

Isometric Words based on Swap and Mismatch Distance *

M. Anselmo¹ G. Castiglione² M. Flores¹
D. Giammarresi³ M. Madonia⁴ S. Mantaci²

¹ Dipartimento di Informatica, Università di Salerno, Italy.

{manselmo, mflores}@unisa.it

² Dipartimento di Matematica e Informatica, Università di Palermo, Italy

{giuseppa.castiglione, sabrina.mantaci}@unipa.it

³ Dipartimento di Matematica. Università Roma “Tor Vergata” Italy.

giammarr@mat.uniroma2.it

⁴ Dipartimento di Matematica e Informatica, Università di Catania, Italy.

madonia@dmi.unict.it

Abstract

An edit distance is a metric between words that quantifies how two words differ by counting the number of edit operations needed to transform one word into the other one. A word f is said isometric with respect to an edit distance if, for any pair of f -free words u and v , there exists a transformation of minimal length from u to v via the related edit operations such that all the intermediate words are also f -free. The adjective “isometric” comes from the fact that, if the Hamming distance is considered (i.e., only mismatches), then isometric words are connected with definitions of isometric subgraphs of hypercubes.

We consider the case of edit distance with swap and mismatch. We compare it with the case of mismatch only and prove some properties of isometric words that are related to particular features of their overlaps.

Keywords: Swap and mismatch distance, Isometric words, Overlap with errors.

1 Introduction

The edit distance is a central notion in many fields of computer science. It plays a crucial role in defining combinatorial properties of families of strings as well as in designing many classical string algorithms that find applications in natural language processing, bioinformatics and, in general, in information retrieval problems. The edit distance is a string metric that quantifies how two strings differ from each other and it is based on counting the minimum number of edit operations required to transform one string into the other one.

*Partially supported by INdAM-GNCS Project 2022 and 2023, FARB Project ORSA229894 of University of Salerno, TEAMS Project of University of Catania and by the MIUR Excellence Department Project MatMod@TOV awarded to the Department of Mathematics, University of Rome Tor Vergata.

Different definitions of edit distance use different sets of edit operations. The operations of insertion, deletion and replacement of a character in the string characterize the Levenshtein distance which is probably the most widely known (cf. [17]). On the other hand, the most basic edit distance is the Hamming distance which applies only to pair of strings of the same length and counts the positions where they have a mismatch; this corresponds to the restriction of using only the replacement operation. For this, the Hamming distance finds a direct application in detecting and correcting errors in strings and it is a major actor in the algorithms for string matching with mismatches (see [13]).

The notion of isometric word (or string) combines the edit distance with the property that a word does not appear as factor in other words. Note that this property is important in combinatorics as well as in the investigation on similarities, or distances, on DNA sequences, where the avoided factor is referred to as an absent word [8, 9, 10]. Isometric words based on Hamming distance were first introduced in [15] as special binary strings that never appear as factors in some string transformations. A string is *f-free* if it does not contain f as factor. A word f is isometric if for any pair of f -free words u and v , there exists a sequence of symbol replacement operations that transform u in v where all the intermediate words are also f -free.

Isometric words are connected with the definition of isometric subgraphs of the hypercubes, called generalized Fibonacci cubes. The hypercube graph Q_n is a graph whose vertices are the (binary) words of length n , and two vertices are adjacent when the corresponding words differ in exactly one symbol. Therefore, the distance between two vertices is the Hamming distance of the corresponding vertex-words. Let $Q_n(f)$ be the subgraph of Q_n which contains only vertices that are f -free. Then, if f is isometric, the distances of the vertices in $Q_n(f)$ are the same as calculated in the whole Q_n . Fibonacci cubes have been introduced by Hsu in [14] and correspond to the case with $f = 11$. In [15, 16, 19, 20, 22] the structure of non-isometric words for alphabets of size 2 and Hamming distance is completely characterized and related to particular properties on their overlaps. The more general case of alphabets of size greater than 2 and Lee distance is studied in [3, 4, 5]. Using these characterizations, in [7] some linear-time algorithms are given to check whether a binary word is Hamming isometric and, for quaternary words, if it is Lee isometric. These algorithms were extended to provide further information on non-isometric words, still keeping linear complexity in [4]. Binary Hamming isometric two-dimensional words have been also studied in [6].

Many challenging problems in correcting errors in strings come from computational biology. Among the chromosomal operations on DNA sequences, in gene mutations and duplication, it seems natural to consider the *swap* operation, consisting in exchanging two adjacent symbols. The Damerau-Levenshtein distance adds also the swap to all edit operations. In [18], Wagner proves that the edit distance with insertion and swap is NP-hard, while each separate case can be solved in polynomial time. Moreover, the general edit distance with insertion, deletion, replacement, and swap, is polynomially solvable. The swap-matching problem has been considered in [1, 12], and algorithms for computing the corresponding edit distance are given in [2, 11].

In this paper, we study the notion of isometric word using the edit distance based on swaps and mismatches. This distance will be referred to by using the *tilde* symbol that somehow evokes the swap operation. The tilde-distance $\text{dist}_{\sim}(u, v)$ of equal-length words u and v is the minimum number of replacement and swap operations to transform u into v . Then, the definition of *tilde-isometric* word comes in a very natural way. A word f is tilde-isometric if for any pair of equal-length words u and v that are f -free, there is a transformation from u to v that uses exactly $\text{dist}_{\sim}(u, v)$ replacement

and swap operations and such that all the intermediate words still avoid f . It turns out that adding the swap operation to the definition makes the situation more complex, but interesting for applications. It is not a mere generalization of Hamming string isometry since special situations arise. A swap operation in fact is equivalent to two replacements, but it counts as one when computing the tilde-distance. Moreover, there could be different ways to transform u into v since particular triples of consecutive symbols can be managed, from left to right, either by first a swap and then a replacement or by a replacement and then a swap. We present some examples of tilde-isometric words that are not Hamming isometric and vice versa. By definition, in order to prove that a given string f is not tilde-isometric one should exhibit a pair of f -free words $(\tilde{\alpha}, \tilde{\beta})$ such that any transformation from $\tilde{\alpha}$ to $\tilde{\beta}$ of length $\text{dist}_{\sim}(\tilde{\alpha}, \tilde{\beta})$ comes through words that contain f . Such a pair is called pair of *tilde-witnesses* for f . We prove some necessary conditions for f to be non-isometric based on the notion of error-overlap and give an explicit construction of the tilde-witnesses in many cases.

