

國立交通大學

應用數學系

碩士論文

DNA 圖的研究



研究生：游舜婷

指導教授：傅恆霖 教授

中華民國九十八年六月

DNA 圖的研究

A Study of DNA Graphs

研究生：游舜婷

Student : Shun-Ting Yu

指導教授：傅恆霖

Advisor : Hung-Lin Fu

國立交通大學

應用數學系

碩士論文



Submitted to Department of Applied Mathematics

College of Science

National Chiao Tung University

in Partial Fulfillment of the Requirements

for the Degree of

Master

in

Applied Mathematics

June 2009

Hsinchu, Taiwan, Republic of China

中華民國九十八年六月

DNA圖的研究


研究生：游舜婷

指導老師：傅恆霖 教授

國立交通大學

應用數學系

摘要



分子生物學主要是研究 DNA 序列及蛋白質的結構。由於序列的特性及讀取序列受到長度的限制，利用建構有向圖的數學模型可以有效地確定 DNA 序列及研究蛋白質的結構。此有向圖是這樣建構的：將每個長度為 k 的核苷酸當成點，對於兩點 x, y ，如果 x 這點後 i 段的核苷酸與 y 這點前 i 段的 DNA 序列要一樣，則 x, y 有一條有向邊 (x, y) 。我們將這類的圖稱作 DNA 圖或 DNA 標記圖。兩者的差別在於核苷酸是否有重複使用。在這篇論文中，我們主要是針對點數較小的圖去刻劃 DNA 圖的特性。

A Study of DNA Graphs

Student: Shun-Ting Yu

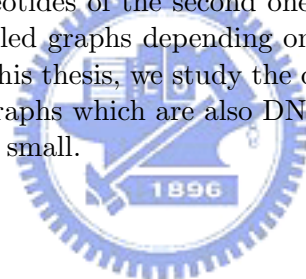
*Department of Applied Mathematics
National Chiao Tung University
Hsinchu, Taiwan 30050*

Advisor: Hung-Lin Fu

*Department of Applied Mathematics
National Chiao Tung University
Hsinchu, Taiwan 30050*

Abstract

Molecular biology aims to study DNA and protein structure, that is the recognition of DNA primary structure. In order to do that, a mathematical model based on graph theory has been developed in recent years. Mainly, suitably defined digraphs are presented. A digraph built from the spectrum (a set of some k -long oligonucleotides) as follows: each oligonucleotide from the spectrum becomes a vertex, two vertices are connected by an arc if the i rightmost nucleotides of the first point overlap with the i leftmost nucleotides of the second one. We refer to these graphs as DNA graphs and DNA labelled graphs depending on whether the oligonucleotides used are distinct or not. In this thesis, we study the digraphs mentioned above and characterize DNA labelled graphs which are also DNA graphs, especially when the order (number of vertices) is small.



Acknowledgement

首先，要感謝我的指導教授傅恆霖老師。這兩年來，修過老師的圖論、組合設計等，使我學到不少東西。在撰寫論文的過程中，老師的指點及意見更是使我獲益良多。在這裡我也要感謝系上的老師們，陳秋媛老師、黃大原老師及翁志文老師，感謝你們這兩年來不厭其煩的教導與關心。

特別感謝惠蘭學姊、敏筠學姊、智懷學長、宏賓學長、元勳學長、榮丰學長、彥婷學姊，你們總是能在我遇到問題的時候，耐心的指導我、幫忙我。感謝研究所的同學們，永潔、玉雯、慧棻、佳玲、宜君、彥琳、文昱、逸軒、哲皓、葉彬、士慶、巧玲、光祥、瑞毅、松育、昱承等，因為有你們，使我這兩年的生活過的更加多彩多姿。

感謝戰友們裴裴、敏筠、小八、筠庭、小吵、顥顥、壞壞、小泡，大家的熱血及努力，讓我們拿下北數盃及系際盃女籃冠軍。

最後我要感謝我的家人，因為有你們的支持與鼓勵，我才能無後顧之憂的專心於我的學業上。在此謝謝爸爸、媽媽及妹妹，感謝你們讓我順利完成碩士學位，謝謝。

Contents

Abstract (in Chinese)	i
Abstract (in English)	ii
Acknowledgement	iii
Contents	iv
List of Figures	v
1 Introduction and Preliminaries	1
2 DNA Graphs	4
2.1 Characterization of directed line-graph	4
2.2 Some properties of the classes S_k^∞	6
2.3 Some properties of the classes S_k^α	10
3 DNA labelled Graphs	12
3.1 The relationship between DNA labelled graphs and DNA graphs	12
3.2 Some properties of DNA labelled graphs	14
3.3 The relationship between DNA labelled graphs and adjoints	18
3.4 An equivalence relation of DNA labelled graphs	18
4 Main Results	21
5 Concluding Remarks	28

