


Research Article

Word Motifs and a Generalized Hamming Distance

Pengyu Liu^{1,2*} , Jingzhou Na³

¹Department of Mathematics and Applied Mathematical Sciences, University of Rhode Island, Kingston, USA

²Department of Cell and Molecular Biology, University of Rhode Island, Kingston, USA

³Department of Mathematics, Simon Fraser University, Burnaby, Canada

E-mail: pengyu.liu@uri.edu

Received: 2 December 2024; **Revised:** 8 January 2025; **Accepted:** 9 January 2025

Abstract: Combinatorics on words is a relatively recent and rich field that involves formal grammar, algebra, geometry, fractals, algorithms, and coding, with initial research focused on repetitions in words. In this paper, we measure the differences between patterns shared by words of the same length. We introduce word motifs to represent collections of words that share the same underlying patterns, and we generalize the Hamming distance for comparing word motifs. A word motif is an equivalence class of words of the same length over an alphabet under the equivalence relation induced by symbol relabeling. We study initial problems in comparing word motifs. We compute the maximal generalized Hamming distance for k word motifs of length n over an alphabet of ℓ symbols, and we demonstrate how to calculate the exact generalized Hamming distance between a pair of word motifs.

Keywords: combinatorics on words, Hamming distance, word patterns, permutations

MSC: 68R15, 05A05

1. Introduction

A word is a sequence with elements from a finite set of symbols called the alphabet. Without loss of generality, we use the set of the first ℓ positive integers $L = \{1, 2, \dots, \ell\}$ as the alphabet and denote the set of all words over the alphabet by L^* . Initial research regarding combinatorics on words focused on repetitions in words [1]. A substitution is a mapping $h : L \rightarrow L^*$ that assigns each symbol in the alphabet to a word. A word pattern is a word $p \in L^*$. Let $w \in L^*$ be a word, and we say that the word w contains the pattern p if there exists a substitution $h : L \rightarrow L^*$ such that $h(p)$ appears in w consecutively. If we recursively apply a substitution h to a pattern p , then we have an L -system [2].

In this paper, we study a different aspect of word patterns, where we focus on words of the same length and over the same alphabet, and we restrict the substitutions to permutations in the symmetric group of order ℓ (symbol relabeling). We set these restrictions for word pattern comparison, which is a common and important task in various fields of research including image comparison in computer vision, pattern recognition in cryptography, signal comparison in coding theory, and sequence analysis in computational biology. For example, comparing word motifs of amino acid sequences allows for identifying protein structure similarities, and comparing word motifs of codes can offer insights into error correction schemes where different permutations of symbols represent equivalent states or messages.

Copyright ©2025 Pengyu Liu, et al.

DOI: <https://doi.org/10.37256/cm.6120256175>

This is an open-access article distributed under a CC BY license
(Creative Commons Attribution 4.0 International License)

<https://creativecommons.org/licenses/by/4.0/>

Here, we introduce word motifs. While the study of repetitions in words focuses on the underlying pattern of a single word, comparing word motifs emphasizes on differentiating underlying patterns shared by all possible words over an alphabet. We say that two words $w_1, w_2 \in L^*$ are equivalent if there exists a symbol relabeling substitution $h \in S_\ell$ such that $h(w_1) = w_2$, where S_ℓ is the symmetric group of order ℓ . Equivalent words share the same underlying pattern and are treated as a single object for comparison. A *word motif* is an equivalence class of words in L^* under the equivalence relation defined above. See the following section for more details. Word motifs are also related to other concepts in combinatorics. A word motif containing length- n words over an alphabet of ℓ symbols with $n \leq \ell$ can be considered as a partition of a set with n elements [3]. In particular, a word motif containing length- n words over an alphabet of n symbols is also called a rhyme scheme for an n -line stanza, and it is known that the number of rhyme schemes for n -line stanzas is the Bell number B_n [4].

Hamming distance is the main tool used to compare sequences or words [5]. The Hamming distance between a pair of words of equal length is defined to be the number of positions with different elements in the word, which is an important tool in coding theory for error detecting. As a word motif is a collection of words sharing the same underlying pattern, we generalize the Hamming distance to compare word motifs by calculating the minimum Hamming distance over all combinations of words in each word motif. In this paper, we study initial problems regarding comparing word motifs. We answer the question of what the maximal generalized Hamming distance is for all pairs of word motifs containing length- n words over an alphabet of ℓ symbols. Then, we show how to compute the exact generalized Hamming distance between a pair of word motifs.

