# The De Bruijn Graph Sequence Mapping Problem with Changes in the Graph

Lucas B. Rocha* Said Sadique Adi* Eloi Araujo*
*Faculdade de Computação*
*Universidade Federal de Mato Grosso de Sul*
Campo Grande, MS, Brazil
*lucas.lb.rocha@gmail.com {said.sadique, francisco.araujo}@ufms.br
*0000-0002-9606-2569 0000-0002-0349-7441 0000-0002-8015-6237

***Abstract*—In computational biology, mapping a sequence $s$ onto a sequence graph $G$ is a significant challenge. One possible approach to addressing this problem is to identify a walk $p$ in $G$ that spells a sequence which is most similar to $s$. This problem is known as the Graph Sequence Mapping Problem (`GSMP`). In this paper, we study an alternative problem formulation, namely the De Bruijn Graph Sequence Mapping Problem (`BSMP`). We focused on addressing the problem involving changes in the graph. We reformulated the problem, taking into account the characteristics of the arcs induced in the De Bruijn graph. This reformulation led to a modification in the problem definition, allowing the application of a polynomial-time algorithm for its resolution.**

***Index Terms*—De Bruijn Graph, Mapping, Sequences**

## I. Introduction

A very relevant task in computational biology is to map a sequence onto another for comparison purposes. Typically, one sequence is compared to a reference sequence, which is considered a high-quality sequence representing a set of sequences [9], [10]. On the other hand, the reference sequence is biased as it represents only a limited set of sequences and it is unable to account for all possible variations. One way to overcome this bias is to represent multiple sequences as another robust structure [6], such as the sequence graph or De Bruijn graph [7], [13], [14].

The *sequence graph* is a graph such that each node is labeled with one or more characters and the *simple sequence graph* is one where each node is labeled with exactly one character [7]. In the *De Bruijn graph* [13], [14] of order $k$, each node is labeled with a distinct sequence of length $k$ and there is an arc from one node to another if and only if there is an overlap of length $k-1$ from the suffix of the first to the prefix of second.

Informally, a walk $p$ in a graph $G$ is a sequence of connected nodes by arcs. Given a sequence graph $G$, a walk $p$ in $G$ can spell a sequence $s'$ by concatenating the characters associated with each node of $p$. The *Graph Sequence Mapping Problem* – `GSMP` consists of finding a walk $p$ in a sequence graph $G$ that spells a sequence as similar as possible to a given sequence $s$.

One of the first articles that addresses `GSMP` in more details was written by Amir *et. al*. in the article entitled *Pattern Matching in Hypertext* [1]. Navarro improved the results of this article and detailed these improvements in the article entitled *Improved Approximate Pattern Matching on Hypertext* [8]. For the approximate mapping, Amir *et. al*. were the first authors in the literature who identified an asymmetry in the location of the changes, showing the importance of understanding whether changes happen only in the pattern, only in the hypertext or in both. Considering the asymmetry identified by Amir *et. al*., the `GSMP` allows three variants:

1) allows changes only in the pattern when mapping the pattern in hypertext;
2) allows changes only in the hypertext when mapping the pattern in hypertext;
3) allows changes in the pattern and hypertext when mapping the pattern in hypertext.

For variant 1, Amir *et. al*. proposed an algorithm that runs in $O(|V| + m \cdot |A|)$ time which was improved by Navarro to run in $O(m(|V|+|A|))$ time. Here, $|V|$ is the number of nodes in the graph, $|A|$ is the number of arcs, and $m$ is the length of the mapped pattern. For variants 2 and 3, Amir *et. al*. proved that the respective problems are NP-complete considering the Hamming and edit distance when the alphabet $\Sigma$ has $|\Sigma| \geq |V|$. In the work entitled *On the Complexity of Sequence-to-Graph Alignment* [3], Jain *et. al* prove that the variants 2 and 3 are NP-complete when the alphabet $\Sigma$ has $|\Sigma| \geq 2$.