2 Preliminaries

Let Σ be a finite alphabet. A word (or string) w of length $|w| = n$, is $w = a_1 a_2 \dots a_n$, where a_1, a_2, \dots, a_n are symbols in Σ . The set of all words over Σ is denoted Σ^* and the set of all words over Σ of length n is denoted Σ^n . Finally, ϵ denotes the *empty word* and $\Sigma^+ = \Sigma^* - \{\epsilon\}$. For any word $w = a_1 a_2 \dots a_n$, the *reverse* of w is the word $w^{rev} = a_n a_{n-1} \dots a_1$. If $x \in \{0, 1\}$, we denote by \bar{x} the opposite of x , i.e. $\bar{x} = 1$ if $x = 0$ and viceversa. Then we define *complement* of w the word $\bar{w} = \bar{a}_1 \bar{a}_2 \dots \bar{a}_n$.

Let $w[i]$ denote the symbol of w in position i , i.e. $w[i] = a_i$. Then, $w[i..j] = a_i \dots a_j$, for $1 \leq i \leq j \leq n$, is a *factor* of w . The *prefix* (resp. *suffix*) of w of length l , with $1 \leq l \leq n - 1$ is $\text{pre}_l(w) = w[1..l]$ (resp. $\text{suf}_l(w) = w[n - l + 1..n]$). When $\text{pre}_l(w) = \text{suf}_l(w) = u$ then u is here referred to as an *overlap* of w of length l ; it is also called border, or bifix. A word w is said *f-free* if w does not contain f as a factor.

An *edit operation* is a function $O : \Sigma^* \rightarrow \Sigma^*$ that transform a word into another one. Among the most common edit operations there are the insertion, the deletion or the replacement of a character and the swap of two adjacent characters. Let OP be a *set of edit operations*. The *edit distance* of two words $u, v \in \Sigma^*$ is the minimum number of edit operations in OP needed to transform u into v . In this paper, we consider the edit distance that uses only replacements and swaps. Note that these two operations do not change the length of the word. We give a formal definition.

Definition 1 Let Σ be a finite alphabet and $w = a_1 a_2 \dots a_n$ a word over Σ .

The replacement operation (or replacement, for short) on w at position i with $x \in \Sigma$, $x \neq a_i$, is defined by

$$R_{i,x}(a_1 a_2 \dots a_{i-1} a_i a_{i+1} \dots a_n) = a_1 a_2 \dots a_{i-1} x a_{i+1} \dots a_n.$$

The swap operation (or swap, for short) on w at position i consists in exchanging characters at positions i and $i + 1$, provided that they are different, $a_i \neq a_{i+1}$,

$$S_i(a_1 a_2 \dots a_i a_{i+1} \dots a_n) = a_1 a_2 \dots a_{i+1} a_i \dots a_n.$$

When the alphabet $\Sigma = \{0, 1\}$ there is only a possible replacement at a given position i , so we write $R_i(w)$ instead of $R_{i,x}(w)$.

Given two equal-length words $u = a_1 \cdots a_n$ and $v = b_1 \cdots b_n$, they have a *mismatch error* (or *mismatch*) at position i if $a_i \neq b_i$ and they have a *swap error* (or *swap*) at position i if $a_i a_{i+1} = b_{i+1} b_i$, with $a_i \neq a_{i+1}$. We say that u and v have an *error* at position i if they have either a mismatch or a swap error.

Note that one swap corresponds to two adjacent mismatches.

A word f is isometric if for any pair of f -free words u and v , there exists a sequence of minimal length of replacement operations that transform u into v where all the intermediate words are also f -free. In this paper we refer to this definition of isometric as *Ham-isometric*. In [21], a word w has a 2-error overlap if there exists l such that $\text{pre}_l(w)$ and $\text{suf}_l(w)$ have two mismatch errors. Then, they prove the following characterization.

Proposition 2 *A word f is Ham-isometric if and only if f has a 2-error overlap.*

3 Tilde-distance and tilde-isometric words

In this section we consider the edit distance based on swap and replacement operations that we call tilde-distance and we denote dist_\sim . First, we give some definitions and notations, together with some examples and the proofs of some preliminary properties.

Definition 3 *Let $u, v \in \Sigma^*$ be words of equal length. The tilde-distance $\text{dist}_\sim(u, v)$ between u and v is the minimum number of replacements and swaps needed to transform u into v .*

Definition 4 *Let $u, v \in \Sigma^*$ be words of equal length. A tilde-transformation τ of length h from u to v is a sequence of words (w_0, w_1, \dots, w_h) such that $w_0 = u$, $w_h = v$, and for any $k = 0, 1, \dots, h-1$, $\text{dist}_\sim(w_k, w_{k+1}) = 1$. Moreover, τ is f -free if for any $i = 0, 1, \dots, h$, the word w_i is f -free.*

A tilde-transformation (w_0, w_1, \dots, w_h) from u to v is associated to a sequence of h operations $(O_{i_1}, O_{i_2}, \dots, O_{i_h})$ such that, for any $k = 1, \dots, h$, $O_{i_k} \in \{R_{i_k, x}, S_{i_k}\}$ and $w_k = O_{i_k}(w_{k-1})$; it can be represented as follows:

$$u = w_0 \xrightarrow{O_{i_1}} w_1 \xrightarrow{O_{i_2}} \dots \xrightarrow{O_{i_h}} w_h = v.$$

With a little abuse of notation, in the sequel we will refer to a tilde-transformation both as a sequence of words and as a sequence of operations. We give some examples.