2. Word motifs

2.1 Definitions

Let n and ℓ be positive integers and $N = \{1, 2, 3, \dots, n\}$ and $L = \{1, 2, 3, \dots, \ell\}$ be the sets of positive integers no greater than n and ℓ respectively. We call N the *index set* and an element in N an *index*. We call L the *alphabet* and an element in L a *symbol*. We define a *word* to be a function $w : N \rightarrow L$, where n is the *length* of the word, and ℓ is the *level* of the word. Note that the level of a word w is the number of possible symbols present in $w(N)$ rather than the actual number of symbols appeared in $w(N)$, and the level of a word is determined by the alphabet L instead of the image $w(N)$. We denote a word as a sequence by $w = [x_1; x_2; \dots; x_n]$, and we say that each x_i is an *element* of the word. We separate the elements of a word by semicolons to indicate that we write a word vertically as a column. In this paper, letters w, v will be used to denote words and x, y will be used to denote the elements in a word.

Let S_ℓ be the symmetric group of order ℓ . Two length- n level- ℓ words $w = [x_1; x_2; \dots; x_n]$ and $v = [y_1; y_2; \dots; y_n]$ are *equivalent* if there exists a permutation $\phi \in S_\ell$ such that $\phi(w) = \phi(x_1); \phi(x_2); \dots; \phi(x_n) = y_1; y_2; \dots; y_n = v$. For instance, among the three length-5 level-3 words in Example (1), the words w_1 and w_2 are equivalent under the permutation $\phi = (123) \in S_3$, that is $\phi(w_1) = w_2$, while no permutation can map w_1 or w_2 to w_3 , so w_3 is not equivalent to w_1 or w_2 . Here, the permutation $\phi = (123)$ is written in cycle notation, and we use the cycle notation for permutations throughout the paper unless otherwise stated.

$$w_1 = \begin{bmatrix} 1 \\ 1 \\ 3 \\ 2 \\ 1 \end{bmatrix} \quad w_2 = \begin{bmatrix} 2 \\ 2 \\ 1 \\ 3 \\ 2 \end{bmatrix} \quad w_3 = \begin{bmatrix} 1 \\ 1 \\ 2 \\ 2 \\ 1 \end{bmatrix} \quad (1)$$

Let $\mathcal{W}_{n, \ell}$ be the set of all length- n level- ℓ words. The equivalence relation induced by permutations in S_ℓ partitions $\mathcal{W}_{n, \ell}$. We define a *word motif* to be an equivalence class of $\mathcal{W}_{n, \ell}$. We denote a word motif by t or $\langle w \rangle$, where w is a word in the word motif t . We say that the word w is a *representative* of the word motif t , and that the word motif t is *generated* by w . We define the *length* of a word motif $\langle w \rangle$ to be the length of w , and similarly, the *level* of the word motif $\langle w \rangle$ to be

the level of w . We denote the set of all length- n level- ℓ word motifs by $\mathcal{T}_{n, \ell}$. For instance, the six words in Example (2) form a length-3 level-3 word motif.

$$w_1 = \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix} \quad w_2 = \begin{bmatrix} 1 \\ 1 \\ 3 \end{bmatrix} \quad w_3 = \begin{bmatrix} 2 \\ 2 \\ 1 \end{bmatrix} \quad w_4 = \begin{bmatrix} 2 \\ 2 \\ 3 \end{bmatrix} \quad w_5 = \begin{bmatrix} 3 \\ 3 \\ 1 \end{bmatrix} \quad w_6 = \begin{bmatrix} 3 \\ 3 \\ 2 \end{bmatrix} \quad (2)$$

We say that a word is *constant* if all of its elements are identical, and a word motif is *constant* if it contains a constant word.

2.2 Enumeration

We count the number of length- n level- ℓ word motifs and denote the number by $|\mathcal{T}_{n, \ell}|$. We say that a length- n level- ℓ word $w : N \rightarrow L$ is in *standard order* if for any $i \in N$ with $w(i) > 1$, there exists a $j < i \in N$ such that $w(j) = w(i) - 1$. We say such a word w is a *standard word* of length n and level ℓ . For instance, only w_3 among words in Example (1) is a standard word, and only w_1 among words in Example (2) is a standard word. It is easy to show that every length- n level- ℓ word can be mapped to a standard word by a permutation in S_ℓ , and that there exists no permutation in S_ℓ that can map one standard word to a different standard word. Hence, every length- n level- ℓ word motif contains one and only one standard word, and counting the number of length- n level- ℓ word motifs is equivalent to counting the number of standard words of length n and level ℓ . We denote the set of all length- n level- ℓ standard words by $W_{n, \ell}^*$ and the number of length- n level- ℓ standard words by $|W_{n, \ell}^*|$. We have $|\mathcal{T}_{n, \ell}| = |W_{n, \ell}^*|$, and the number $|W_{n, \ell}^*|$ is computed using Stirling numbers of the second kind [6].

Theorem 1 (Arndt-Sloane [6]) The number of length- n level- ℓ words that are in standard order is given by Formula (3).

$$|W_{n, \ell}^*| = \sum_{m=1}^{\ell} \sum_{i=0}^m \frac{(-1)^i (m-i)^n}{i! (m-i)!} \quad (3)$$

Here, we present an alternative enumeration from the perspective of word motifs. We note that a word motif $t \in_{n, \ell}$ can be considered as the orbit of a word $w \in t$ under the action of the symmetric group S_ℓ on individual elements of w . Thus, we can compute $|\mathcal{T}_{n, \ell}|$ with the orbit-counting theorem, also known as Cauchy-Frobenius lemma or Burnside's lemma [7].