The first time the `GSMP` was addressed using a De Bruijn graph as input was in the article entitled *Read Mapping on De Bruijn Graphs* [2]. In this work, Limasset *et. al*. propose the following problem, called here *De Bruijn Graph Sequence Mapping Problem* – `BSMP`: given a De Bruijn graph $G_k$ and a sequence $s$, the goal is to find a path $p$ in $G_k$ such that the sequence $s'$ spelled by $p$ have at most $d$ differences between $s$ and $s'$ with $d \in \mathbb{N}$. The `BSMP` was proved to be NP-complete considering the Hamming distance, leading to the development of a seed-and-extended heuristic by the mentioned authors. Note that for the `BSMP` it does not make sense to talk about the three variants mentioned above since there are no node repetitions.

Recently, the `BSMP` was addressed for walks in the article entitled *On the Hardness of Sequence Alignment on De Bruijn Graphs* [4]. Gibney *et. al*. proved that the problem is NP-

Supported by CAPES and UFMS.

complete when the changes occur in the graph and they proved that there is no algorithm faster than $O(|A| \cdot m)$ for De Bruijn graphs when we have changes only occur in $s$ in which $|A|$ is a number of arcs and $m$ is the length of $s$.

In a recently published article entitled *Heuristic for the De Bruijn Mapping Problem* [15], we dedicated ourselves to developing heuristics for Variant 1 (sequence changes) of BSMP. However, in this new article, we focus on exploring Variant 2 (graph changes) for BSMP. This work is motivated by the study by Gibney *et. al*, who demonstrated that the problem is NP-complete for Variant 2. Nevertheless, we present a reformulation of BSMP that enables the application of a polynomial algorithm to solve the problem.

## II. Preliminaries

In this section, we describe some necessary concepts such as computational definitions and problem definition that are used in this paper.

### A. Sequence, distance, graphs and matching

Let $\Sigma$ be an **alphabet** with a finite number of characters. We denote a sequence (or string) $s$ over $\Sigma$ by $s[1]s[2]\dots s[n]$ in which each character $s[i] \in \Sigma$. We say that the **length** of $s$, denoted by $|s|$, is $n$ and that $s$ is a $n$**-length** sequence.

We say that the sequence $s[i]s[i+1]\dots s[j]$ is a **substring** of $s$ and we denote it by $s[i,j]$. A substring of $s$ with length $k$ is a $k$-length sequence and also called $k$**-mer** of $s$. For $1 \leq j \leq n$ in which $n = |s|$, a substring $s[1,j]$ is called a **prefix** of $s$ and a substring $s[j,n]$ is called a **suffix** of $s$.

Given five sequences $s, t, x, w, z$, we define $st$ the **concatenation** of $s$ and $t$ and this concatenation contains all the characters of $s$ followed by the characters of $t$. If $s$ and $t$ are $n$- and $m$-length sequences respectively, $st$ is a $(n+m)$-length sequence. If $s = xw$ ($x$ is a prefix and $w$ is a suffix of $s$) and $t = wv$ ($w$ is a prefix and $z$ is a suffix of $t$), we say the substring $w$ is an **overlap** of $s$ and $t$.

The **Hamming distance** $d_h$ of two $n$-length sequences $s$ and $t$ is defined as

$$d_h(s,t) = \sum_{i=1}^{n} s[i] \stackrel{?}{=} t[i],$$

where $s[i] \stackrel{?}{=} t[i]$ is equal to 1 if $s[i] \neq t[i]$, and 0 otherwise. In this context, we also say that $s$ and $t$ have $d_h(s,t)$ differences.

A **graph** is an ordered pair $(V, A)$ of two sets in which $V$ is a set of elements called **nodes** (of the graph) and $A$ is a set of ordered pairs of nodes, called **arcs** (of the graph). Given a graph $G$, a **walk** in $G$ is a sequence of nodes $p = v_1, \dots, v_k$, such that for each pair $v_i$,$v_{i+1}$ of nodes in $p$ there is a arc $(v_i, v_{i+1}) \in A$. A **path** in $G$ is a walk with no repeated nodes. Given a walk $p = v_1, \dots, v_k$, $|p| = k-1$ is the **length** of $p$. For graphs with costs associated with their arcs, the **cost** of a walk $p$ is the sum of the cost of all arcs of all consecutive pairs of nodes $(v_i, v_{i+1})$ in $p$. A **shortest path** from $v_1$ to $v_k$ is one whose cost is minimum (a path of **minimum cost**).