Example 5 *Let $u = 1011, v = 0110$. Below we show two different tilde-transformations from u to v . Note that the length of τ_1 corresponds to $\text{dist}_\sim(u, v) = 2$.*

$$\tau_1 : 1011 \xrightarrow{S_1} 0111 \xrightarrow{R_4} 0110 \quad \tau_2 : 1011 \xrightarrow{R_1} 0011 \xrightarrow{R_2} 0111 \xrightarrow{R_4} 0110$$

Furthermore, consider the following tilde-transformations of $u' = 100$ into $v' = 001$:

$$\tau'_1 : 100 \xrightarrow{S_1} 010 \xrightarrow{S_2} 001 \quad \tau'_2 : 100 \xrightarrow{R_1} 000 \xrightarrow{R_2} 001$$

Note that both τ'_1 and τ'_2 have the same length equal to $\text{dist}_\sim(u', v') = 2$. Interestingly, in τ'_1 the symbol in position 2 is changed twice.

The next lemma shows that, in the case of a two letters alphabet, we can restrict to tilde-transformations where each character is changed at most once.

Lemma 6 *Let $u, v \in \{0, 1\}^m$ with $m \geq 1$. Then, there exists a tilde-transformation of u into v of length $\text{dist}_\sim(u, v)$ such that for any $i = 1, 2, \dots, m$, the character in position i is changed at most once.*

Proof: Let $u = a_1 \dots a_m$ and $v = b_1 \dots b_m$ and let τ be a tilde-transformation of u into v of length $d = \text{dist}_\sim(u, v)$. Suppose that, for some i , the character in position i is changed more than once by τ and let O_t and O_s be the first and the second operation, respectively, that modify the character in position i . Observe that the character in position i can be changed by the operations R_i, S_{i-1} or S_i .

Suppose that $O_t = S_i$ and $O_s = R_i$. Then, the symbol a_i is changed twice and two operations S_i and R_i could be replaced by a single R_{i+1} . This would yield a tilde-transformation of u into v of length strictly less than d ; this is a contradiction to the definition of tilde-distance. Similarly for the cases where $O_t = R_i$ and $O_s = S_i$, $O_t = S_{i-1}$ and $O_s = R_i$, $O_t = R_i$ and $O_s = S_{i-1}$.

Finally, if $O_t = S_{i-1}$ and $O_s = S_i$ then the three characters in positions $i-1, i$ and $i+1$ are changed, but the one in position i is changed twice. Hence, the two swap operations S_{i-1} and S_i can be replaced by R_{i-1} and R_{i+1} yielding a tilde-transformation of u into v of same length d which instead involves positions $i-1$ and i just once (see τ'_2 in Example 5). \square

Remark 7 *Lemma 6 only applies to a binary alphabet. Indeed, if $\Sigma = \{0, 1, 2\}$, and take $u = 012$ and $v = 120$, then $\text{dist}_\sim(012, 120) = 2$ because there is the tilde-transformation $012 \xrightarrow{S_1} 102 \xrightarrow{S_2} 120$. Instead, in order to change each character at most once, three replacement operations are needed.*

Definition 8 *Let Σ be a finite alphabet and $u, v \in \Sigma^+$. A tilde-transformation from u to v is minimal if its length is equal to $\text{dist}_\sim(u, v)$ and characters in each position are modified at most once.*

Lemma 6 guarantees that, in the binary case, a minimal tilde-transformation always exists. In the sequel, this will be the most investigated case. Let us now define isometric words based on the swap and mismatch distance.

Definition 9 *Let $f \in \Sigma^n$, with $n \geq 1$, f is tilde-isometric if for any pair of f -free words u and v of length $m > n$, there exists a minimal tilde-transformation from u to v that is f -free. It is tilde-non-isometric if it is not tilde-isometric.*

In order to prove that a word is tilde-non-isometric it is sufficient to exhibit a pair (u, v) of words contradicting the Definition 9. Such a pair will be referred to as tilde-witnesses for f . Some examples follow.

Definition 10 *A pair (u, v) of words in Σ^m is a pair of tilde-witnesses for f if:*

1. u and v are f -free
2. $\text{dist}_\sim(u, v) \geq 2$
3. there exists no minimal tilde-transformation from u to v that is f -free.

Example 11 The word $f = 1010$ is tilde-non-isometric because $u = 11000$ and $v = 10110$ are tilde-witnesses for f . In fact, the only possible minimal tilde-transformations from u to v are $11000 \xrightarrow{S_2} 10100 \xrightarrow{R_4} 10110$ and $11000 \xrightarrow{R_4} 11010 \xrightarrow{S_2} 10110$ and in both cases 1010 appears as factor after the first step.

Remark 12 When a transformation contains a swap and a replacement that are adjacent, there could exist many distinct minimal tilde-transformations that involve different sets of operations. For instance, the pair (u, v) , with $u = 010$ and $v = 101$, has the following minimal tilde-transformations:

$$010 \xrightarrow{S_1} 100 \xrightarrow{R_3} 101 \qquad 010 \xrightarrow{S_2} 001 \xrightarrow{R_1} 101$$

This fact cannot happen when only replacements are allowed. For this reason studying tilde-isometric words is more complicated than the Hamming case.

Example 11 shows a tilde-non-isometric word. Proving that a given word is tilde-isometric is much harder since it requires to give evidence that no tilde-witnesses exist. We will now prove that word 111000 is isometric with *ad-hoc* technique.

Example 13 The word $f = 111000$ is tilde-isometric. Suppose by the contrary that f is tilde-non-isometric and let (u, v) be a pair of tilde-witnesses for f of minimal tilde-distance. If u and v have only mismatch errors, this is the case of the Hamming distance and results from this theory [4, 19] show that $u = 11100\mathbf{1}1000$ and $v = 1110\mathbf{1}01000$; these are not tilde-witnesses since $\text{dist}_{\sim}(u, v) = 1$.

