ELSEVIER

Discrete Applied Mathematics 127 (2003) 85–94

DISCRETE APPLIED MATHEMATICS

Recognizing DNA graphs is difficult

Rudi Pendavingh[a], Petra Schuurman[a], Gerhard J. Woeginger[b,c,*1]

[a] *Department of Mathematics and Computing Science, Eindhoven University of Technology, P.O.Box 513, NL-5600 MB Eindhoven, Netherlands*

[b] *Department of Mathematics, University of Twente, P.O. Box 217, 7500 AE Enschede, Netherlands*

[c] *Institut für Mathematik, Technische Universität Graz, Steyrergasse 30, A-8010 Graz, Austria*

Received 17 April 2000; accepted 7 February 2001

Abstract

DNA graphs are the vertex induced subgraphs of De Bruijn graphs over a four letter alphabet. In this paper, we prove the NP-hardness of various recognition problems for subgraphs of De Bruijn graphs; in particular, the recognition of DNA graphs is shown to be NP-hard. As a consequence, two open questions from a recent paper by Błażewicz et al. (Discrete Appl. Math. 98, (1999) 1) are answered in the negative.

Keywords: Graph theory; Recognition algorithm; Computational complexity; NP-hardness; De Bruijn graph; DNA graphs; DNA computing

1. Introduction

Błażewicz et al. [3] introduced the concept of DNA graphs to model the computing and reconstruction phase of DNA chain sequencing by hybridization. For more information on the biological background, we refer the reader to Błażewicz et al. [3] or to Bains and Smith [1]. The graph theoretical background is fairly easy to describe: Let $G=(V,A)$ be a directed and simple graph that may contain loops. An (α,k)-*labeling* of G assigns a label $L(v)=[\ell_1(v),\ell_2(v),\ldots,\ell_k(v)]$ to every vertex $v\in V$ such that

(L1) All entries $\ell_i(v)$ in the labels of all vertices $v\in V$ are from an alphabet of size α (e.g., from the set $\{1,\ldots,\alpha\}$).

* Corresponding author. Department of Mathematics, University of Twente, P.O. Box 217, 7500 AE Enschede, Netherlands.

E-mail addresses: rudi@win.tue.nl (R. Pendavingh), petra@win.tue.nl (P. Schuurman), gwoegi@opt.math.tu-graz.ac.at (G.J. Woeginger).

[1] Supported by the START program Y43-MAT of the Austrian Ministry of Science.

0166-218X/03/$ - see front matter
PII: S0166-218X(02)00287-1

(L2) For $u, v \in V$ with $u \neq v$, $L(u) \neq L(v)$. Thus, different vertices get different labels.
(L3) For $u, v \in V$, $[\ell_2(u), \ldots, \ell_k(u)] = [\ell_1(v), \ldots, \ell_{k-1}(v)]$ holds if and only if there is an arc $(u, v) \in A$. In other words, an arc is encoded by the fact that the last $k-1$ entries of the label of the tail-vertex are equal to the first $k-1$ entries of the label of the head-vertex.

For integers $\alpha \geqslant 2$ and $k \geqslant 1$, we denote by $\mathscr{L}_k^\alpha$ the class of all directed simple graphs that possess an (α, k)-labeling. The set $\bigcup_{k=1}^{\infty} \mathscr{L}_k^\alpha$ is denoted by $\mathscr{L}_\infty^\alpha$, and the set $\bigcup_{\alpha=2}^{\infty} \mathscr{L}_k^\alpha$ is denoted by $\mathscr{L}_k^\infty$. Błażewicz et al. [3] call a graph G a *DNA graph* if and only if $G \in \mathscr{L}_\infty^4$. The four letters in the underlying alphabet correspond to the four nucleotide bases adenine (A), guanine (G), cytosine (C), and thymine (T).

Fifty years ago and working in a somewhat different line of research, De Bruijn [4] studied the subwords of certain circular sequences. To this end he investigated the combinatorial structure of directed graphs whose vertex set consists of the α^k possible words of length k over an alphabet of size α. Two vertices are connected by an arc if and only if the last $k-1$ letters of tail-word are equal to the first $k-1$ letters of the head word. Such a graph is nowadays called a *De Bruijn graph*, and it is denoted by $B(\alpha, k)$. De Bruijn graphs are used in communication networks and in VLSI design; cf. e.g. Samathan and Pradhan [8]. It is straightforward to see that a graph is a member of the above defined class $\mathscr{L}_k^\alpha$ if and only if it is the vertex induced subgraph of the De Bruijn graph $B(\alpha, k)$ with word length k and alphabet size α. Moreover, a graph is a DNA graph if and only if it is the vertex induced subgraph of $B(4, k)$ for some k.