An undirected graph $G = (V, A)$ is called a **bipartite graph** $H = (V_1 \cup V_2, A)$ if there is a partition of its vertices $V$ into two disjoint sets $V_1$ and $V_2$, such that for each edge $(u, v) \in A$, $u \in V_1$ and $v \in V_2$. A bipartite graph is called a **complete bipartite graph** if there is an edge from every vertex $u \in V_1$ to every vertex $v \in V_2$.

A **matching** $E$ in $H$ consists of a set of $m$ edges of $H$ that do not share common vertices. The **cost of the matching** $E$ is $C(E) = \sum_{e \in E} z(e)$, where $z(e)$ is the cost of an edge $e = \{x, y\} \in A$. A matching is said to be of **minimum cost** if there is no matching with a lower cost. A matching $E$ is **maximum** if there is no other matching larger than $E$, i.e., if there is no matching $E'$ such that $|E'| > |E|$. Given a set E with all maximum matchings, a matching $E \in$ E is said to be **maximum and of minimum cost** if there is no matching $E'$ such that $C(E') < C(E)$.

Given a bipartite graph $H = (V_1 \cup V_2, A)$, find a minimum-cost matching $E$. This problem is extensively studied in the literature and can be solved using, for example, Edmonds' algorithm (with a time complexity of $O((|V_1| \times |V_2|)^3)$), Hopcroft-Karp algorithm (with a time complexity of $O(\sqrt{(|V_1| \times |V_2|) \times |A|)})$, and the Hungarian algorithm (with a time complexity of $O((|V_1| \times |V_2|)^3)$) [11], [12], [17].

A **sequence graph** is a graph $(V, A)$ with a sequence of characters, built on an alphabet $\Sigma$, associated with each of its nodes. A **simple sequence graph** is a graph in which each node is labeled with only one character. Given a set $S = \{r_1, \dots, r_m\}$ of sequences and an integer $k \geq 2$, a **De Bruijn graph** is a graph $G_k = (V, A)$ such that:

- $V = \{d \in \Sigma^k |$ such that $d$ is a substring of length $k$ ($k$-mer) of $r \in S$ and $d$ labels only one node$\}$;
- $A = \{(d, d') |$ the suffix of length $k-1$ of $d$ is a prefix of $d'\}$.

Note that we can define the De Bruijn graph just by its set of vertices and the arcs are for each pair of $k$-mer $d$ and $d'$ whose $k-1$ length prefix of d is equal to the $k-1$ length suffix of $d'$.

In this paper, informally for readability, we consider the node label as node. Given a walk $p = v_1, v_2, \dots, v_n$ in a De Bruijn graph $G_k$, the **sequence spelled** by $p$ is the sequence $v_1 v_2[k] \dots v_n[k]$, given by the concatenation of the $k$-mer $v_1$ with the last character of each $k$-mer $v_2, \dots, v_n$. For a walk $p = v_1, v_2, \dots, v_n$ in a simple sequence graph $G$, the sequence spelled by $p$ is $v_1 v_2 \dots v_n$. A **mapping** of a sequence $s$ onto a simple sequence graph or a De Bruijn graph $G$ is a walk $p$ in $G$ whose editing cost between $s$ and the sequence spelled by $p$ is minimum.

Given the definitions above, we state the following problem for simple ssequence graphs (GSMP) and for De Bruijn graphs (BSMP) when we have changes only in the sequence, respectively:

*Problem 1 (**Graph Sequence Mapping Problem** – GSMP):* Given a $m$-sequence $s$ and a simple sequence graph $G$, find a mapping of $s$ onto $G$.

