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 do 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 poses a significant challenge. One possible approach to tackling this problem is to find a walk p in G that spells a sequence most similar to s. This challenge is formally known as the Graph Sequence Mapping Problem (GSMP). In this paper, we delve into an alternative problem formulation known as the De Bruijn Graph Sequence Mapping Problem (BSMP). Both problems have three variants: changes only in the sequence, changes in the graph, and changes in both the sequence and the graph. We concentrate on addressing the variant involving changes in the graph. In the literature, when this problem does not allow the De Bruijn graph to induce new arcs after changes, it becomes NP-complete, as proven by Gibney *et. al* [4]. However, we reformulate the problem by considering the characteristics of the arcs induced in the De Bruijn graph. This reformulation alters the problem definition, thereby enabling the application of a polynomial-time algorithm for its resolution. Approaching the problem with this arc-inducing characteristic is pioneering, and the algorithm proposed in this work is groundbreaking in the literature.**

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

I. Introduction

Mapping a sequence onto another for comparative purposes is a important task in computational biology. Typically, a comparison is made between one sequence and a reference sequence, acknowledged as a high-quality representation of a sequence set [9], [10]. However, the reference sequence is inherently biased as it only reflects a limited set of sequences and cannot encompass all potential variations. To address this bias, one approach is to represent multiple sequences through a more robust structure, such as a sequence graph or De Bruijn graph [6], [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.

Focusing exclusively on variant 1, it has been thoroughly studied in recent years by various authors, and this variant admits a polynomial-time algorithm. For example, Amir *et. al* [1] proposed a polynomial algorithm that runs in time $O(m(|V| \cdot \log |V| + |A|))$, and Navarro [8] improved it to $O(m(|V| + |A|))$. In the case of variants 2 and 3, Amir *et. al* proved that the respective problems are NP-complete when considering Hamming and edit distance, especially when the alphabet Σ has $|\Sigma| \geq |V|$. In the work titled *On the Complexity of Sequence-to-Graph Alignment* [3], Jain *et. al* demonstrated that variants 2 and 3 are NP-complete even when the alphabet Σ 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

Supported by CAPES and UFMS.

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.

Considering that we extensively explored this issue in our recent paper titled *Heuristics for the De Bruijn Mapping Problem* [15], specifically focusing on variant 1 for the De Bruijn graph, our current focus is on investigating variant 2, which involves changes only in the graph structure. The BSMP was previously addressed in the context of walks in an article titled *On the Hardness of Sequence Alignment on De Bruijn Graphs* [4], where Gibney *et. al*. established its NP-completeness when changes occur in the graph.

Our current work is inspired by the study by Gibney *et. al* that demonstrated the NP-completeness of the problem for variant 2. However, we introduce a reformulation of BSMP based on a specific characteristic of the De Bruijn graph, which involves inducing new arcs when a change is made to the graph. This reformulation enables the application of a polynomial algorithm to solve the problem in this particular variant. In the study conducted by Gibney *et. al*, they do not account for this characteristic, resulting in the problem being NP-complete. In contrast, our study is pioneering in considering the induction of new arcs in the De Bruijn graph, making this polynomial algorithm groundbreaking for this 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 Σ be an **alphabet** with a finite number of characters. We denote a sequence (or string) s over Σ 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 Σ, 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 $v_1 v_2 \ldots v_n$. 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 sequence 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 δ 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 Gibney *et. al* 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 δ 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 δ 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.

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).

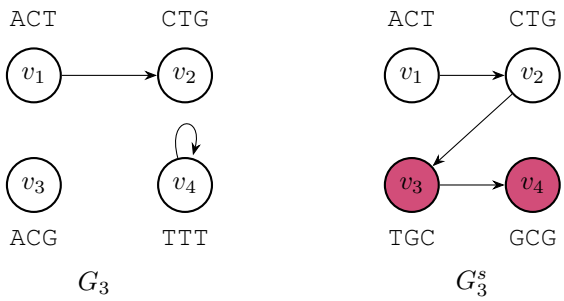

Figure 1. Example of an s-transformation S of the De Bruijn graph G_3 to the De Bruijn graph G_3^s, considering the sequence $s = \texttt{ACTGCG}$. Note that the path $p = v_1, v_2, v_3, v_4$ in G_k^s induces the sequence s. The De Bruijn graphs are depicted with their induced arcs.

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.