In this paper, we will study the computational complexity of the membership problems for the classes $\mathscr{L}_k^\alpha$, $\mathscr{L}_\infty^\alpha$, and $\mathscr{L}_k^\infty$ for various values of $\alpha \geqslant 2$ and $k \geqslant 1$. Let us start our discussion with the classes $\mathscr{L}_k^\alpha$.

- For $k = 1$, the membership problem for $\mathscr{L}_1^\alpha$ is trivial: A directed simple graph $G = (V, A)$ is in $\mathscr{L}_1^\alpha$ if and only if $A = V \times V$ and $|V| \leqslant \alpha$.
- For $k = 2$, Błażewicz et al. [3] design a polynomial time algorithm that decides for any input graph G and any input parameter α, whether G is a member of $\mathscr{L}_2^\alpha$.
- For any fixed number $k \geqslant 3$, we will prove in this paper (cf. Theorem 3.3) that it is NP-hard to decide for an input graph G and an input parameter α whether $G \in \mathscr{L}_k^\alpha$.
- For $\alpha = 2$, the complexity of the membership problem for $\mathscr{L}_k^2$ is unknown. We conjecture that it is polynomially solvable.
- For any fixed number $\alpha \geqslant 3$, we will prove in this paper (cf. Theorem 4.3) that it is NP-hard to decide for an input graph G and an input parameter k whether $G \in \mathscr{L}_k^\alpha$.

Note that if α and k both are not part of the input, then the membership problem for $\mathscr{L}_k^\alpha$ is easy. In this case the size of class $\mathscr{L}_k^\alpha$ is a fixed constant that does not depend on the input, and we simply may search through all of it. Next, we turn to the membership problems for the classes $\mathscr{L}_\infty^\alpha$ and $\mathscr{L}_k^\infty$. Some of these problems are fairly close to the membership problems for the classes $\mathscr{L}_k^\alpha$.

- Błażewicz et al. [3] give a polynomial time algorithm that takes a graph G and a parameter k as input, and correctly decides whether G is in $\mathscr{L}_k^\infty$.
- For $\alpha = 2$, the complexity of the membership problem for $\mathscr{L}_\infty^2$ is unknown.

- For any fixed number $\alpha \geqslant 3$, we will prove in this paper (cf. Theorem 4.4) that it is NP-hard to decide for an input graph G whether $G \in \mathscr{L}^{\alpha}_{\infty}$.

At the end of [3], the authors pose five open questions on the computational complexity of recognizing graphs in various classes $\mathscr{L}^{\alpha}_{k}$. *Question* 1 considers a graph G with a given (α, k)-labeling, and it asks to find the largest possible label length q such that G is in $\mathscr{L}^{\beta}_{q}$ for some appropriate value of β. This question has been answered by Błażewicz et al. [2] who design a polynomial time algorithm for it. *Question* 2 concerns the complexity of the membership problem in $\mathscr{L}^{2}_{k}$. This question remains open. *Question* 3 asks for a polynomial time algorithm for the following problem: Given an integer k and a graph G, find the smallest integer α such that G is in $\mathscr{L}^{\alpha}_{k}$. Our Theorem 4.3 answers this question in the negative; the problem is NP-hard. *Question* 4 asks for a polynomial time algorithm for the membership problem for $\mathscr{L}^{\alpha}_{k}$. Theorems 3.3 and 4.3 answer this question in the negative, and they show that the problem is NP-hard. *Question* 5 asks for a polynomial time algorithm for the following problem: Given a graph G with a given (α, k)-labeling, determine all integers q such that G is in $\mathscr{L}^{\alpha}_{q}$. This question remains open.

Organization of the paper. In Section 2 we state some notation, we recall the definition of the graph coloring problem, and we derive some facts on De Bruijn graphs. Section 3 deals with the problem variants for fixed label lengths, and Section 4 deals with the problem variants for fixed sizes of the alphabet.

2. Preliminaries

Throughout the paper, labels will sometimes be considered as words over an appropriate alphabet, and then the entries of the labels will be considered as the letters of these words. For words w_1 and w_2, we define in the usual way their *concatenation* $w_1 \cdot w_2$ that results from appending word w_2 at the right end of word w_1.

All our NP-hardness proofs will be done by reductions from the graph coloring problem (cf. [5]). This coloring problem remains NP-hard even if the color bound $\gamma \geqslant 3$ is not part of the input.