Theorem 2 The number of length- n level- ℓ word motifs is given by Formula (4). In particular, we have $|\mathcal{T}_{n, 2}| = 2^{n-1}$.

$$|\mathcal{T}_{n, \ell}| = \frac{\ell^n}{\ell!} + \sum_{m=2}^{\ell} \frac{(\ell-m)^n}{(\ell-m)!} \sum_{i=0}^m \frac{(-1)^i}{i!} \quad (4)$$

Proof. The orbit-counting theorem states that the number of orbits can be computed with Formula (5), where $\mathcal{W}_{n, \ell}^\phi = w \in \{\mathcal{W}_{n, \ell} \mid \phi(w) = w\}$ is the set of fixed words by the permutation $\phi \in S_\ell$ [7].

$$|\mathcal{T}_{n, \ell}| = \frac{1}{|S_\ell|} \sum_{\phi \in S_\ell} |\mathcal{W}_{n, \ell}^\phi| \quad (5)$$

The identity in S_ℓ fixes all words in $\mathcal{W}_{n, \ell}$, so the number of words fixed by the identity is ℓ^n . Let $S_{\ell|m} \subset S_\ell$ be the subset of permutations that derange m symbols, where $2 \leq m \leq \ell$. The number of permutations in $S_{\ell|m}$ can be computed by counting the number of ways of selecting m symbols from the symbol set L and multiplying by the number of derangements of the m symbols. See Formula (6).

$$S_{\ell|m} = \binom{\ell}{m} \left(m! \sum_{i=0}^m \frac{(-1)^i}{i!} \right) = \frac{\ell!}{(\ell-m)!} \sum_{i=0}^m \frac{(-1)^i}{i!} \quad (6)$$

We note that a permutation $\phi \in S_{\ell|m}$ fixes all and only words without the m symbols that ϕ deranges, and there are $(\ell-m)^n$ such words in $\mathcal{W}_{n, \ell}^\phi$. Therefore, we have Formula (7).

$$\sum_{\phi \in S_\ell} |\mathcal{W}_{n, \ell}^\phi| = \ell^n + \ell! \sum_{m=2}^{\ell} \frac{(\ell-m)^n}{(\ell-m)!} \sum_{i=0}^m \frac{(-1)^i}{i!} \quad (7)$$

Then Formula (4) follows applying Formula (7) to Formula (5). □

Recall that when $n \leq \ell$, there is an one-to-one correspondence between word motifs and partitions of set with n elements. So, we have $|\mathcal{T}_{n, \ell}| = B_n$ for any $n \leq \ell$, where B_n is the n -th Bell number. If we set $n = \ell$, then Formula (4) is equivalent to Dobiński's formula for Bell numbers [8]. When $n > \ell$, the number $|\mathcal{T}_{n, \ell}|$ equals the number of partitions of an n -element set with at most ℓ subsets.

3. Hamming distance of word motifs

3.1 Basic definitions

Let $W = \{w_1, w_2, \dots, w_k\} \subset \mathcal{W}_{n, \ell}$ be a set of words with $k \geq 2$. Recall that a word is a function from the index set $N = \{1, 2, 3, \dots, n\}$ to the alphabet $L = \{1, 2, 3, \dots, \ell\}$. We define the *Hamming distance* (or simply the *distance*) of words in W to be the number of indices in N whose images in w_1, w_2, \dots, w_k are not identical. We denote the distance of the words in W by $d(W)$ or $d(w_1, w_2, \dots, w_k)$. For instance, the distance of the three length-5 level-3 words in Example (8) is $d(w_1, w_2, w_3) = 4$, because only the fourth elements in the words are identical.

$$w_1 = \begin{bmatrix} 1 \\ 1 \\ 3 \\ 2 \\ 1 \end{bmatrix} \quad w_2 = \begin{bmatrix} 3 \\ 3 \\ 1 \\ 2 \\ 3 \end{bmatrix} \quad w_3 = \begin{bmatrix} 1 \\ 1 \\ 2 \\ 2 \\ 1 \end{bmatrix} \quad (8)$$

Note that the order of words in W is irrelevant in computing the distance. So to simplify our arguments, we fix the order of words and write $W = [w_1, w_2, \dots, w_k]$ as a sequence of words. Here, we separate the words in W by colons to indicate that we list the words horizontally. For example, in computing the distance of the three words w_1, w_2 and w_3 in Example (8), we write the set W of the three words as in Example (9).

$$W = [w_1, w_2, w_3] = \begin{bmatrix} 1 & 3 & 1 \\ 1 & 3 & 1 \\ 3 & 1 & 2 \\ 2 & 2 & 2 \\ 1 & 3 & 1 \end{bmatrix} \quad (9)$$

