true w.r.t. the total lengths.
- The ratios relating to the total lengths are smaller than the corresponding ratios relating to the cardinalities. This shows that the words in $\mathscr{D}(x,y)$ are among the smallest of the words in $M(x) \triangle M(y)$.

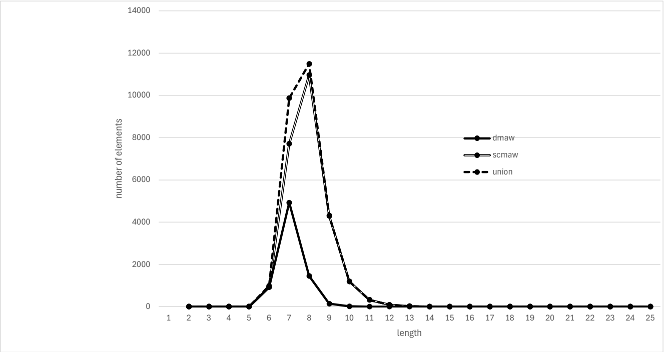

**Figure 4:** Distributions of MAWs lengths in $\mathscr{D}(x,y)$, $M(x) \triangle M(y)$ and $M(x) \cup M(y)$ for 4-letters alphabet and length 8500

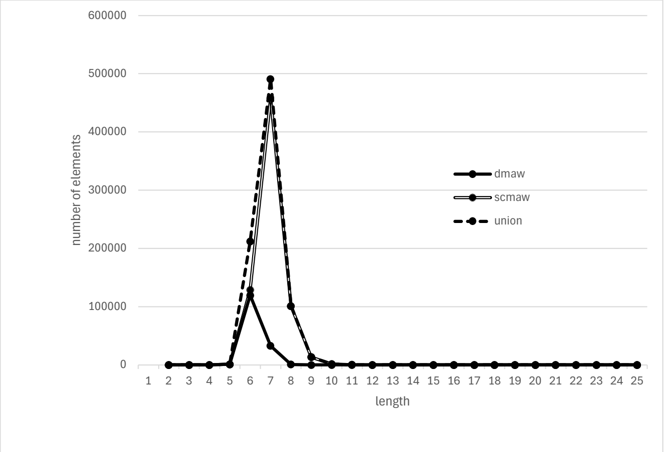

**Figure 5:** Distributions of MAWs lengths in $\mathscr{D}(x, y)$, $M(x) \triangle M(y)$ and $M(x) \cup M(y)$ for 8-letters alphabet and length 136000

| $\lvert\Sigma\rvert$ | Lengths | $avg_1 \times 100\%$ (s.d) | $avg_2 \times 100\%$ (s.d) |
|---|---|---|---|
| 2 | 8500 | 37.71% (0.49) | 34.46% (0.50) |
| | 17000 | 37.67% (0.34) | 34.64% (0.34) |
| | 34000 | 37.76% (0.22) | 34.93% (0.22) |
| | 68000 | 37.76% (0.19) | 35.11% (0.19) |
| | 136000 | 37.83% (0.13) | 35.36% (0.13) |
| 4 | 8500 | 29.28% (0.32) | 26.22% (0.31) |
| | 17000 | 28.77% (0.20) | 25.93% (0.20) |
| | 34000 | 29.33% (0.16) | 26.62% (0.16) |
| | 68000 | 28.74% (0.09) | 26.20% (0.09) |
| | 136000 | 29.25% (0.08) | 26.81% (0.08) |
| 8 | 8500 | 23.51% (0.31) | 20.67% (0.28) |
| | 17000 | 21.95% (0.12) | 19.07% (0.11) |
| | 34000 | 19.58% (0.10) | 17.51% (0.10) |
| | 68000 | 23.59% (0.09) | 21.17% (0.09) |
| | 136000 | 21.94% (0.04) | 19.48% (0.04) |

**Table 2**
For each pair of words $x, y \in D_{\Sigma,m}$, the ratios $R_1(x, y) = \frac{\lvert\mathscr{D}(x,y)\rvert}{\lvert M(x) \triangle M(y)\rvert}$ and $R_2(x, y) = \frac{s(\mathscr{D}(x,y))}{s(M(x) \triangle M(y))}$, resp. have been computed. Afterwards, the average values $avg_1 = avg_{x,y \in D_{\Sigma,m}}(R_1(x, y))$ and $avg_2 = avg_{x,y \in D_{\Sigma,m}}(R_2(x, y))$, resp., have been computed and reported in terms of percentage in columns 3 and 4, respectively, with the relative standard deviation (s.d.) in parenthesis. Since the sandard deviation is everywhere very small, this means that data are clustered tightly around the mean.