2.1. Graph Coloring

Input: An undirected graph $H = (X, E)$ with $|X| = n$ vertices, and a color bound $3 \leqslant \gamma \leqslant n$.

Question: Does H have a feasible γ-coloring, i.e., does there exist a function $f : X \to \{1, \ldots, \gamma\}$ such that $f(x) \neq f(y)$ for all edges $(x, y) \in E$?

In the rest of this section, we investigate when a De Bruijn graph $B(\alpha, k)$ is a subgraph of another De Bruijn graph $B(\beta, h)$. This will lead to a gadget for our NP-hardness proof in Theorem 4.4.

Lemma 2.1. *Let $\alpha, \beta \geqslant 2$ and $k, h \geqslant 2$ be integers such that $2\alpha > \beta$. If the De Bruijn graph $B(\alpha, k)$ is a subgraph of $B(\beta, h)$, then $B(\alpha, k-1)$ is a subgraph of $B(\beta, h-1)$.*

Proof. To simplify presentation, we will sometimes identify vertices with their labels. Assume that $B(\alpha,k)$ is a subgraph of $B(\beta,h)$, and let $f:\{1,\ldots,\alpha\}^k \to \{1,\ldots,\beta\}^h$ denote the corresponding injection between the two label sets. Consider an arbitrary label v of length $k-1$ over $\{1,\ldots,\alpha\}$. There are α distinct vertices in $B(\alpha,k)$ whose labels end with the $k-1$ entries in v; we denote this vertex set by S_v. Moreover, there are α distinct vertices in $B(\alpha,k)$ whose labels start with the $k-1$ entries in v; we denote this vertex set by T_v. Note that the sets S_v and T_v are not necessarily disjoint. However, for $u \neq v$ we have $S_v \cap S_u = \emptyset$ and $T_v \cap T_u = \emptyset$.

In the graph $B(\alpha,k)$ there is an arc from every vertex in S_v to every vertex in T_v. Now consider the vertices in $f(S_v) = \{f(s) \mid s \in S_v\}$ and in $f(T_v) = \{f(t) \mid t \in T_v\}$. Since $B(\alpha,k)$ is a subgraph of $B(\beta,h)$, all labels of vertices in $f(S_v)$ must end with the same $h-1$ entries $[\ell_2,\ldots,\ell_h] \doteq w$, and all labels of vertices in $f(T_v)$ must start with the same $h-1$ entries w. Summarizing, these observations define for every label v in $\{1,\ldots,\alpha\}^{k-1}$ a unique corresponding label $w \doteq g(v)$ in $\{1,\ldots,\beta\}^{h-1}$. We claim

(i) that the function $g:\{1,\ldots,\alpha\}^{k-1} \to \{1,\ldots,\beta\}^{h-1}$ is an injection, and
(ii) that g certifies that $B(\alpha,k-1)$ is a subgraph of $B(\beta,h-1)$.

To see (i), suppose that for $u,v \in \{1,\ldots,\alpha\}^{k-1}$ with $u \neq v$ and for $w \in \{1,\ldots,\beta\}^{h-1}$, we have $g(u) = g(v) = w$. As f is an injection, $f(S_u \cup S_v)$ is a set of 2α labels, all with last $h-1$ entries equal to w, thus with 2α distinct first entries. This contradicts the assumption that $2\alpha > \beta$.

To see (ii), consider an arc from vertex u to vertex v in the graph $B(\alpha,k-1)$. Then the label of u starts with an entry x followed by a sequence y of $k-2$ entries, and the label of v starts with the sequence y followed by an entry z. Consider the vertex w in $B(\alpha,k)$ with label $x \cdot y \cdot z$, and let the label of $f(w)$ be $x' \cdot y' \cdot z'$ where x' and z' are single entries and where y' is a sequence of $h-2$ entries. With this notation we have $f(w) = f(x \cdot y \cdot z) = x' \cdot y' \cdot z'$, and hence $g(u) = g(x \cdot y) = x' \cdot y'$ and $g(v) = g(y \cdot z) = y' \cdot z'$. Consequently, there is also an arc in the graph $B(\beta,h-1)$ going from vertex $g(u)$ to vertex $g(v)$. □

Theorem 2.2. *Let $\alpha,\beta \geqslant 2$ and $k,h \geqslant 1$ be integers such that $2\alpha > \beta$. If the De Bruijn graph $B(\alpha,k)$ is a subgraph of the De Bruijn graph $B(\beta,h)$, then $h = k$.*