Let $K = \{1, 2, \dots, k\}$ be the set of k positive integers. We assume $k \geq 2$ throughout the paper unless otherwise stated. We define a *cross section* of $W = [w_1, w_2, \dots, w_k]$ to be a word $c_i : K \rightarrow L$ consisting of the i -th elements of the words in W , that is $c_i = [w_1(i), w_2(i), \dots, w_k(i)]$. We denote the set of length- k level- ℓ cross sections by $\mathcal{C}_{k, \ell}$, to distinguish the horizontally written words (cross sections) from the vertically written words in $\mathcal{W}_{k, \ell}$. Similarly, we say a cross section is *constant* if all of its elements are identical. Let $s(W)$ be the number of constant cross sections in W . We have $d(W) = n - s(W)$. For instance, the words $c_1 = [1, 3, 1]$, $c_2 = [1, 3, 1]$, $c_3 = [3, 1, 2]$, $c_4 = [2, 2, 2]$ and $c_5 = [1, 3, 1]$ are cross sections of W displayed in Example (9), and c_4 is the only constant cross section in W , so $s(W) = 1$ and $d(W) = 4$.

Let $T = \{t_1, t_2, \dots, t_k\} \subset \mathcal{T}_{n, \ell}$ be a set of word motifs. Recall that a word motif t is an equivalence class of words under permutations, and t can be denoted by w with a representative $w \in t$. We define the *Hamming distance* (or simply the *distance*) of word motifs in T by Formula (10) and denote the distance by $d(T)$ or $d(t_1, t_2, \dots, t_k)$.

$$d(t_1, t_2, \dots, t_k) = \min_{w_i \in t_i} d(w_1, w_2, \dots, w_k) \quad (10)$$

Suppose t_1, t_2, \dots, t_k are respectively generated by words w_1, w_2, \dots, w_k in $\mathcal{W}_{n, \ell}$, and we write the set $W = [w_1, w_2, \dots, w_k]$ as a sequence. We can analogously write the set of word motifs as $T = \langle W \rangle = [\langle w_1 \rangle, \langle w_2 \rangle, \dots, \langle w_k \rangle] = [t_1, t_2, \dots, t_k]$. The distance of $T = t_1, t_2, \dots, t_k$ can also be defined using permutations. Let $\Phi = [\phi_1, \phi_2, \dots, \phi_k] \in S_\ell^k$ be a sequence of permutations, where S_ℓ^k is the Cartesian product of k symmetric groups of order ℓ . We define $\Phi(W) = [\phi_1(w_1), \phi_2(w_2), \dots, \phi_k(w_k)]$ and the distance for a set $T = \langle W \rangle$ of word motifs by Formula (11).

$$d(\langle W \rangle) = \min_{\Phi \in S_\ell^k} d(\Phi(W)) \quad (11)$$

Note that the distance of W can be computed by counting the number of constant cross sections in W . So the distance of $T = \langle W \rangle$ can also be computed by Formula (12).

$$d(\langle W \rangle) = n - \max_{\Phi \in S_\ell^k} s(\Phi(W)) \quad (12)$$

For instance, the sequence of permutations $\Phi = [(1), (13), (1)]$ maps W in Example (9) to $\Phi(W)$ in Example (13), and there are 4 constant cross sections in $\Phi(W)$. It is easy to check that the maximal number of constant cross sections in $\Phi(W)$ is 4 for any $\Phi \in S_\ell^k$. So the distance for the set $\langle W \rangle$ of word motifs is 1.

$$\Phi(W) = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 3 & 3 & 2 \\ 2 & 2 & 2 \\ 1 & 1 & 1 \end{bmatrix} \quad (13)$$

3.2 Metric spaces of word motifs

It is well known that the Hamming distance between two words in $\mathscr{W}_{n, \ell}$ is a metric. Namely, the Hamming distance has the following three properties.

1. Identity: For any words $w, v \in \mathscr{W}_{n, \ell}$, we have $d(w, v) = 0$ if and only if $w = v$.
2. Symmetry: For any words $w, v \in \mathscr{W}_{n, \ell}$, we have $d(w, v) = d(v, w)$;
3. Triangle inequality: For any words $w_1, w_2, w_3 \in \mathscr{W}_{n, \ell}$, the inequality $d(w_1, w_3) + d(w_2, w_3) \geq d(w_1, w_2)$ holds.