Therefore, u and v have a swap error in some position i ; suppose $u[i..i+1] = 01$. The minimality of $\text{dist}_{\sim}(u, v)$ implies that $S_i(u)$ is not f -free. Then, a factor 111000 appears in $S_i(u)$ from position $i-2$, and $u[i-2..i+3] = 110100$. Since v is f -free, then there is another error in u involving some positions in $[i-2..i+3]$. It cannot be neither a swap (since there are no adjacent different symbols that are not changed yet), nor a mismatch in positions $i-1, i+2$ (since the corresponding replacement cannot let f occur). Then, it is a mismatch in position $i-2$ or $i+3$. Consider the case of a mismatch in position $i+3$ (the other case is analogous). Then, $u[i+3..i+8] = 011000$ and there is another error in $[i+4..i+8]$, in fact, in position $i+6$ or $i+8$. Continuing with similar reasoning, one falls back to the previous situation. This is a contradiction because the length of u is finite.

From now on, we consider only the binary alphabet $\Sigma = \{0, 1\}$ and we study isometric binary words beginning by 1, in view of the following lemma whose proof can be easily inferred by combining the definitions.

Lemma 14 Let $f \in \{0, 1\}^n$. The following statements are equivalent:

1. f is tilde-isometric
2. f^{rev} is tilde-isometric
3. \overline{f} is tilde-isometric.

Let us conclude the section by comparing tilde-isometric with Ham-isometric words. Although the tilde-distance is more general than the Hamming distance, they are incomparable, as stated in the following proposition.

Proposition 15 *There exists a word which is tilde-isometric but Ham-non-isometric, and a word which is tilde-non-isometric, but Ham-isometric.*

Proof: The word $f = 111000$ is tilde-isometric (see Example 13), but f is Ham-non-isometric by Proposition 2.

Conversely, $f' = 1010$ is tilde-non-isometric (see Example 11), but Ham-isometric by Proposition 2.

□

4 Tilde-isometric words and tilde-error overlaps

In this section we focus on the word property of being tilde-non-isometric and connect it to the number of errors in its overlaps. The idea reminds the characterization for Ham-isometric words recalled in Proposition 2 but the swap operation changes all the perspectives as pointed also in Proposition 15.

Definition 16 *Let $f \in \{0, 1\}^n$. Then, f has a q -tilde-error overlap of length l , with $1 \leq l \leq n - 1$ and $0 \leq q \leq l$, if $\text{dist}_{\sim}(\text{pre}_l(f), \text{suf}_l(f)) = q$.*

In other words, if f has a q -tilde-error overlap of length l then there exists a minimal tilde-transformation τ from $\text{pre}_l(f)$ to $\text{suf}_l(f)$ of length q . In the sequel, when $q = 2$, in order to specify the kind of errors, a 2-tilde-error overlap is referred to be of type RR if τ consists of two replacements, of type SS in case of two swaps, of type RS in case of replacement and swap, and of type SR in case of swap and replacement. If the two errors are in positions i and j , with $i < j$ and we say that f has a 2-tilde-error overlap in i and j or, equivalently, that i and j are the *error positions* of the 2-tilde-error overlap.

Let $f \in \{0, 1\}^n$ have a 2-tilde-error overlap in positions i and j with $i < j$, of shift r and length $l = n - r$. The following situations can occur (see Fig.1), for some $w_1, w_2, w_3, w_4 \in \{0, 1\}^*$ and $|w_1| = r$.

$$\text{RR: } f = w_2 f[i] w f[j] w_3 w_4 = w_1 w_2 \overline{f[i]} w \overline{f[j]} w_3$$

$$\text{SR: } f = w_2 f[i] f[i+1] w f[j] w_3 w_4 = w_1 w_2 f[i+1] f[i] w \overline{f[j]} w_3$$

$$\text{RS: } f = w_2 f[i] w f[j] f[j+1] w_3 w_4 = w_1 w_2 \overline{f[i]} w f[j+1] f[j] w_3$$

$$\text{SS: } f = w_2 f[i] f[i+1] w f[j] f[j+1] w_3 w_4 = w_1 w_2 f[i+1] f[i] w f[j+1] f[j] w_3$$

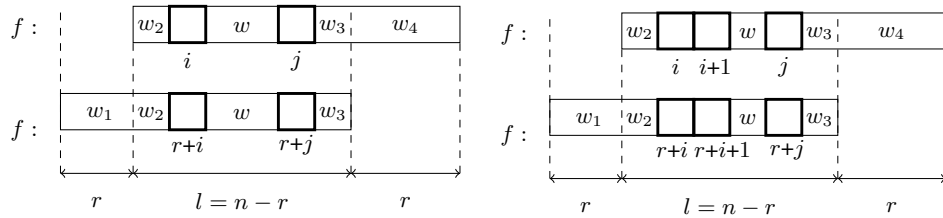


Figure 1: A word f and its 2-tilde-error overlap of type RR (a) and SR (b)

If $w = \epsilon$ we say that the two errors are *adjacent*. In particular, in the case of a 2-tilde-error overlap in positions i and j , of type RR and RS, the two errors are adjacent if $j = i + 1$. Note that in case of adjacent errors of type RR with $f[i] \neq f[i + 1]$, we have a 1-tilde-error overlap that is a swap and that we call of type S. For 2-tilde error overlap of type SR and SS in positions i and j , the two errors are adjacent if $j = i + 2$.

Remark 17 Let $f \in \{0, 1\}^n$ be a tilde-non-isometric word and (u, v) , with $u, v \in \Sigma^m$, be a pair of tilde-witnesses for f , with minimal $d = \text{dist}_{\sim}(u, v)$ among all pairs of tilde-witnesses of length m .

Let $\{O_{i_1}, O_{i_2}, \dots, O_{i_d}\}$ be the set of operations of a minimal tilde-transformation from u to v , $1 \leq i_1 < i_2 < \dots < i_d \leq m$. Then, for any $j = 1, 2, \dots, d - 1$, $O_{i_j}(u)$ has an occurrence of f in the interval $[k_j..(k_j + n - 1)]$, which contains at least one position modified by O_{i_j} . In fact, if O_{i_j} is a swap operation then it changes two positions at once, positions i_j and $i_j + 1$, and the interval $[k_j..(k_j + n - 1)]$ may contain both positions or just one. Note that when only one position is contained in the interval, such position is at the boundary of the interval. This means that, although an error in a position at the boundary of a given interval may appear as caused by a replacement, this can be actually caused by a hidden swap involving positions over the boundary.