Proof. The proof is done by contradiction. Suppose that $k \neq h$. Then by Lemma 2.1 we may assume that $k = 1$ or $h = 1$. In case $k = 1$ and $h \geqslant 2$, observe that the De Bruijn graph $B(\alpha,1)$ contains two loops that are connected by an arc, whereas $B(\beta,h)$ does not contain such a configuration. Hence, $B(\alpha,1)$ cannot be a subgraph of $B(\beta,h)$. In case $h = 1$ and $k \geqslant 2$, observe that the De Bruijn graph $B(\beta,1)$ has only β vertices, whereas $B(\alpha,k)$ has $\alpha^k \geqslant q2\alpha > \beta$ vertices. Hence, $B(\alpha,k)$ cannot be a subgraph of $B(\beta,1)$. □

We note that since $B(2,2)$ is a subgraph of the complete directed graph on 4 vertices $B(4,1)$, the requirement that $2\alpha > \beta$ in the statement of Theorem 2.2 is necessary.

3. The problem variant with a fixed label length

Let $H=(X,E)$ and γ be an arbitrary instance of Graph Coloring. Let $x_1,\ldots,x_n$ be an enumeration of all vertices in X and let $e_1,\ldots,e_m$ be an enumeration of all edges in E. Without loss of generality we assume that H does not contain any isolated vertices. We will now construct in polynomial time a directed graph $G=(V,A)$ and an alphabet size α, such that G is in $\mathscr{L}_3^\alpha$ if and only if H is γ-colorable.

- For every vertex $x_i \in X$ with $1 \leqslant i \leqslant n$, the graph G contains two corresponding vertices $v(x_i)$ and $v'(x_i)$. There is a loop at every vertex $v(x_i)$. From every vertex $v(x_i)$ with $1 \leqslant i \leqslant n$ there is an outgoing arc to $v'(x_i)$.
- For all $1 \leqslant i \neq j \leqslant n$, there is a directed path with three arcs from $v(x_i)$ to $v(x_j)$ through the two intermediate vertices $v_1(x_i,x_j)$ and $v_2(x_i,x_j)$. From every vertex $v_2(x_i,x_j)$ there is an outgoing arc to $v'(x_j)$, and an outgoing arc to every vertex $v_1(x_j,x_k)$ with $1 \leqslant k \leqslant n$ and $k \neq j$.
- For every edge $e_s \in E$ with $1 \leqslant s \leqslant m$, the graph G contains a corresponding vertex $v(e_s)$. There is a loop at every $v(e_s)$.
- For all $1 \leqslant s \neq t \leqslant m$, there is a directed path with three arcs from $v(e_s)$ to $v(e_t)$ through the two intermediate vertices $v_1(e_s,e_t)$ and $v_2(e_s,e_t)$. From every vertex $v_2(e_s,e_t)$ there is an outgoing arc to every vertex $v_1(e_t,e_u)$ with $1 \leqslant u \leqslant m$ and $u \neq t$.
- If vertex x_i is incident to edge e_s in H, then G contains a directed path from $v'(x_i)$ to $v(e_s)$ via vertices $v_1(x_i,e_s)$ and $v_2(x_i,e_s)$. From every vertex $v_2(x_i,e_s)$ there is an outgoing arc to every $v_1(e_s,e_t)$ with $1 \leqslant t \leqslant m$ and $t \neq s$.

This completes the construction of graph $G=(V,A)$. Finally, we define the alphabet size $\alpha = n+m+\gamma$. To simplify the presentation of the following arguments, we assume that the α letters in the alphabet are the vertices $x_1,\ldots,x_n$ and the edges $e_1,\ldots,e_m$ in H, together with γ colors $c_1,\ldots,c_\gamma$.

Lemma 3.1. *If $G=(V,A)$ has an $(\alpha,3)$-labeling, then the graph H is γ-colorable.*

Proof. Consider an $(\alpha,3)$-labeling of G. It is easy to see that the label of a vertex with a loop must consist of three identical letters. Without loss of generality we assume that every vertex $v(x_i)$ is labeled by $[x_i,x_i,x_i]$ and that every vertex $v(e_s)$ is labeled by $[e_s,e_s,e_s]$. From this we derive that vertex $v_1(x_i,x_j)$ has label $[x_i,x_i,x_j]$, and that vertex $v_2(e_t,e_s)$ has label $[e_t,e_s,e_s]$.