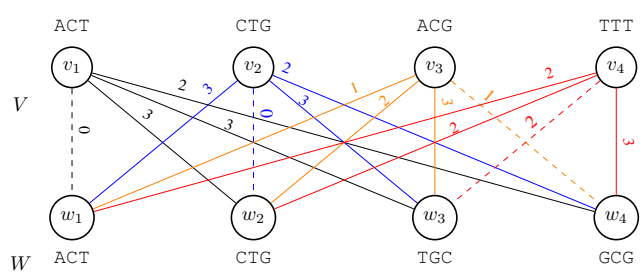

Figure 2. Example of a complete bipartite graph $H = (V \cup W, A)$ for the k-mers of the De Bruijn graph G_3, where $V = \{v_1, v_2, v_3, v_4\}$ represents the k-mers of G_3, and $W = K(s) = \{w_1, w_2, w_3, w_4\}$ is the set of k-mers obtained from the sequence $s = \texttt{ACTGCG}$. Each edge $e = (v, w) \in A$ 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.

The redefinition of this problem, considering the arcs induced in the De Bruijn graph, is pioneering and has no approach in the literature for this problem in this specific manner. Therefore, the algorithm proposed in this work is the first algorithm in the literature to address this problem.

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_k^s. There exists a matching E in G_k^s called a matching induced by S such that $C(E) \leq c(S)$.

Proof. Let S be an s-transformation of G_k^s 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_k^s 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_k^s 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_k^s 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.

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 ← unique k-mers (without repetition) from s;
if |V| ≥ |W| then
    V' ← V;
    G ← BIPARTITE(V, W);        ▷ G is a bipartite graph considering V and W
    L ← RUN HUNGARIAN ALGORITHM ON H;   ▷ L is a maximum matching of minimum cost
    for each arc {u, v} ∈ L do
        assuming v ∈ V';
        V' ← V' \ {v};          ▷ Removing v from V'
        V' ← V' ∪ {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 ← {};
for v ∈ V do
    for w ∈ W do
        c ← d_H(v, w);          ▷ edge cost
        A ← A ∪ {v, w} where {v, w} has cost c;
return H = (V ∪ W, A);
```

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'.

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 $O((|V| + |W|)^3)$. The implementation is available on github: github.com/LucasBarbosaRocha.

V. TESTS

In these tests, the objective is to use the developed tool named BMTC and evaluate the cost of modifications made in the graph, as well as the execution time of the tool. As discussed earlier, our proposal of BMTC is pioneering, as modifications in the graph induce new arcs, and therefore, there is no similar tool in the literature for comparison. In these tests, we used sequences with an average length of 5566 to map onto the graph, and De Bruijn graphs with k-mers ranging from 1464 to 11033. To assess BMTC, we performed the following operation in each test:

- Given all k-mers of the sequence s to be mapped onto the graph;
- Build a De Bruijn graph with these k-mers;
- To edit $\frac{\text{length of } s}{k}$ k-mers in the graph (note that this step may reduce the number of k-mers in the graph, as a modification can generate an already-existing k-mer);
- Run BMTC.

Given the previous steps, we have control over how many k-mers were altered in the graph, considering that a single modification was made to each k-mer in the graph, resulting in $\frac{\text{length of } s}{k}$ modifications in the graph. When executing BMTC, we obtain the pairing cost, and as expected, the cost is at most $\frac{\text{length of } s}{k}$. In these tests, we ran BMTC with k values equal to 10, 20, and 30. The results are presented in Table I. For each k value, we conducted 100 tests.

For $k = 10$, the number of k-mers in the graph varies between 1482 and 10933, with an average of 5519 k-mers. For $k = 20$, the number of k-mers in the graph varies between 1474 and 11043, with an average of 5547 k-mers. As for $k = 30$, the number of k-mers in the graph varies between 1464 and 11033, with an average of 5537 k-mers. It is observed in the table that, based on the average length of the sequence 5566, for $k = 10$, we have an average of 556 modified k-mers, with the value returned by BMTC being 372 (on average). For $k = 20$, we have 278 modified k-mers, with an average cost of 191 returned by BMTC for making modifications to the graph. As for $k = 30$, we have 185 modified k-mers, with an average cost of 126 for the modifications.

In the tests, we observed that the application executes in seconds, and it is also noted that increasing the value of k results in an increase in execution time. However, this application demonstrates promising performance, and the adopted strategy proves effective in executing modifications in the graph.

Table I
AVERAGE COSTS OBTAINED WITH THE BMTC TOOL AND THE AVERAGE EXECUTION TIME (IN SECONDS): TESTS CONDUCTED WITH MAPPED SEQUENCES OF AN AVERAGE LENGTH OF 5566.

k	Average number of k-mers in the graph	k-mers modified	Cost	Time
10	5519	556	372	6,83
20	5547	278	191	8,69
30	5537	185	126	13,43

VI. 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

This research was supported by CAPES and UFMS.

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.