| $\lvert\Sigma\rvert$ | Lengths | $avg_3 \times 100\%$ (s.d.) | $avg_4 \times 100\%$ (s.d.) |
|---|---|---|---|
| 2 | 8500 | 36.37% (0.44) | 33.35% (0.46) |
| | 17000 | 36.32% (0.30) | 33.50% (0.31) |
| | 34000 | 36.41% (0.20) | 33.77% (0.21) |
| | 68000 | 36.41% (0.17) | 33.94% (0.18) |
| | 136000 | 36.49% (0.12) | 34.18% (0.12) |
| 4 | 8500 | 26.29% (0.25) | 23.78% (0.24) |
| | 17000 | 26.09% (0.15) | 23.69% (0.16) |
| | 34000 | 26.33% (0.12) | 24.10% (0.13) |
| | 68000 | 26.06% (0.07) | 23.92% (0.07) |
| | 136000 | 26.27% (0.06) | 24.27% (0.06) |
| 8 | 8500 | 18.91% (0.17) | 16.99% (0.17) |
| | 17000 | 18.83% (0.08) | 16.65% (0.08) |
| | 34000 | 16.81% (0.06) | 15.17% (0.06) |
| | 68000 | 18.94% (0.05) | 17.32% (0.05) |
| | 136000 | 18.83% (0.02) | 16.96% (0.03) |

**Table 3**

For each pair of words $x, y \in D_{\Sigma,m}$, the ratios $R_3(x,y) = \frac{|\mathscr{D}(x,y)|}{|M(x) \cup M(y)|}$ and $R_4(x,y) = \frac{s(\mathscr{D}(x,y))}{s(M(x) \cup M(y))}$, resp., have been computed. Afterwards the average values $avg_3 = avg_{x,y \in D_{\Sigma,m}}(R_3(x,y))$ and $avg_4 = avg_{x,y \in D_{\Sigma,m}}(R_4(x,y))$, resp. have been computed and reported in terms of percentage in columns 3 and 4, respectively, with the relative standard deviation (s.d.) in parenthesis. Since the sandard deviation is everywhere very small, this means that data are clustered tightly around the mean.

In conclusion, these experiments are aimed to remark that the great numerical difference of the data dimension in these sets, make the distance $\delta$ interesting, from a computational point of view, compared to the distance based on the symmetric difference. In fact, computing a distance based on $\mathscr{D}(x,y)$ could be more efficient than computing the distance on $M(x) \triangle M(y)$ since we would have a smaller set, provided that one can get directly the words of the set $\mathscr{D}(x,y)$ without explicitely producing all the words in $M(x), M(y), F(x), F(y)$.

## References

[1] S. Mantaci, A. Restivo, M. Sciortino, Distance measures for biological sequences: Some recent approaches, Int. J. Approx. Reason. 47 (2008) 109–124. URL: https://doi.org/10.1016/j.ijar.2007.03.011. doi:10.1016/J.IJAR.2007.03.011.

[2] F. Mignosi, A. Restivo, M. Sciortino, Forbidden factors and fragment assembly, RAIRO Theor. Informatics Appl. 35 (2001) 565–577.

[3] M. Crochemore, F. Mignosi, A. Restivo, Automata and forbidden words, Inf. Process. Lett. 67 (1998) 111–117.

[4] P. Charalampopoulos, M. Crochemore, G. Fici, R. Mercas, S. P. Pissis, Alignment-free sequence comparison using absent words, Inf. Comput. 262 (2018) 57–68.

[5] M. S. Rahman, A. Alatabbi, M. Crochemore, M. S. Rahman, Absent words and the (dis)sim-

ilarity analysis of dna sequences: An experimental study, BMC Research Notes. 9:186 450 (2016) 1–8.

[6] S. Chairungsee, M. Crochemore, Using minimal absent words to build phylogeny, Theor. Comput. Sci. 450 (2012) 109–116.

[7] G. Castiglione, S. Mantaci, A. Restivo, Some investigations on similarity measures based on absent words, Fundam. Informaticae 171 (2020) 97–112.

[8] A. Ehrenfeucht, D. Haussler, A new distance metric on strings computable in linear time, Discrete Applied Mathematics 20 (1988) 191–203.

[9] M. Béal, M. Crochemore, Fast detection of specific fragments against a set of sequences, volume 13911 of *Lecture Notes in Computer Science*, Springer, 2023, pp. 51–60.

[10] P. Bonizzoni, C. D. Felice, Y. Pirola, R. Rizzi, R. Zaccagnino, R. Zizza, Can formal languages help pangenomics to represent and analyze multiple genomes?, in: V. Diekert, M. V. Volkov (Eds.), Developments in Language Theory - 26th International Conference, DLT 2022, Tampa, FL, USA, May 9-13, 2022, Proceedings, volume 13257 of *Lecture Notes in Computer Science*, Springer, 2022, pp. 3–12.

[11] K. Okabe, T. Mieno, Y. Nakashima, S. Inenaga, H. Bannai, Linear-time computation of generalized minimal absent words for multiple strings, volume 14240 of *Lecture Notes in Computer Science*, Springer, 2023, pp. 331–344.

[12] S. L. Pizzuto, Similarity measures based on minimal absent words, Master's thesis, DAMI, University of Palermo, in preparation.