Now consider vertex $v'(x_i)$. Since it is a successor of $v(x_i)$, its label must be of the form $[x_i,x_i,\sigma]$ where σ is some letter from the alphabet. The letter σ cannot be equal to x_i, since then $v'(x_i)$ and $v(x_i)$ would have the same label in contradiction to property (L2). The letter σ can also not be equal to x_j with $i \neq j$, since then vertices $v'(x_i)$ and $v_1(x_i,x_j)$ would have the same label. Since x_i is not an isolated vertex in H, there is a path from $v'(x_i)$ to some vertex $v(e_s)$ through the intermediate vertices $v_1(x_i,e_s)$ and $v_2(x_i,e_s)$. Then the label of vertex $v_1(x_i,e_s)$ equals $[x_i,\sigma,e_s]$, and the label of vertex $v_2(x_i,e_s)$ equals $[\sigma,e_s,e_s]$. We conclude that σ can neither be equal to e_s (since then $v_2(x_i,e_s)$ and $v(e_s)$ had the same label) nor can it be equal to some e_t with $t \neq s$ (since

then $v_2(x_i, e_s)$ and $v_2(e_t, e_s)$ had the same label). The only remaining possibility for σ is that it is one of the colors $c_1, \ldots, c_\gamma$.

We now define $f(x_i) = \sigma_i$ for $1 \leqslant i \leqslant n$, where $[x_i, x_i, \sigma_i]$ is the label of vertex $v'(x_i)$. By the above discussion, every σ_i is one of the colors $c_1, \ldots, c_\gamma$. We claim that this yields a feasible coloring. Indeed, suppose that $f(x_i) = f(x_j) = \sigma$ where x_i and x_j are connected to each other by an edge e_s in H. Then the label of $v_2(x_i, e_s)$ equals $[\sigma, e_s, e_s]$, and the label $v_2(x_j, e_s)$ also equals $[\sigma, e_s, e_s]$, a contradiction to property (L2). □

Lemma 3.2. *If the graph H is γ-colorable, then $G = (V, A)$ has an $(\alpha, 3)$-labeling.*

Proof. Consider a feasible γ-coloring of H that assigns to every vertex $x_i \in X$ a color σ_i from the colors $c_1, \ldots, c_\gamma$. We define the following labeling:

- For $1 \leqslant i \leqslant n$, the label of $v(x_i)$ is $[x_i, x_i, x_i]$ and the label of $v'(x_i)$ is $[x_i, x_i, \sigma_i]$.
- For all $1 \leqslant i \neq j \leqslant n$, the label of $v_1(x_i, x_j)$ is $[x_i, x_i, x_j]$ and the label of $v_2(x_i, x_j)$ is $[x_i, x_j, x_j]$.
- For $1 \leqslant s \leqslant m$, the label of $v(e_s)$ is $[e_s, e_s, e_s]$.
- For all $1 \leqslant s \neq t \leqslant m$, the label of $v_1(e_s, e_t)$ is $[e_s, e_s, e_t]$ and the label of $v_2(e_s, e_t)$ is $[e_s, e_t, e_t]$.
- If vertex x_i is incident to edge e_s, then the label of $v_1(x_i, e_s)$ is $[x_i, \sigma_i, e_s]$, and the label of $v_2(x_i, e_s)$ is $[\sigma_i, e_s, e_s]$.

Clearly, this labeling assigns distinct labels to distinct vertices. To verify that the labeling also fulfills property (L3), we divide the vertices of G into five classes: The first class V_1 contains all vertices $v(x_i)$, $v_1(x_i, x_j)$, and $v_2(x_i, x_j)$; the entries in the labels of all these vertices are from $x_1, \ldots, x_n$. The second class V_2 contains all vertices $v'(x_i)$; the first two entries in the labels of these vertices are from $x_1, \ldots, x_n$, and the last entry is from $c_1, \ldots, c_\gamma$. The third class V_3 contains all vertices $v_1(x_i, e_s)$; their labels consist of some vertex x_i, followed by a color, followed by an edge. The fourth class V_4 contains all vertices $v_2(x_i, e_s)$; their labels consist of some color, followed by two edges. The fifth class V_5 contains all vertices $v(e_s)$, $v_1(e_s, e_t)$, and $v_2(e_s, e_t)$; their labels are triples of edges.

By the definition of the labeling, there can only be arcs within V_1, arcs from V_1 to V_2, arcs from V_2 to V_3, arcs from V_3 to V_4, arcs from V_4 to V_5, and arcs within V_5. It is now straightforward but somewhat tedious to verify that all arcs within V_1 and within V_5, from V_1 to V_2, and from V_4 to V_5 are correctly encoded. Moreover, every vertex x_i has a unique color σ_i, and therefore every label $[x_i, \sigma_i, e_s]$ of a vertex in V_3 only allows a unique predecessor and a unique successor. □

Combining the statements of Lemmas 3.1 and 3.2 now yields the statement of Theorem 3.3 below for $k = 3$. The construction is easily extended to the cases with a fixed $k \geqslant 4$. The main idea is to replace in our construction all the connecting paths with three arcs by new connecting paths with k arcs. The details are left to the reader.