*Problem 2 (**De Bruijn Graph Sequence Mapping Problem** – BSMP):* Given a $m$-sequence $s$ and a De Bruijn graph of order $k$, $G_k$, find a mapping of $s$ onto $G_k$.

## III. Reformulating the De Bruijn Mapping Problem

Given a De Bruijn graph with an alphabet $|\Sigma| \geq 4$, a pattern $P$, and an integer $\delta \geq 0$, the BSMP, studied by Gibney et al. [4], determines whether there exists a walk $p$ in the De Bruijn graph. This walk must have a Hamming distance of zero concerning the sequence induced by $p$ and the pattern $P$, with a maximum of $\delta$ substitutions in the graph. In this context, the authors use a simple sequence graph to represent the De Bruijn graph. It is essential to note that in the visualization of the De Bruijn graph as a simple sequence graph, the $k$-mers are represented by walks of length $k$. In other words, each walk of length $k$ induces a distinct $k$-mer.

It is important to highlight that the NP-completeness of this problem for a simple sequence graph was already known, as explored earlier by Amir *et. al* and Jain *et. al*. The reduction studied by the authors starts from a directed graph $D$ that contains a Hamiltonian cycle. From $D$, they construct the De Bruijn graph as a simple sequence graph $G$ and demonstrate that this graph is indeed a De Bruijn graph. Finally, the authors' reduction proves that if there is a Hamiltonian cycle in $D$, then there exists a walk $p$ in $G$ with at most $\delta$ modifications in $G$, such that the sequence induced by $p$ and the pattern $P$ have a Hamming distance of zero. It is also demonstrated that if there is a walk $p$ in $G$ with at most $\delta$ modifications to $G$, then there exists a Hamiltonian cycle in $D$.

When exclusively considering modifications in the graph, they approach the problem in a way that does not compromise the fundamental structure of the graph, i.e., vertices and arcs are not modified. This characteristic is crucial when analyzing the NP-completeness of the problem, as, as mentioned earlier, the BSMP has been proven to be NP-complete in simple sequence graphs in the past, where modifications naturally preserved the structure of vertices and arcs.

On the other hand, when specifically addressing De Bruijn graphs, it is possible to explore the induction of new arcs after a modification in the graph. Given this characteristic, in this work, we explore modifications in the graph allowing for a change that can induce new arcs, implying a change in its underlying structure. This flexibility allows us, when the induction of new arcs is considered, to find solutions in polynomial time.

The main distinction lies in the approach to graph modifications. While for Gibney et al., the impossibility of altering the structure contributes to the NP-completeness of the problem, here, we acknowledge the possibility of changes that impact not only the labels of the vertices but also the topology of the graph. This permissiveness offers the advantage of admitting strategies to find solutions in polynomial time when the introduction of new arcs is allowed. The BSMP is approached differently in this work compared to the work of Gibney et al. In summary, we emphasize that the difference lies in the manipulation of the De Bruijn graph.

Given that the De Bruijn graph allows the induction of arcs through its $k$-mers, we consider the definitions in the next two sections to redefine the problem taking into account this characteristic of induced arcs.

### A. The s-transformation of a De Bruijn graph

Let $s$ be a sequence and $G_k$ a De Bruijn graph, where $V = \{v_1, \ldots, v_n\}$ is the set of vertices of $G_k$. We denote by $\boldsymbol{K(s)} = \{w_i, \ldots, w_m\}$ the set of $m$ unique subchains of length $k$ extracted from $s$, with $n \geq m$. An ***s*-transformation $\boldsymbol{S}$** of $G_k$ is the substitution of each vertex ($k$-mer) $v_i$ by $w_i$ (denoted as $v_i \rightarrow w_i$), resulting in a new graph $G_k^s$ where there exists a walk that induces $s$. A cost function $\boldsymbol{c(S)}$ is defined as the sum of Hamming distances between the corresponding pairs $v_i$ and $w_i$, i.e., $c(S) = \sum_i d_H(v_i, w_i)$. The set $\boldsymbol{\mathcal{S}}$ is defined as the set of all possible $s$-transformations of $G_k$, and $D_H(s, G_k) = \min_{S \in \mathcal{S}} c(S)$.