Proposition 18 If $f \in \{0, 1\}^n$ is tilde-non-isometric then

1. either f has a 1-tilde-error overlap of type S
2. or f has a 2-tilde-error overlap.

Proof: Let f be a tilde-non-isometric word, (u, v) be a pair of tilde-witnesses for f , and $\{O_{i_1}, O_{i_2}, \dots, O_{i_d}\}$ as in Remark 17. Then, for any $j = 1, 2, \dots, d - 1$, $O_{i_j}(u)$ has an occurrence of f in the interval $[k_j..k_j + n - 1]$, which contains at least one position modified by O_{i_j} . Note that, this occurrence of f must disappear in a tilde-transformation from u to v , because v is f -free. Hence, the interval $[k_j..k_j + n - 1]$ contains a position modified by another operation in $\{O_{i_1}, O_{i_2}, \dots, O_{i_d}\}$. By the pigeonhole principle, there exist $s, t \in \{i_1, i_2, \dots, i_d\}$, such that $O_s(u)$ has an occurrence of f in $[k_s..k_s + n - 1]$ that contains at least one position modified by O_t and $O_t(u)$ has an occurrence of f in $[k_t..k_t + n - 1]$ that contains at least one position modified by O_s . Without loss of generality, suppose that $k_s < k_t$. The intersection of $[k_s..k_s + n - 1]$ and $[k_t..k_t + n - 1]$ intercepts a prefix of f in $O_t(u)$ and a suffix of f in $O_s(u)$ of some length l . Such an intersection can contain either two, or three, or four among the positions modified by O_s and O_t , of which at least one is modified by O_s and at least one by O_t .

Consider the case that the intersection of $[k_s..k_s + n - 1]$ and $[k_t..k_t + n - 1]$ contains two among the positions modified by O_s and O_t , and denote them i and j , with $1 \leq i < j \leq l$. If the positions are not adjacent, then f has a 2-tilde-error overlap (of type RR). Otherwise, if $f[i] \neq f[i + 1]$ then f has a 1-tilde-error overlap of type S. If $f[i] = f[i + 1]$ then f has a 2-tilde-error overlap (of type RR).

Suppose that the intersection of $[k_s..k_s + n - 1]$ and $[k_t..k_t + n - 1]$ contains three among the positions modified by O_s and O_t . In this case, at least one of the two operations must be a swap; suppose O_s is a swap. Then, O_t could be either a replacement on the third position, or a swap if the third position is at the boundary of $[k_t..k_t + n - 1]$. In any case, f has a 2-tilde-error overlap (of type SR or SS).

Suppose now that the intersection of $[k_s..k_s + n - 1]$ and $[k_t..k_t + n - 1]$ contains four among the positions modified by O_s and O_t . In this case, each of O_s and O_t involves two positions, and f has a 2-tilde-error overlap of type SS.

□

5 Construction of tilde-witnesses

As already discussed in Section 4, in order to prove that a word is tilde-non-isometric it is sufficient to exhibit a pair of tilde-witnesses. Proposition 18 states that if a word is tilde-non-isometric then it has either a 1-tilde-error overlap of type swap or a 2-tilde-error overlap. In this section we show the construction of tilde-witnesses for a word, starting from its error overlaps. Let us start with the case of a 1-tilde-error overlap.

Proposition 19 *If f has a 1-tilde-error overlap of type S, then it is tilde-non-isometric.*

Proof: Let f have a 1-tilde-error overlap in position i of type S with shift r . The pair (u, v) with:

$$u = \text{pre}_r(f)R_i(f) \quad v = \text{pre}_r(f)R_{i+1}(f)$$

is a pair of tilde-witnesses for f . In fact, one can prove that they satisfy the conditions in Definition 10.

□

Example 20 *The word $f = 101$ has a 1-tilde-error overlap of type S in position 1 therefore it is tilde-non-isometric. In fact, following the proof of previous proposition, the pair (u, v) with $u = 1001$ and $v = 1111$ is a pair of tilde-witnesses.*

Let us now introduce some special words which often will serve as tilde-witnesses. Let $f \in \{0, 1\}^n$ have a 2-tilde-error overlap of shift r in positions i and j , then

$$\tilde{\alpha}_r = \text{pre}_r(f)O_i(f) \text{ and } \tilde{\beta}_r = \text{pre}_r(f)O_j(f) \quad (1)$$

As an example, using the previous notations for errors of type SR we have that

$$\tilde{\alpha}_r(f) = w_1w_2f[i+1]f[i]wf[j]w_3w_4 \text{ and } \tilde{\beta}_r(f) = w_1w_2f[i]f[i+1]w\overline{f[j]}w_3w_4 \quad (2)$$

Lemma 21 *Let $f \in \{0, 1\}^n$ have a 2-tilde-error overlap of shift r , then $\tilde{\alpha}_r(f)$ is f -free.*