Theorem 3.3. *For any fixed $k \geqslant 3$ the following problem is* NP*-hard: Decide for an input graph G and an input parameter α whether $G \in \mathscr{L}_k^\alpha$.*

4. The problem variant with a fixed alphabet size

Let $H=(X,E)$ and γ be an arbitrary instance of Graph Coloring. By adding at most $|E|-1$ independent edges to H, we can make $|E|$ a perfect power of two without increasing the chromatic number of H. Then by adding some isolated vertices (again without increasing the chromatic number of H), we can make $|X|=|E|$. Summarizing this yields that we may assume without loss of generality that $n=|X|=|E|=2^z$ holds for some integer $z \geqslant 4$.

Let $x_1,\ldots,x_n$ be an enumeration of all vertices in X and let $e_1,\ldots,e_n$ be an enumeration of all edges in E. We will now construct in polynomial time a directed graph $G=(V,A)$ and a label length k, such that G is in $\mathscr{L}_k^\gamma$ if and only if H is γ-colorable.

- For every vertex $x_i \in X$ with $1 \leqslant i \leqslant n$, the graph G contains a corresponding *complete binary out-tree* $T(x_i)$ of height z. The root of this tree is the vertex $v(x_i)$. Every interior vertex in $T(x_i)$ has a left and a right out-going arc that connect it to its two children. All 2^z leaves in $T(x_i)$ are at the same distance z from the root. Enumerating the leaves in $T(x_i)$ from left to right, they are called $v(x_i,e_1),v(x_i,e_2),\ldots,v(x_i,e_n)$.
- For every edge $e_s \in E$ with $1 \leqslant s \leqslant n$, the graph G contains a corresponding vertex $v(e_s)$.
- If vertex x_i is incident to edge e_s in H, then G contains a directed path $P(x_i,e_s)$ from $v(x_i,e_s)$ to $v(e_s)$. The path $P(x_i,e_s)$ consists of $z+2$ arcs and of $z+1$ new vertices. The interior vertices on this path are not connected to any other parts of graph G.

This completes the construction of graph $G=(V,A)$. We define the label length $k=2z+3$.

To simplify the presentation of the following arguments, we assume that the alphabet Σ consists of the letters $0,1,\ldots,\gamma-1$. Some of the labels will be concatenated from the *binary* representations of certain integers. For an integer i in the range $1 \leqslant i \leqslant 2^z$, we define $BIN(i)$ to be the z-digit binary representation of $i-1$. This representation always contains an appropriate number of leading zeroes so that its length exactly equals z. For non-negative integers i and for letters $\sigma \in \Sigma$, we will use σ^i to represent the word consisting of exactly i letters σ.

Lemma 4.1. *If $G=(V,A)$ has a (γ,k)-labeling, then the graph H is γ-colorable.*

Proof. Consider a (γ,k)-labeling of G. For any vertex $x_i \in X$, we define its color $f(x_i)$ to be the $(2z+2)$th digit of the label of the root $v(x_i)$. Since the alphabet size is γ, this coloring uses only γ colors. We claim that this yields a feasible coloring.

Indeed, suppose that $f(x_i)=f(x_j)=\sigma$ where x_i and x_j are connected to each other by an edge e_s in H. We consider the predecessors of $v(e_s)$ on the paths $P(x_i,e_s)$ and $P(x_j,e_s)$, and call these vertices $w(x_i,e_s)$ and $w(x_j,e_s)$, respectively. There is a directed path of $2z+1$ arcs from $v(x_i)$ to $w(x_i,e_s)$. As a consequence of property (L3), the $(2z+2)$th digit of the label of $v(x_i)$ must be equal to the first digit of the label of $w(x_i,e_s)$. Hence, this first digit equals σ. By analogous reasoning we get that the first digit of the label of $w(x_j,e_s)$ also equals σ. Since $w(x_i,e_s)$ and $w(x_j,e_s)$ both are predecessors of $v(e_s)$, the last $k-1$ digits of their labels must agree with the first $k-1$ digits of the label of $v(e_s)$. But now the labels of $w(x_i,e_s)$ and $w(x_j,e_s)$ agree in the

first digit and also in the last $k-1$ digits, and hence they are equal to each other. This contradicts property (L2). □

Lemma 4.2. *If the graph H is γ-colorable, then $G=(V,A)$ has a (γ,k)-labeling.*

Proof. Consider a feasible γ-coloring of H that assigns to every vertex $x_i \in X$ a color σ_i from Σ. We define the following partial labeling:

- For $1 \leqslant i \leqslant n$, the label of $v(x_i)$ is $2^{z-1} \cdot 0 \cdot BIN(i) \cdot 2 \cdot \sigma_i \cdot 1$.
- For $1 \leqslant s \leqslant n$, the label of $v(e_s)$ is $1 \times BIN(s) \cdot 2^{z+2}$.