It is trivial to check that the generalized Hamming distance of two word motifs in $\mathscr{T}_{n, \ell}$ satisfies the first two properties. Here, we show that the triangle inequality holds for word motifs. Consider three word motifs $t_1 = \langle w_1 \rangle$, $t_2 = \langle w_2 \rangle$ and $t_3 = \langle w_3 \rangle$ in $\mathscr{T}_{n, \ell}$. Suppose that $d(t_1, t_3) = d(\phi_1(w_1), w_3)$ for a permutation $\phi_1 \in S_\ell$ and that $d(t_2, t_3) = d(\phi_2(w_2), w_3)$ for a permutation $\phi_2 \in S_\ell$. We have $d(t_1, t_3) + d(t_2, t_3) = d(\phi_1(w_1), w_3) + d(\phi_2(w_2), w_3) \geq d(\phi_1(w_1), \phi_2(w_2))$. Note that $d(t_1, t_2)$ is defined to be the minimal distance of $d(\phi_a(w_1), \phi_b(w_2))$ over all $\phi_a, \phi_b \in S_\ell$. We have $d(\phi_1(w_1), \phi_2(w_2)) \geq d(t_1, t_2)$ and $d(t_1, t_3) + d(t_2, t_3) \geq d(t_1, t_2)$. Therefore, we have the following theorem.

Theorem 3 The generalized Hamming distance defined for a pair of word motifs in $\mathscr{T}_{n, \ell}$ is a metric, and all word motifs in $\mathscr{T}_{n, \ell}$ together with the generalized Hamming distance for two word motifs form a metric space.

3.3 Bounds of the generalized Hamming distance

It is clear that the minimal distance for a set of word motifs is 0. The *maximal distance* of k word motifs in $\mathscr{T}_{n, \ell}$, denoted by $D_{n, \ell, k}$, is defined by Formula (14).

$$D_{n, \ell, k} = \max_{t_i \in \mathscr{T}_{n, \ell}} d(t_1, t_2, \dots, t_k) \quad (14)$$

For any set $W = [w_1, w_2, \dots, w_k]$ of words in $\mathscr{W}_{n, \ell}$, we can always apply a sequence of permutations $\Phi \in S_\ell^k$ to W such that every word in $\Phi(W)$ has the same first element. Hence, the maximal distance $D_{n, \ell, k}$ has a trivial upper bound.

Lemma 1 For any positive integers n, ℓ and $k \geq 2$, we have $D_{n, \ell, k} \leq n - 1$.

We characterize the sets of k word motifs in $\mathscr{T}_{n, \ell}$ that have distance $n - 1$. Let $c = [a_1, a_2, \dots, a_k]$ and $c' = [b_1, b_2, \dots, b_k]$ be two cross sections in $\mathscr{C}_{k, \ell}$. We say that c and c' are *identical* if $a_i = b_i$ for all $1 \leq i \leq k$. We say that c and c' are *incompatible* if $a_i \neq b_i$ for all $1 \leq i \leq k$. We say that c and c' are *connected* (by a sequence of permutations) if they are either identical or incompatible.

Lemma 2 Let $T = \langle W \rangle = [\langle w_1 \rangle, \langle w_2 \rangle, \dots, \langle w_k \rangle]$ be a set of k word motifs in $\mathscr{T}_{n, \ell}$. Then $d(T) = n - 1$ if and only if W contains no pair of connected cross sections.

Proof. If there are two identical cross sections $c = c' = [a_1, a_2, \dots, a_k]$ in W , then the sequence of permutations $\Phi^* = [(a_1), (a_1 a_2), \dots, (a_1 a_k)]$ creates two constant cross sections in $\Phi^*(W)$. So, we have $d(T) < n - 1$, contradicting the assumption. If there are two incompatible cross sections $c = [a_1, a_2, \dots, a_k]$ and $c' = [b_1, b_2, \dots, b_k]$ in W , then the sequence of permutations $\Phi^* = [\phi_1^*, \phi_2^*, \dots, \phi_k^*]$ creates two constant cross sections in $\Phi^*(W)$, where ϕ_i^* is given by the two-line notation in Formula (15).

$$\phi_i^* = \begin{pmatrix} a_i & b_i & \cdots \\ a_1 & b_1 & \cdots \end{pmatrix} \quad (15)$$

Note that c and c' are incompatible, so we have $a_1 \neq b_1$ and $a_i \neq b_i$, which guarantee that the sequence of permutations Φ^* is well defined. Similarly, two constant cross sections in $\Phi(W)$ imply $d(T) < n - 1$, which contradicts the assumption.

Conversely, assume $d(T) < n - 1$, then there exists a pair of cross sections $c = [a_1, a_2, \dots, a_k]$ and $c' = [b_1, b_2, \dots, b_k]$ in W such that $\Phi(c)$ and $\Phi(c')$ are constant in $\Phi(W)$ for a sequence of permutations $\Phi = [\phi_1, \phi_2, \dots, \phi_k] \in S_\ell^k$. We claim that c and c' are either identical or incompatible. If they are not identical or incompatible, then there exists an index $i \in N$ such that $a_i = b_i$, and there also exists a different index $j \in N$ such that $a_j \neq b_j$. Since the cross sections $\Phi(c)$ and $\Phi(c')$ are constant in $\Phi(W)$, we have $\phi_i(a_i) = \phi_j(a_j)$ and $\phi_i(b_i) = \phi_j(b_j)$. Furthermore, because $a_i = b_i$, we have $\phi_i(a_i) = \phi_i(b_i)$. These imply that $\phi_j(a_j) = \phi_j(b_j)$, which contradicts $a_j \neq b_j$. \square

The proof of Lemma 2 can be generalized for a set of pairwise connected cross sections.