Proof: Suppose that $f \in \{0, 1\}^n$ has a 2-tilde-error overlap. If it is of type RR then $\tilde{\alpha}_r(f)$ is f -free by Claim 1 of Lemma 2.2 in [19], also in the case of adjacent errors. If it is of type SR, of shift r in positions i and j , with $i < j$, then, by Equation (1), we have $\tilde{\alpha}_r(f) = w_1S_i(f)$ then $\tilde{\alpha}_r[r+k] = f[k]$, for any $1 \leq k \leq n$, with $k \neq i$ and $k \neq i+1$. If f occurs in $\tilde{\alpha}_r$ in position r_1+1 we have that $1 < r_1 < r$ (if $r_1 = 1$ then $f[i] = f[i+1]$ and there is no swap error at position i) and $\tilde{\alpha}_r[r_1+1 \dots r_1+n] = f[1 \dots n]$. Finally, by Equation (2), we have that $\tilde{\alpha}_r[k] = f[k]$, for $k \neq r+j$. In conclusion, we have that $f[i] = \tilde{\alpha}_r[r_1+i] = f[r_1+i]$ (trivially, $r_1+i \neq r+j$). Furthermore $f[r_1+i] = \tilde{\alpha}_r[r+r_1+i]$ ($r_1+i \neq i$ and $r_1+i \neq i+1$ because $r_1 > 1$). But $\tilde{\alpha}_r[r+r_1+i] = f[r+i]$ then we have the contradiction that $f[i] = f[r+i]$. If the 2-tilde-error is of type RS, SS the proof is similar. For clarity, note that, also in the case of adjacent errors, supposing that f occurs in $\tilde{\alpha}_r$ leads to a contradiction in $f[i]$ that is not influenced by j .

□

Note that while $\tilde{\alpha}_r$ is always f -free, $\tilde{\beta}_r$ is not. Indeed, the property $\tilde{\beta}_r$ not f -free is related to a condition on the overlap of f . We give the following definition.

Definition 22 Let $f \in \{0, 1\}^n$ and consider a 2-tilde-error overlap of f , with shift r and error positions i, j , with $1 \leq i < j \leq n - r$. The 2-tilde-error overlap satisfies *Condition $^\sim$* if it is of type RR or SS and:

$$\begin{cases} r \text{ is even} \\ j - i = r/2 \\ f[i..(i + r/2 - 1)] = f[j..(j + r/2 - 1)] \end{cases} \quad (\text{Condition}^\sim)$$

Lemma 23 Let $f \in \{0, 1\}^n$ have a 2-tilde-error overlap of shift r , then $\tilde{\beta}_r(f)$ is not f -free iff the 2-tilde-error overlap satisfies *Condition $^\sim$* .

Proof: Suppose that $f \in \{0, 1\}^n$ has a 2-tilde-error overlap that satisfies *Condition $^\sim$* . Note that a 2-tilde-error overlap of type RS or SR cannot satisfy *Condition $^\sim$* . Now, if the 2-tilde-error overlap is of type RR, then the fact that $\tilde{\beta}_r(f)$ is not f -free can be shown as in the proof of Claim 2 of Lemma 2.2 in [19]. If the 2-tilde-error overlap is of type SS, then that proof must be suitably modified. More precisely, let i, j , with $1 \leq i < j \leq n - r$, be the error positions of the 2-tilde-error overlap of shift r that satisfies *Condition $^\sim$* .

Let $f[i] = f[j] = x$, $f[i + r] = f[j + r] = \bar{x}$, $f[i + 1] = f[j + 1] = \bar{x}$ and $f[i + 1 + r] = f[j + 1 + r] = x$.

It is possible to show that, for some $k_1, k_2 \geq 0$, we can write

$$f = \rho(uw)^{k_1} u w u w \bar{u} w \bar{u} (w \bar{u})^{k_2} \sigma$$

where $u = x\bar{x}$, $w = f[i + 2..j - 1]$ (w is empty, if $j = i + 2$) and ρ and σ are, respectively, a suffix and a prefix of w . Now, we have

$$\tilde{\beta}_r(f) = \rho(uw)^{k_1+1} u w u w \bar{u} w \bar{u} (w \bar{u})^{k_2+1} \sigma$$

and, hence, $\tilde{\beta}_r(f)$ is not f -free.

Assume now that $\tilde{\beta}_r(f)$ is not f -free and suppose that a copy of f occurs in $\tilde{\beta}_r(f)$ at position $r_1 + 1$. A reasoning similar to the one used in the proof of Lemma 21, shows that, if i and j are the error positions, then $j - i = r_1$ and $j - i = r - r_1$. Hence $r = 2r_1$ is even and $j - i = r/2$. Therefore, $f[i + t] = f^b[i + t] = f[i + t + r/2] = f[j + t]$, for $0 \leq t \leq r/2$, i.e. $f[i..(i + r/2 - 1)] = f[j..(j + r/2 - 1)]$ and the 2-tilde-error overlap satisfies *Condition $^\sim$* .

□

In the rest of the section we deal with the construction of tilde-witnesses in the case of 2-tilde-error overlaps. We distinguish the cases of non-adjacent and adjacent errors. Non-adjacent errors can be dealt with standard techniques, while the case of adjacent ones may show new issues. For example, when f has a 2-error-overlap of type SR with error block **101** (aligned with **010**) then it can be also considered of type RS.

Moreover, note that all the adjacent pairs of errors can be listed as follows, up to complement and reverse. A 2-error overlap of type SS may have (error) block **1010** or **1001**; of type SR or RS may have block **100**, **101** or **110**; of type RR block **11** (block **10** aligned with **01** corresponds to one swap). Note that, for some error types, we need

also to distinguish sub-cases related to the different characters adjacent to those error blocks. We collect all the cases in the following proposition. For lack of space, the proof is detailed only in the case 2. In the remaining cases, the proofs are sketched by exhibiting a pair of words that can be shown to be a pair of tilde-witnesses.

Theorem 24 *Let $f \in \{0, 1\}^n$. Any of the following conditions, up to complement and reverse, is sufficient for f being tilde-non-isometric.*

1. f has a 2-tilde-error overlap with not adjacent error positions
2. f has a 2-tilde-error overlap of type SS with adjacent error positions
3. f has a 2-tilde-error overlap with block **101** (of type SR or RS)
4. f has a 2-tilde-error overlap with block **100** (of type SR or RS) in the particular case that $f = x\mathbf{100}1z = yx\mathbf{011}$, for some $x, y, z \in \{0, 1\}^*$
5. f has a 2-tilde-error overlap RR in the particular case that f starts with **110** and ends with **100**.

Proof: We provide, for each case in the list, a pair of tilde-witnesses for f .