All other vertices are on some path with $2z+2$ arcs from some root $v(x_i)$ to some $v(e_s)$. If a vertex v is j arcs away from $v(x_i)$ and $2z+2-j$ arcs away from $v(e_s)$, then its first $2z+3-j$ letters are the last $2z+3-j$ letters of the label of $v(x_i)$, and its last $j+1$ letters are the first $j+1$ letters of the label of $v(e_s)$. This completely determines the label of v. Summarizing, the label of every vertex on path $P(x_i,e_s)$ is an appropriate subword of length $2z+3$ of the word

$$2^{z-1} \cdot 0 \cdot BIN(i) \cdot 2 \cdot \sigma_i \cdot 1 \cdot BIN(s) \cdot 2^{z+2}.$$

For an illustration, the reader may want to verify that the label of the leaf vertex $v(x_i,e_s)$ in $T(x_i)$ equals $BIN(i) \cdot 2 \cdot \sigma_i \cdot 1 \cdot BIN(s)$. Now we state some simple but important observations on these labels.

First: By considering a label $L(v)$, one can easily determine whether the corresponding vertex v is in one of the trees $T(x_i)$, or lies on one of the paths $P(x_i,e_s)$, or is one of the vertices $v(e_s)$. Indeed, if a label $L(v)$ has a 0 or a 1 as last digit, then the vertex v is contained in some tree $T(x_i)$. And if the label $L(v)$ ends with a 2, then the vertex v lies on one of the paths $P(x_i,e_s)$, or it is one of the vertices $v(e_s)$; and v is one of the vertices $v(e_s)$ if and only if its label $L(v)$ ends with a block of 2's of length $z+2$.

Second: If a vertex v is contained in some tree $T(x_i)$, then its label $L(v)$ starts with a block of 2s that is followed by a 0 that in turn is followed by z binary digits that uniquely identify the index i. For a non-leaf vertex v, the only possible labels of successors of v result by removing the first digit from $L(v)$ and by appending a 0 or a 1 at its right end; this correctly encodes the tree structure of $T(x_i)$. For the leaves $v(x_i,e_s)$, there is at most one possible successor label that results by removing the first digit (0 or 1) and by appending a 2. Such a successor vertex exists if and only if x_i is incident to e_s. Hence, also the initial arcs of the paths $P(x_i,e_s)$ are correctly encoded by the labeling.

Finally: Let us consider a vertex v that lies on one of the paths $P(x_i,e_s)$. Its label $L(v)$ ends with a non-empty block of 2s that is preceded by z binary digits that uniquely identify the index s. These z binary digits are preceded by a digit 1, which in turn is preceded by the digit σ_i. Since the edge e_s is incident to two vertices x_i and x_j with different colors in the γ-coloring, we have $\sigma_i \neq \sigma_j$. Hence, the index s together with the letter σ_i uniquely identifies the path on which the vertex v lies. The only possible label for a successor results by removing the first digit, and by appending a 2. Hence, all the arcs on all the paths $P(x_i,e_s)$ are correctly encoded. □

Combining the statements of Lemmas 4.1 and 4.2 now yields that there exists a γ-coloring for H if and only if there exists a (γ, k)-labeling for G. Since γ-coloring remains NP-hard for every fixed $\gamma \geqslant 3$, we get the following theorem.

Theorem 4.3. *For any fixed $\alpha \geqslant 3$ the following problem is* NP*-hard*: *Decide for an input graph G and an input parameter k whether $G \in \mathscr{L}^{\alpha}_{k}$.*

In fact our reduction also proves that it is hard to find labelings whose alphabet size is close to optimal, since all the inapproximability results from graph coloring (cf. eg. [6,7]) immediately carry over to the labeling problem. E.g. since it is NP-hard to find a 4-coloring for a 3-colorable graph [7], it is NP-hard to find a $(4, k)$-labeling for a graph G in $\mathscr{L}^{3}_{k}$. Since (unless P = NP) there is no constant factor approximation algorithm for graph coloring [6], there also is no constant factor approximation algorithm for minimizing the alphabet size α under a given label length k.