Lemma 3 Let $W = [w_1, w_2, \dots, w_k]$ be a set of k words in $\mathcal{W}_{n, \ell}$. If W contains m pairwise connected cross sections, then there exists a sequence of permutations $\Phi^* \in S_\ell^k$ such that the m pairwise connected cross sections are mapped to m constant cross sections by Φ^* .

Proof. For any $1 \leq i \leq m$, let $c^i = [a_1^i, a_2^i, \dots, a_k^i]$ be one of the m pairwise connected cross sections in W . Since connected cross sections can be identical, we assume that there are $u \leq m$ unique cross sections among them. Without loss of generality, let c^1, c^2, \dots, c^u be the unique cross sections. Because they are pairwise incompatible, the elements $a_j^1, a_j^2, \dots, a_j^u$ are u distinct symbols for any $1 \leq j \leq k$. We define a sequence of permutations $\Phi^* = [\phi_1^*, \phi_2^*, \dots, \phi_k^*] \in S_\ell^k$ by the two-line notation in Formula (16)

$$\phi_j^* = \begin{pmatrix} a_j^1 & a_j^2 & \dots & a_j^{u-1} & a_j^u & \dots \\ a_1^1 & a_1^2 & \dots & a_1^{u-1} & a_1^u & \dots \end{pmatrix} \quad (16)$$

Note that c^1, c^2, \dots, c^u being pairwise incompatible implies that the sequence of permutations Φ^* is well defined. It is trivial to check that the m connected cross sections are mapped to m constant cross sections by Φ^* . \square

We examine the connectedness of cross sections in $\mathcal{C}_{k, \ell}$. There are in total ℓ^k cross sections in $\mathcal{C}_{k, \ell}$. We divide $\mathcal{C}_{k, \ell}$ into ℓ subsets based on their first elements, and we denote the subset of cross sections with first element i by $\mathcal{C}_{k, \ell|i}$. Let c be a cross section in $\mathcal{C}_{k, \ell|i}$ and c' be a cross section in $\mathcal{C}_{k, \ell|i+1}$. We say that c is *linked* to c' if $c' = \psi(c)$, where $\psi = (12 \dots \ell) \in S_\ell$. Let c be a cross section in $\mathcal{C}_{k, \ell|1}$. We define a *link* $\mathcal{L}(c)$ generated by c to be a subset of $\mathcal{C}_{k, \ell}$ such that $\mathcal{L}(c) = c, \psi(c), \psi^2(c), \dots, \psi^{\ell-1}(c)$. For instance, we display a link of $\mathcal{C}_{5, 3}$ in Example (17).

$$c = [1, 2, 3, 1, 1] \quad \psi(c) = [2, 3, 1, 2, 2] \quad \psi^2(c) = [3, 1, 2, 3, 3] \quad (17)$$

It is clear that for any cross section $c \in \mathcal{C}_{k, \ell|1}$, every subset $\mathcal{C}_{k, \ell|i}$ has one and only one element in $\mathcal{L}(c)$. It is also trivial that for different cross sections $c, c' \in \mathcal{C}_{k, \ell|1}$, we have $\mathcal{L}(c) \cap \mathcal{L}(c') = \emptyset$. So the links partition $\mathcal{C}_{k, \ell}$, and there are ℓ^{k-1} links in $\mathcal{C}_{k, \ell}$. Moreover, since ψ increases every symbol by 1 in \mathbb{Z}_ℓ (where we set $0 = \ell$), every pair of cross sections in a link is incompatible, hence connected.

We say that a set W of k words in $\mathcal{W}_{n, \ell}$ is *complete* if $\Phi(W)$ contains one and only one constant cross section for any sequence of permutations $\Phi \in S_\ell^k$. We say that W is *semi-complete* if $\Phi(W)$ contains at most one constant cross section for any sequence of permutations $\Phi \in S_\ell^k$. Note that there exist totally ℓ^{k-1} cross sections in $\mathcal{C}_{k, \ell|1}$. We list these cross sections in lexicographic order of the elements in them and construct a set $M = [w_1^*, w_2^*, \dots, w_k^*]$ of words in $\mathcal{W}_{\ell^{k-1}, \ell}$ such

that the cross sections in M are the cross sections from $\mathcal{C}_{k, \ell|1}$ in lexicographic order. See Formula (18). For a positive integer $r < \ell^{k-1}$, we define M_r to be the set of k words in $\mathcal{W}_{r, \ell}$ that contains only the first r cross sections in M .

$$M = w_1^*, w_2^*, \dots, w_k^* = \begin{bmatrix} 1 & 1 & \dots & 1 & 1 \\ 1 & 1 & \dots & 1 & 2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 1 & \dots & 1 & \ell \\ 1 & 1 & \dots & 2 & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & \ell & \dots & \ell & \ell \end{bmatrix} \quad (18)$$

Lemma 4 The set M of k words in $\mathcal{W}_{\ell^{k-1}, \ell}$ is complete, and the set M_r of k words in $\mathcal{W}_{r, \ell}$ is semi-complete.