Case 1. If the 2-tilde-error overlap does not satisfy *Condition \sim* , following Definition 10, one can prove that the pair $(\tilde{\alpha}_r, \tilde{\beta}_r)$ as in Equation (1) is a pair of tilde-witnesses for f . Otherwise, one can prove that $(\tilde{\eta}_r, \tilde{\gamma}_r)$ with $\tilde{\eta}_r = \text{pre}_r(f)O_i(f)\text{suf}_{r/2}(f)$ and $\tilde{\gamma}_r = \text{pre}_r(f)O_j(O_t(f))\text{suf}_{r/2}(f)$ is a pair of tilde-witnesses for f .

Case 2. Proved in Lemma 25.

Case 3. We have $f = w_2\mathbf{101}w_3w_4 = w_1w_2\mathbf{010}w_3$, for some $w_1, w_2, w_3, w_4 \in \{0, 1\}^*$. The pair $(\tilde{\alpha}_r, \tilde{\beta}_r)$, with $\tilde{\alpha}_r = w_1w_2\mathbf{011}w_3w_4$ and $\tilde{\beta}_r = w_1w_2\mathbf{100}w_3w_4$, is a pair of tilde-witnesses, following Definition 10.

Case 4. We have $f = w_2\mathbf{1001}w_3 = w_1w_2\mathbf{011}$, for some $w_1, w_2, w_3 \in \{0, 1\}^*$. In this case we need a different technique to construct the pair of tilde-witnesses $(\tilde{\alpha}_r, \tilde{\delta}_r)$. We set $\tilde{\alpha}_r = w_1w_2\mathbf{0101}w_3$ and $\tilde{\delta}_r = w_1w_2\mathbf{1010}w_3$. Here we prove that $\tilde{\delta}_r$ is f -free. Indeed, suppose that a copy of f occurs in $\tilde{\delta}_r$ starting from position r_1 . Some considerations, related to the definition of $\tilde{\delta}_r$ and to the structure of f , show that either $r_1 = 2$ or $r_1 = 3$, and one can prove that this leads to a contradiction.

Case 5. We have $f = \mathbf{110}w_1 = w_2\mathbf{100}$, for some $w_1, w_2 \in \{0, 1\}^*$. By following Definition 10, one can prove that the pair $(\tilde{\alpha}_r, \tilde{\delta}_r)$, with $\tilde{\alpha}_r = w_2\mathbf{1010}w_1$ and $\tilde{\delta}_r = w_2\mathbf{0101}w_1$ is a pair of tilde-witnesses. Remark that, in such a case, the pair $(\tilde{\alpha}_r, \tilde{\beta}_r)$ of Equation 1 is not a pair of tilde-witnesses because $\text{dist}_{\sim}(\tilde{\alpha}_r, \tilde{\beta}_r) = 1$.

□

Let us prove in details that Case 2. of previous theorem holds.

Lemma 25 *If f has a 2-tilde-error overlap of type SS, where the errors are adjacent, then f is tilde-non-isometric.*

Proof: Let $f \in \{0, 1\}^n$ have a 2-tilde-error overlap of shift r and type SS, where the errors are adjacent. Then, two cases can occur (up to complement):

Case 1: $f = w_2\mathbf{1010}w_3w_4 = w_1w_2\mathbf{0101}w_3$, $|w_1| = r$

If the 2-tilde-error overlap does not satisfy *Condition \sim* , then $(\tilde{\alpha}_r, \tilde{\beta}_r)$, with $\tilde{\alpha}_r = w_1w_2\mathbf{0110}w_3w_4$ and $\tilde{\beta}_r = w_1w_2\mathbf{1001}w_3w_4$, is a pair of tilde-witnesses, following Definition 10. In fact:

1. $\tilde{\alpha}_r$ is f -free thanks to Lemma 21 and $\tilde{\beta}_r$ is f -free, by Lemma 23.
2. $\text{dist}_{\sim}(\tilde{\alpha}_r, \tilde{\beta}_r) = 2$, straightforward.
3. a minimal tilde-transformation from $\tilde{\alpha}_r$ to $\tilde{\beta}_r$ consists of two swaps S_i and S_j with $i = |w_1 w_2| + 1$ and $j = i + 2$. If S_i is applied to $\tilde{\alpha}_r$ as first operation, then f appears as a suffix, whereas if S_j is applied first to $\tilde{\alpha}_r$, then f appears as a prefix.

If the 2-tilde-error overlap satisfies Condition $_{\sim}$, then $w_3 = \mathbf{10}w'_3$ and, following Definition 10, $(\tilde{\eta}_r, \tilde{\gamma}_r)$ is a pair of tilde-witnesses, where $\tilde{\eta}_r = w_1 w_2 \mathbf{011010}w'_3 w_4 w_5$ and $\tilde{\gamma}_r = w_1 w_2 \mathbf{100101}w'_3 w_4 w_5$, with $w_5 = \text{suffix}_{r/2}(f)$.

1. One can prove that $\tilde{\eta}_r$ and $\tilde{\gamma}_r$ are f -free.
2. $\text{dist}_{\sim}(\tilde{\eta}_r, \tilde{\gamma}_r) = 3$.
3. a minimal tilde-transformation from $\tilde{\eta}_r$ to $\tilde{\gamma}_r$ consists of three swap operations S_i, S_j and S_t with $i = |w_1 w_2| + 1, j = i + 2, t = j + 2$.
If S_i is applied to $\tilde{\eta}_r$ as first operation, then f occurs at position $|w_1| + 1$, if S_j is applied first then f appears as a prefix, whereas if S_t is applied first then f appears as a suffix.

Case 2: $f = w_2 \mathbf{1001}w_3 w_4 = w_1 w_2 \mathbf{0110}w_3, |w_1| = r$

The pair $(\tilde{\alpha}_r, \tilde{\beta}_r)$, with $\tilde{\alpha}_r = w_1 w_2 \mathbf{0101}w_3 w_4$ and $\tilde{\beta}_r = w_1 w_2 \mathbf{1010}w_3 w_4$, is a pair of tilde-witnesses, following Definition 10. In such a case, by Lemma 23, $\tilde{\beta}_r$ is f -free. In fact, the Condition $_{\sim}$ never holds, since $f[i]$ is different from $f[j]$.