Based on the previously presented definitions, consider, for example, the sequence $s$ = ACTGCG and the De Bruijn graph $G_3$ represented only by its vertices $V = \{$ACT, CTG, ACG, TTT$\}$. In this case, we have $K(s) = \{$ACT, CTG, TGC, and GCG$\}$. Figure 1 shows a graphical representation with the induced arcs of $G_3$. In the same figure, the graph $G_3^s$ is represented, obtained after applying an $s$-transformation $S$ to $G_3$ considering the $k$-mers of $K(s)$, where $c(S) = 6$ is the minimum cost. This reflects the cost of replacing, in $G_3$, the $k$-mer ACG with TGC and the $k$-mer TTT with GCG. In $G_3^s$, the distance between $s$ and the sequence induced by $p = v_1, v_2, v_3, v_4$ is zero.

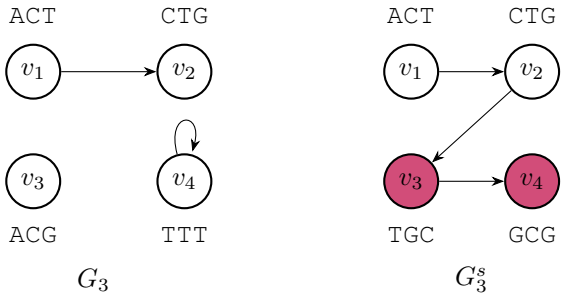


Figure 1. Example of an $s$-transformation $S$ of the De Bruijn graph $G_3$ to the De Bruijn graph $G_3^s$, considering the sequence $s$ = ACTGCG. Note that the path $p = v_1, v_2, v_3, v_4$ in $G_k^s$ induces the sequence $s$. The De Bruijn graphs are depicted with their induced arcs.

### B. Bipartition and pairing between two sets of k-mers

Given a sequence $s$ and a De Bruijn graph $G_k$, to determine an $s$-transformation of $G_k$, we construct a complete bipartite graph $H = (V \cup W, A)$ where $V$ corresponds to the set of vertices ($k$-mers from $G_k$), and $W$ is the set $K(s)$ associated with $s$. For each vertex ($k$-mer) $v \in V$ and $w \in W$, there exists an edge $e = \{v, w\} \in A$ with a cost $z(e) = d_H(v, w)$ (the Hamming distance between the $k$-mer of $v$ and the $k$-mer of $w$). To determine an $s$-transformation of $G_k$, we use a maximum cardinality matching of minimum cost $E$ in $H$. Each edge $\{v, w\} \in E$ represents the cost of replacing in $G_k$ the vertex $v$ (from $V$) with $w$ (from $W$).

An example of this complete bipartite graph is shown in Figure 2, corresponding to the De Bruijn graph $G_3$ with $V = \{\texttt{ACT}, \texttt{CTG}, \texttt{ACG}, \texttt{TTT}\}$ and $W = K(s) = \{\texttt{ACT}, \texttt{CTG}, \texttt{TGC}, \texttt{GCG}\}$ obtained from the sequence $s =$ACTGCG. In this example, the cost of each edge $e = \{v, w\}$, where $v \in V$ and $w \in W$, is determined by $z(e) = d_H(v, w)$. For the same bipartite graph presented in Figure 2, an example of a maximum cardinality matching of minimum cost is $E = \{\{v_1, w_1\}, \{v_2, w_2\}, \{v_3, w_4\}, \{v_4, w_3\}\}$ (dashed edges), with a total cost equal to 3.