Proof. Because w_1^* is a constant word and there is no identical cross sections in M or M_r , no pair of cross sections in M or M_r is connected. By Lemma 1 and Lemma 2, we have $d(\langle M \rangle) = \ell^{k-1} - 1$, and $d(\langle M_r \rangle) = r - 1$. So M and M_r are semi-complete.

Let $M' = [w_2^*, w_3^*, \dots, w_k^*]$ be a set of $k - 1$ words in $\mathcal{W}_{\ell^{k-1}, \ell}$. The cross sections in M' contain all combinations of assigning ℓ symbols to $k - 1$ positions. So there are ℓ unique constant cross sections in M' , namely $[1, 1, \dots, 1]$, $[2, 2, \dots, 2]$, ..., $[\ell, \ell, \dots, \ell]$. For any sequence of permutations $\Phi' = [\phi_2, \phi_3, \dots, \phi_k] \in S_\ell^{k-1}$, these constant cross sections persist in $\Phi'(M')$. So for any sequence of permutations $\Phi = [\phi_1, \phi_2, \dots, \phi_k]$ in S_ℓ^k , no matter what $\phi_1(1) \in L$ in $\Phi(M)$ is, there always exists a constant cross section in $\Phi(M)$. Therefore, M is complete. \square

Theorem 4 The maximal distance of k word motifs of length n and level ℓ is given by Formula (19).

$$D_{n, \ell, k} = \left\lceil n - \frac{n}{\ell^{k-1}} \right\rceil \quad (19)$$

Proof. For any set W of k words in $\mathcal{W}_{n, \ell}$, there exists a link $\mathcal{L}(c)$ generated by a cross section $c \in \mathcal{C}_{k, \ell|1}$ such that W has at least $\lceil n/\ell^{k-1} \rceil$ cross sections of $\mathcal{L}(c)$ due to the pigeonhole principle. These cross sections in $\mathcal{L}(c)$ are connected, so they are constant cross sections in $\Phi(W)$ for a sequence of permutations $\Phi \in S_\ell$ by Lemma 3. Therefore, we have $s(\Phi(W)) \geq \lceil n/\ell^{k-1} \rceil$ and $D_{n, \ell, k} \leq n - \lceil n/\ell^{k-1} \rceil$.

Suppose that $n = m\ell^{k-1} + r$ for integers $m \geq 0$ and $0 \leq r < \ell^{k-1}$. Consider the set $M_n = [M; M; \dots; M; M_r]$ of k words in $\mathcal{W}_{n, \ell}$ constructed by vertically concatenating m copies of M and one copy of M_r . By Lemma 4, the set M is complete and the set M_r is semi-complete. So, there are at most $m + 1$ constant cross sections in $\Phi(M_n)$ for any sequence of permutations $\Phi \in S_\ell^k$. Therefore, we have $d(\langle M_n \rangle) = n - (m + 1) = n - \lceil n/\ell^{k-1} \rceil$ and $D_{n, \ell, k} = n - \lceil n/\ell^{k-1} \rceil$. \square

3.4 Computing exact generalized Hamming distances

Let $T = \langle W \rangle = [\langle w_1 \rangle, \langle w_2 \rangle, \dots, \langle w_k \rangle]$ be a set of word motifs in $\mathcal{T}_{n, \ell}$ and $C(W) = [c_1; c_2; \dots; c_n]$ be the sequence of all cross sections in W . We say that a subset $X(W) \subset C(W)$ is *maximally connected* if all the cross sections in $X(W)$ are pairwise connected, and any cross section in $C(W) - X(W)$ is connected to some cross section in $X(W)$. We use $\mathcal{X}_i(W)$ to denote the set of all maximally connected subsets of $C(W)$ that contain c_i , and we use $\mu(\mathcal{X}_i(W))$ to denote the maximum cardinality of elements in $\mathcal{X}_i(W)$. Lemma 2 implies that two cross sections that are not connected can not both be mapped to constant cross sections by any sequence of permutations $\Phi \in S_\ell^k$. So, the minimum number of cross sections in $C(W) - X(W)$ over all maximally connected subset $X(W) \subset C(W)$ gives the exact distance $d(\langle W \rangle)$, and we have the following Lemma 5, where Formula (20) can be used to compute the exact distance.