□

The following example uses Lemma 25.

Example 26 *The word $f = \mathbf{10010110} = \mathbf{10010110}$ has a 2-tilde-error overlap of type SS and shift $r = 4$ in positions 1, 3. By Lemma 25, the pair $(\tilde{\alpha}_4, \tilde{\beta}_4)$ with $\tilde{\alpha}_4 = \mathbf{100101010110}$ and $\tilde{\beta}_4 = \mathbf{100110010110}$ is a pair of tilde-witnesses. Then f is tilde-non-isometric. Note that f is Ham-isometric.*

Theorem 24 lists the conditions for a word f being tilde non-isometric and the proof provides all the corresponding pairs of witnesses. The construction of $(\tilde{\alpha}_r, \tilde{\beta}_r)$ and $(\tilde{\eta}_r, \tilde{\gamma}_r)$, used so far, is inspired by an analogous construction for the Hamming distance (cf. [19]) and it is here adapted to the tilde-distance. On the contrary, in the cases 4 and 5 a new construction is needed because the usual pair of witnesses does not satisfy any more Definition 10. The construction of $\tilde{\delta}_r$ is peculiar of the tilde-distance. It solves the situation expressed in Remark 17 when a mismatch error may appear as caused by a replacement, but it is actually caused by a hidden swap involving adjacent positions.

In conclusions, the swap and mismatch distance we adopted in this paper opens up new scenarios and presents interesting new situations that surely deserve further investigation.

References

- [1] Amihood Amir, Richard Cole, Ramesh Hariharan, Moshe Lewenstein, and Ely Porat. Overlap matching. *Inf. Comput.*, 181(1):57–74, 2003.

- [2] Amihood Amir, Estrella Eisenberg, and Ely Porat. Swap and mismatch edit distance. *Algorithmica*, 45(1):109–120, 2006.
- [3] Marcella Anselmo, Manuela Flores, and Maria Madonia. Quaternary n -cubes and isometric words. In *Combinatorics on Words*, volume 12842 of *Lect. Notes Comput. Sci.*, pages 27–39, 2021.
- [4] Marcella Anselmo, Manuela Flores, and Maria Madonia. Fun slot machines and transformations of words avoiding factors. In *11th International Conference on Fun with Algorithms*, volume 226 of *LIPICs*, pages 4:1–4:15, 2022.
- [5] Marcella Anselmo, Manuela Flores, and Maria Madonia. On k -ary n -cubes and isometric words. *Theor. Comput. Sci.*, 938:50–64, 2022.
- [6] Marcella Anselmo, Dora Giammarresi, Maria Madonia, and Carla Selmi. Bad pictures: Some structural properties related to overlaps. In *DCFS 2020*, volume 12442 of *Lect. Notes Comput. Sci.*, pages 13–25, 2020.
- [7] Marie-Pierre Béal and Maxime Crochemore. Checking whether a word is Hamming-isometric in linear time. *Theor. Comput. Sci.*, 933:55–59, 2022.
- [8] Marie-Pierre Béal, Filippo Mignosi, and Antonio Restivo. Minimal forbidden words and symbolic dynamics. In *STACS 96, 13th Annual Symposium on Theoretical Aspects of Computer Science*, volume 1046 of *Lecture Notes in Computer Science*, pages 555–566, 1996.
- [9] Giuseppa Castiglione, Sabrina Mantaci, and Antonio Restivo. Some investigations on similarity measures based on absent words. *Fundam. Informaticae*, 171(1-4):97–112, 2020.
- [10] Panagiotis Charalampopoulos, Maxime Crochemore, Gabriele Fici, Robert Mercas, and Solon P. Pissis. Alignment-free sequence comparison using absent words. *Inf. Comput.*, 262:57–68, 2018.
- [11] Yair Dombb, Ohad Lipsky, Benny Porat, Ely Porat, and Asaf Tsur. The approximate swap and mismatch edit distance. *Theor. Comput. Sci.*, 411(43):3814–3822, 2010.
- [12] Simone Faro and Arianna Pavone. An efficient skip-search approach to swap matching. *Comput. J.*, 61(9):1351–1360, 2018.
- [13] Zvi Galil and Kunsoo Park. An improved algorithm for approximate string matching. *SIAM J. Comput.*, 19:989–999, 01 1990.
- [14] W.-J. Hsu. Fibonacci cubes-a new interconnection topology. *IEEE Transactions on Parallel and Distributed Systems*, 4(1):3–12, 1993.
- [15] Aleksandar Ilić, Sandi Klavžar, and Yoomi Rho. The index of a binary word. *Theor. Comput. Sci.*, 452:100–106, 2012.
- [16] Sandi Klavžar and Sergey V. Shpectorov. Asymptotic number of isometric generalized Fibonacci cubes. *Eur. J. Comb.*, 33(2):220–226, 2012.
- [17] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Cybern Control Theory*, 10:707–710, 1966.

- [18] Robert A. Wagner. On the complexity of the extended string-to-string correction problem. In William C. Rounds *et al*, editor, *Proceedings of the 7th Annual ACM Symposium on Theory of Computing, 1975*, pages 218–223, 1975.
- [19] Jianxin Wei. The structures of bad words. *Eur. J. Comb.*, 59:204–214, 2017.
- [20] Jianxin Wei, Yujun Yang, and Xuena Zhu. A characterization of non-isometric binary words. *Eur. J. Comb.*, 78:121–133, 2019.
- [21] Jianxin Wei, Yujun Yang, and Xuena Zhu. A characterization of non-isometric binary words. *Eur. J. Comb.*, 78:121–133, 2019.
- [22] Jianxin Wei and Heping Zhang. Proofs of two conjectures on generalized Fibonacci cubes. *Eur. J. Comb.*, 51:419 – 432, 2016.