Theorem 4.4. *For any fixed $\alpha \geqslant 3$ the following problem is* NP*-hard*: *Decide for an input graph G whether $G \in \mathscr{L}^{\alpha}_{\infty}$.*

Proof. Let $H = (X, E)$ be an arbitrary instance of Graph γ-Coloring; throughout the proof we assume that $\gamma \geqslant 3$ is a fixed constant. We repeat the polynomial time construction of above: We make $n = |X| = |E| = 2^z$ for some integer z, and we construct a directed graph $G = (V, A)$, such that G is in $\mathscr{L}^{\gamma}_{2z+3}$ if and only if the undirected graph H is γ-colorable. Finally, we define the graph $G' = (V', A')$ to be the disjoint union of this graph G and of the De Bruijn graph $B(\gamma - 1, 2z + 3)$. The De Bruijn graph $B(\gamma - 1, 2z + 3)$ has

$$(\gamma - 1)^{2z+3} \leqslant \gamma^{2 \log n + 3} = \gamma^3 n^{2 \log \gamma}$$

vertices. Hence, the size of G' is polynomial in the size of $H = (X, E)$, and the whole construction can be done in polynomial time.

We claim that G' is in $\mathscr{L}^{\gamma}_{\infty}$ if and only if G is in $\mathscr{L}^{\gamma}_{2z+3}$. First, assume that G' is in $\mathscr{L}^{\gamma}_{\infty}$. Then G' is a vertex induced subgraph of $B(\gamma, k)$ for some k. Since on the other hand $B(\gamma - 1, 2z + 3)$ is a subgraph of G', Theorem 2.2 now yields that $k = 2z + 3$ must hold. Therefore, the vertex induced subgraph G of G' indeed is in $\mathscr{L}^{\gamma}_{2z+3}$. Next, assume that G is in $\mathscr{L}^{\gamma}_{2z+3}$. Then we can reuse the $(\gamma, 2z + 3)$-labeling that we defined in Lemma 4.2 for G. Moreover, the De Bruijn graph $B(\gamma - 1, 2z + 3)$ can be $(\gamma, 2z + 3)$-labeled in the natural way by all words of length $2z + 3$ over the alphabet $\{0, 1, \ldots, \gamma - 1\} \setminus \{2\}$. In the labels for vertices in G, every substring of length $z + 3$ contains at least once the digit 2. In the labels for vertices in $B(\gamma - 1, 2z + 3)$, the digit 2 does not show up at all. Consequently, these labelings do not generate any cross-edges between the graphs G and $B(\gamma - 1, 2z + 3)$, and they together yield a $(\gamma, 2z + 3)$-labeling for G'.

If we combine these statements with the statements in Lemmas 4.1 and 4.2, we get that G' is in $\mathscr{L}^{\gamma}_{\infty}$ if and only if the graph H is γ-colorable. □

Corollary 4.5. *It is* NP*-hard to decide whether an input graph G is a DNA graph.*

References

[1] W. Bains, G.C. Smith, A novel method for nucleic acid sequence determination, J. Theoret. Biol. 135 (1988) 303–307.
[2] J. Błażewicz, P. Formanowicz, M. Kasprzak, D. Kobler, On the recognition of De Bruijn graphs and their induced subgraphs, manuscript, September, 1999.
[3] J. Błażewicz, A. Hertz, D. Kobler, D. de Werra, On some properties of DNA graphs, Discrete Appl. Math. 98 (1999) 1–19.
[4] N.G. de Bruijn, A combinatorial problem. Koninklijke Nederlandse Akademie van Wetenschappen te Amsterdam, Proc. 49 (1946) 758–764.
[5] M.R. Garey, D.S. Johnson, Computers and Intractability: A Guide to the Theory of NP-Completeness, Freeman, San Francisco, 1979.
[6] J. Håstad, Some optimal inapproximability results, Proceedings of the 29th Annual ACM Symposium on the Theory of Computing (STOC'97), 1997, pp. 1–10.
[7] S. Khanna, N. Linial, S. Safra, On the hardness of approximating the chromatic number, in: Proceedings of the Second Israeli Symposium on Theory and Computing Systems (ISTCS'93), 1993, pp. 250–260.
[8] M.R. Samathan, D.K. Pradhan, The De Bruijn multiprocessor network: a versatile parallel processing and sorting network for VLSI, IEEE Trans. Comput. 38 (1989) 567–581.