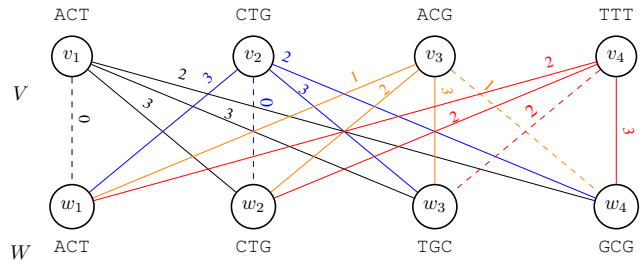


Figure 2. Example of a complete bipartite graph $H = (V \cup W, A)$ for the $k$-mers of the De Bruijn graph $G_3$, where $V = \{v_1, v_2, v_3, v_4\}$ represents the $k$-mers of $G_3$, and $W = K(s) = \{w_1, w_2, w_3, w_4\}$ is the set of $k$-mers obtained from the sequence $s =$ ACTGCG. Each edge $e = (v, w) \in A$ connects a vertex $v \in V$ to a vertex $w \in W$, with an associated cost $z(e) = d_H(v, w)$. Dashed edges represent a maximum matching of minimum cost.

### C. Reformulating the problem

Based on the previous information about the $s$-transformation, we redefine the problem DE BRUIJN GRAPH SEQUENCE MAPPING PROBLEM WITH GRAPH CHANGES as follows:

*Problem 3 (**De Bruijn Graph Sequence Mapping Problem** – BSMP):* Given as input a sequence $s$ and a De Bruijn graph $G_k$ represented only by its set of vertices, determine an $s$-transformation $S$ of $G_k$ of minimum cost.

## IV. ALGORITHM FOR THE PROBLEM OF MAPPING SEQUENCES INTO DE BRUIJN GRAPHS WITH CHANGES IN THE GRAPH

Given a De Bruijn graph $G_k$, represented exclusively by its set of vertices $V$, and a sequence $s$, the proposed algorithm for the version with changes in the labels of the De Bruijn graph vertices, called DE BRUIJN GRAPH MAPPING TOOL WITH GRAPH CHANGES – BMTC, follows the steps below:

1) Initially, a set $W$ is created containing all unique $k$-mers from the sequence $s$;
2) Next, a bipartite graph $H$ is created. In this context, connections are established from each $k$-mer $v \in V$ to all $k$-mers $w \in W$, with the edge cost $\{u, v\}$ given by $d_H(v, w)$;
3) The Hungarian algorithm is applied to the graph $H$ to find a maximum matching with minimum cost;
4) Based on the obtained matching, the $k$-mers in $V$ that have undergone changes are identified, resulting in the generation of a new set of vertices $V'$, representing the modified De Bruijn graph.

### A. Proof

To prove that the previous steps work, consider the following lemmas.

*Lemma 1:* $K(s) \subseteq V(G_k)$ if and only if there exists a walk in $G_k$ that spells out $s$.

*Lemma 2:* Let $S$ be an $s$-transformation of $G'_k$ resulting in $G^s_k$. There exists a matching $E$ in $G^s_k$ called a matching induced by $S$ such that $C(E) \leq c(S)$.

**Proof**. Let $S$ be an $s$-transformation of $G^s_k$ in $G_k$, meaning for the vertex set $V = \{v_1, \ldots, v_n\}$, $S$ is a substitution $v_i \to w_i$ for each $i$ transforming $G_k$ into $G^s_k$ where there exists a walk that spells out $s$. Given that there exists a walk that spells out $s$ by Lemma 1, we have for $G^s_k$ a $K(s) = \{u_i, \ldots, u_m\}$ such that $u_i = w_i, \ldots, u_m = w_m$ and $c(S) = \sum_i d_H(v_i, w_i)$. Let $E = \{u_i v_i : 1 \leq i \leq m\}$, then

$$\begin{aligned} C(E) &= \sum_{i=1}^{m} z(u_i, v_i) \\ &= \sum_{i=1}^{m} d_H(v_i, w_i) \leq c(S). \end{aligned}$$

□

Given the two previous lemmas, consider the following theorem:

*Theorem 1:* Let $s$ be a sequence, $G_k$ a Bruijn graph, and $E^*$ a minimum-cost matching in the bipartite graph $H = (V(G_k) \cup K(s), A)$. Then, $d_H(s, G_k) = C(E^*)$.