Lemma 5 Let $W = [w_1, w_2, \dots, w_k]$ be a set of k words in $\mathcal{W}_{n, \ell}$.

$$d(\langle W \rangle) = n - \max_{\Phi \in \mathcal{S}_\ell^k} s(\Phi(W)) = n - \max_{1 \leq i \leq n} \mu(\mathcal{X}_i(W)) \quad (20)$$

Let $T = \langle W \rangle = [\langle w_1 \rangle, \langle w_2 \rangle]$ be a set of two word motifs in $\mathcal{T}_{n, \ell}$ and $C(W) = [c_1; c_2; \dots; c_n]$ be the sequence of all cross sections in W . To compute the exact distance of the word motifs in $\langle W \rangle$, we construct an $\ell \times \ell$ matrix A_W as follows. For any cross section $c = [a, b] \in \mathcal{C}_{2, \ell}$ where $a, b \in L$, if c appears m times in $C(W)$, then the entry of A_W at row a and column b is $A_W(a, b) = m$. Moreover, any sequence of permutations $\Phi = [\phi_1, \phi_2] \in \mathcal{S}_\ell^2$ can be written as a permutation matrix P_Φ , where ϕ_1 permutes rows and ϕ_2 permutes columns. Note that the diagonal entries of A_W record the number of constant cross sections in W . Thus, Formula (12) can be written as Formula (21).

$$d(\langle W \rangle) = n - \max_{\Phi \in \mathcal{S}_\ell^2} \text{tr}(P_\Phi A_W) \quad (21)$$

This is equivalent to the linear assignment problem, and we can use the Hungarian algorithm or Kuhn-Munkres algorithm to compute the exact distance between two word motifs in polynomial time [9, 10]. In general, computing the exact distance of a set $T = \langle W \rangle = \langle w_1 \rangle, \langle w_2 \rangle, \dots, \langle w_k \rangle$ of word motifs in $\mathcal{T}_{n, \ell}$ is equivalent to the k -dimensional assignment problem, and algorithms to solve the assignment problem, for example in [11], can be used to compute the exact distance of word motifs.

4. Discussion

We have formally introduced word motifs. A word motif is a collection of all words of the same length that share the same underlying pattern. This paper focused on comparing word motifs, where we generalized the Hamming distance and found the maximal generalized Hamming distance for k word motifs of length n over an alphabet of ℓ symbols to be $n - \lceil n/\ell^{k-1} \rceil$. This is the upper bound of the generalized Hamming distance for all sets of k length- n level- ℓ word motifs. The exact generalized Hamming distance between two word motifs can be computed using the Hungarian algorithm or Kuhn-Munkres algorithm, where the computational complexity increases linearly to the length of word n , cubically to the size of the alphabet ℓ , and exponentially to the number of word motifs k . If we need to compare two word motifs over different alphabets, we only need to consider the union of the alphabets as the new alphabet.

There remains considerable scope to modify the definitions of word motifs and the distances, and to ask various questions. In general, words of different lengths can also share the same underlying pattern, and our methods can be generalized to compare word motifs of different lengths. It is interesting to investigate how classic concepts in the fields such as subword complexity, pattern avoidance, or automaton-based word matching can be generalized for word motifs, and how methods regarding word motifs can help the study of these concepts. Regarding the generalized Hamming distance, it is also interesting to investigate its relationship to other distance measures in combinatorics, and its behavior under additional constraints such as all symbols in the alphabet must be present in a word or adjacent element in a word must have different symbols.

Acknowledgement

P.L. was supported by the start-up funds of the University of Rhode Island, and J.N. was supported by a doctoral scholarship from the China Scholarship Council. The authors thank anonymous reviewers for their helpful comments and suggestions.

Conflict of interest

The authors declare no competing financial interest

References

- [1] Berstel J, Perrin D. The origins of combinatorics on words. *European Journal of Combinatorics*. 2007; 28(3): 996-1022.
- [2] Lindenmayer A. Mathematical models for cellular interactions in development II. Simple and branching filaments with two-sided inputs. *Journal of Theoretical Biology*. 1968; 18(3): 300-315.
- [3] Rota GC. The number of partitions of a set. *The American Mathematical Monthly*. 1964; 71: 498-504.
- [4] Riordan J. A budget of rhyme scheme counts. *Annals of the New York Academy of Sciences*. 1979; 319(1): 455-465.
- [5] Hamming RW. Error detecting and error correcting codes. *The Bell System Technical Journal*. 1950; 29(2): 147-160.
- [6] Arndt J, Sloane NJA. *Counting Words that are in "Standard Order"*. 2016. Available from: <https://oeis.org/A278984/a278984.txt> [Accessed 30 November 2024].
- [7] Gallian J. *Contemporary Abstract Algebra*. 7th ed. Belmont, CA: Brooks/Cole, Cengage Learning; 2010.
- [8] Pitman J. Some probabilistic aspects of set partitions. *The American Mathematical Monthly*. 1997; 104(3): 201-209.
- [9] Kuhn H. The Hungarian method for the assignment problem. *Naval Research Logistics*. 1955; 2(1-2): 83-97.
- [10] Munkres J. Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics*. 1957; 5(1): 32-38.
- [11] Li J, Kirubarajan T, Tharmarasa R, Brown D, Pattipati K. A dual approach to multi-dimensional assignment problems. *Journal of Global Optimization*. 2021; 81(3): 691-716.