**Proof**. For each edge $\{x, y\} \in E^*$, $x \in V(G_k)$, $y \in K(s)$ replace $x$ with $y$ to obtain $G^s_k$ and, as a consequence of Lemma 1, an $s$-transformation $S$. Note that $c(S) = C(E^*)$. Therefore,

$$d_H(s, G_k) \leq c(S) = C(E^*).$$

Let $E$ be a matching induced by an $s$-transformation of minimum cost such that $C(E) \leq c(S)$, which Lemma 2 guarantees to exist. Therefore,

$$\begin{aligned} C(E^*) &\leq C(E) \\ &\leq c(S) = d_H(s, G_k). \end{aligned}$$

Hence, $d_H(s, G_k) = C(E^*)$. □

### B. Algorithm

The steps for the algorithm for the problem of sequence mapping onto De Bruijn graphs with graph changes are described in Algorithm 1.

### C. Complexity

For Algorithm 1 consider the follow values:

- $O(|s|)$ to identify all unique $k$-mers;
- $O((|V| + |W|) \cdot k)$ to create the bipartite graph with edge costs;
- $O((|V| + |W|)^3)$ to run the Hungarian algorithm;
- $O(|L|)$ to iterate through all edges in the matching and replace $k$-mers in $V'$.

**Algorithm 1** DE BRUIJN GRAPH MAPPING TOOL WITH GRAPH CHANGES – BMTC

**Require:** a sequence $s$ and the set $V$ of $k$-mers of a De Bruijn graph $G_k$.

**Ensure:** a set $V'$ of $k$-mers representing the modifications made to $V$.

$W \leftarrow$ unique $k$-mers (without repetition) from $s$;
**if** $|V| \geq |W|$ **then**
  $V' \leftarrow V$;
  $G \leftarrow$ BIPARTITE$(V, W)$; ▷ $G$ is a bipartite graph considering $V$ and $W$
  $L \leftarrow$ RUN HUNGARIAN ALGORITHM ON $H$; ▷ $L$ is a maximum matching of minimum cost
  **for** each arc $\{u, v\} \in L$ **do**
    assuming $v \in V'$;
    $V' \leftarrow V' \setminus \{v\}$; ▷ Removing $v$ from $V'$
    $V' \leftarrow V' \cup \{u\}$; ▷ Adding $u$ to $V'$
  **return** $(V, A)$
**return** $|V|$ must be greater than or equal to $|W|$;

**Algorithm 2** BIPARTITE

**Require:** two sets of vertices ($k$-mers) $V$ and $W$.

**Ensure:** a bipartite graph $H = (V \cup W, A)$ with edge costs.

$A \leftarrow \{\}$;
**for** $v \in V$ **do**
  **for** $w \in W$ **do**
    $c \leftarrow d_H(v, w)$; ▷ edge cost
    $A \leftarrow A \cup \{v, w\}$ where $\{v, w\}$ has cost $c$;
**return** $H = (V \cup W, A)$;

Therefore we have the following time complexity:

$$
\begin{aligned}
&= O(|s| - k) + O((|V| + |W|) \cdot k) + O((|V| + |W|)^3) + O(|L|) \\
&\leq O(|W|) + O((|V| + |W|) \cdot k) + O((|V| + |W|)^3) + O(2 \cdot |V|) \\
&= O(|W|) + O(|V| \cdot k + |W| \cdot k) + O((|V| + |W|)^3) + O(2 \cdot |V|) \\
&= O(|W|(k+1)) + O(|V|(k+2)) + O((|V| + |W|)^3)
\end{aligned}
$$

and the time complexity of Algorithm 1 is $O((|V|+|W|)^3)$.

## V. DISCUSSION, PERSPECTIVES AND CONCLUSIONS

In this paper, we delve into the De Bruijn Problem Mapping with changes in the graph. We explore the work of Gibney *et. al* [4], which demonstrated the NP-completeness of the problem for De Bruijn graphs. In this work, we allow a change in the graph not only to modify the labels of the vertices but also the topology of the graph. The fundamental distinction lies in the flexibility allowed, exploring the induction of new arcs after a modification in the De Bruijn graph. While Gibney *et. al* focus on modifications that preserve the fundamental structure, this work considers changes that can induce new arcs, enabling polynomial-time solutions.

Considering this feature of inducing new arcs, we redefine the problem, introducing concepts such as the *s-transformation* of a De Bruijn graph and the *Bipartition and matching between two sets of k-mers*. We develop an algorithm called BMTC, which utilizes the Hungarian algorithm to find a maximum-cost minimum matching in a bipartite graph, resulting in a modified set of vertices for the De Bruijn graph.

The proof that the algorithm works is based on lemmas establishing the relationship between the graph transformation and matchings in the bipartite graph. The theorem demonstrates that the cost of the maximum matching found in the bipartite graph is equal to the Hamming distance between the given sequence and the original graph. BMTC offers an innovative approach to the problem, allowing changes in the De Bruijn graph that may result in new arcs, proving advantageous for finding polynomial-time solutions.

It is noteworthy that the literature still lacks depth in this aspect. The most recent work by Gibney *et. al* [ [4]] addresses the problem but in a different manner. By introducing the possibility of inducing new arcs in case of changes in the De Bruijn graph, this work proposes a polynomial-time solution to the problem. However, further exploration of the potential practical applications of this solution is still needed.


## ACKNOWLEDGMENT


## REFERENCES


[1] A. Amir, et al.: Pattern matching in hypertext. Journal of Algorithms, (1997).
[2] A. Limasset, et al.: Read mapping on de bruijn graphs. In: BMC bioinformatics, vol. 17, no. 1, pp. 1–12, (2016).
[3] C. Jain, et al.: On the complexity of sequence-to-graph alignment. Journal of Computational Biology, 640–654 (2019).
[4] D. Gibney and S. Thankachan and S. Aluru.: On the hardness of sequence alignment on De Bruijn graphs. Journal of Computational Biology, 1377-1396 (2022)
[5] D. S. Hirschberg: A linear space algorithm for computing maximal common subsequences. In: Communications of the ACM, vol. 18, no. 6, pp. 341–343, (1975).
[6] E. Garrison, et al.: Variation graph toolkit improves read mapping by representing genetic variation in the reference. In: Nature biotechnology, pp. 875–879, (2018).
[7] E. W. Myers: The fragment assembly string graph. In: Bioinformatics, (2005).
[8] G. Navarro.: improved approximate pattern matching on hypertext. In: Theoretical Computer Science, pp. 455–463, (1998).
[9] G. Holley and P. Melsted.: Bifrost: highly parallel construction and indexing of colored and compacted de bruijn graphs. In: Genome biology, pp. 1–20, (2020).
[10] H. Li and N. Homer.: A survey of sequence alignment algorithms for next-generation sequencing. Brief Bioinform., 11, 473–483, (2010)
[11] J. Edmonds (1965). Paths, trees, and flowers. Canadian Journal of mathematics, 17:449–467
[12] J. E. Hopcroft and R. M. Karp (1973). An nˆ5/2 algorithm for maximum matchings in bipartite graphs. SIAM Journal on computing, 2(4):225–231
[13] N. G. De Bruijn: A combinatorial problem. In: Proc. Koninklijke Nederlandse Academie van Wetenschappen, vol. 49, pp. 758–764, (1946).
[14] P. A. Pevzner, et al.: An eulerian path approach to dna fragment assembly. In: Proceedings of the national academy of sciences, pp. 9748–9753, (2001).
[15] Rocha, L. B., Adi, S. S., e Araujo, E.: Heuristics for the De Bruijn Mapping Problem. In: Computational Science and Its Applications – ICCSA, (2023).
[16] U. Manber and S. Wu.: Approximate string matching with arbitrary costs for text and hypertext. In: Advances In Structural And Syntactic Pattern Recognition, pp. 22–33, World Scientific, (1992).
[17] H. W. Kuhn (1955). The hungarian method for the assignment problem. Naval research logistics quarterly, 2(1-2):83–97.