List of Figures

1	The graphs G_1, G_2 and G_3	4
2	The graphs D_1, D_2 and D_3	5
3	$H \in S_3^\infty$ but $H \notin S_k^\infty$ for $k \geq 4$	8
4	D is a DNA labelled graph but not a DNA graph.	13
5	The graph D_i	22
6	$D \in S_{2,1}^4$ but $D \notin S_2^4$	22
7	D is weakly connected and satisfying the sufficient condition of Theorem 4.2, but D is not a DNA graph.	26
8	D is weakly connected and satisfying the sufficient condition of Theorem 4.2, but D is not a DNA graph.	27



1 Introduction and Preliminaries

It is well known that DNA (deoxyribonucleic acid) is a double helix in which the two coiled strands (chains) are composed of only four different molecule types—nucleotides. Every nucleotide consists of phosphate, sugar and one of the following bases: adenine (abbreviated A), thymine (T), guanine (G) and cytosine (C). The two chains are held together by hydrogen bonds which exist only between the pairs of complementary bases, which are A-T and G-C. It follows that knowing one chain, the other (complementary) can be easily reconstructed.

A DNA sequence in molecular biology may be viewed as a sequence of characters from the DNA alphabet $\{A, T, G, C\}$. One of the methods of recognition of the primary structure of DNA sequences is hybridization. This method consists of two phases: biochemical and computational. In the biochemical phase a set of (possibly all) subchains constituting the DNA chain which is to be read, is found. Then, in the computational phase these subchains are to be put in order to form the desired chain. The first approach to reconstructing an unknown sequence based on graph theory has been described by Lysov et al.[4, 6, 7]. In order to begin the computational phase with the approach, one needs a digraph which is built from the spectrum (a set of some k -long oligonucleotides) as follows: each oligonucleotide from the spectrum becomes a vertices, two vertices are connected by an arc if the i rightmost nucleotides of the first point overlap with the i leftmost nucleotides of the second one. In such graphs either Hamiltonian [4] or Eulerian paths [6] corresponding to the DNA chains, are looked for. We will refer to these graphs as DNA graphs or DNA labelled graphs. By definition of DNA graph and DNA labelled graph in the following, a DNA graph is a DNA labelled graph but a DNA labelled graph is not necessary a DNA graph. In [9], Wang et al. give some conditions to characterizes which DNA labelled graphs are DNA graphs. We also give some discussions about this problem in section 4.

The following definitions will be used.

Definition 1.1. A graph is a p -graph if given any pair x, y of vertices (x possibly equal

to y), there are at most p parallel arcs from x to y .

For integers $k \geq 2$ and $\alpha \geq 1$, let $\mathbb{Z}_\alpha = \{0, 1, \dots, \alpha - 1\}$ and $\mathbb{Z}_\alpha^k = \{(a_1, \dots, a_k) \mid a_j \in \mathbb{Z}_\alpha, 1 \leq j \leq k\}$.

Definition 1.2. Let $k \geq 2$, $1 \leq i \leq k$ and $\alpha \geq 1$ be three integers. We say that a 1-graph D can be $(k, i; \alpha)$ -labelled if there exists a mapping $l : x \rightarrow l(x) = (l_1(x), \dots, l_k(x))$ from $V(D)$ to \mathbb{Z}_α^k such that

$$(x, y) \in E(D) \Leftrightarrow (l_{k-i+1}(x), \dots, l_k(x)) = (l_1(y), \dots, l_i(y)).$$

We call such a mapping a $(k, i; \alpha)$ -labelling of D and use $S_{k,i}^\alpha$ to denote the class of 1-graphs that can be $(k, i; \alpha)$ -labelled.

Since DNA uses only four letters $\{A, T, C, G\}$, we consider the special case $\alpha = 4$. We give the definition of DNA labelled graph in the following:

Definition 1.3. A digraph is a DNA labelled graph if and only if there are k, i ($k \geq 2$, $1 \leq i \leq k$) such that $D \in S_{k,i}^4$.

This implies that $\bigcup_{k=2}^{\infty} \bigcup_{i=1}^k S_{k,i}^4$ is the set of all DNA labelled graphs. Moreover, in 2008, Wang et al.[9] prove that every graph in $S_{k,i}^\alpha$ is a DNA labelled graph where $k, i, \alpha \in \mathbb{N}$ satisfying $k \geq 2$, $1 \leq i \leq k$ and $\alpha \geq 1$.

Definition 1.4. Let $k \geq 2$ and $\alpha \geq 1$ be two integers. We say that a 1-graph D can be (k, α) -labelled if there exists an mapping $l : x \rightarrow l(x) = (l_1(x), \dots, l_k(x))$ from $V(D)$ to \mathbb{Z}_α^k such that

- (a). l is a $(k, k - 1; \alpha)$ -labeling of D ;
- (b). all labels are different; (i.e. $l(x) \neq l(y)$ if $x \neq y \forall x, y \in V(D)$).

We call such a mapping a (k, α) -labelling of D and use S_k^α to denote the class of 1-graphs that can be (k, α) -labelled.

Definition 1.5. A digraph D is a DNA graph if and only if there exists some $k \geq 2$ such that $D \in S_k^4$.

Definition 1.6. The directed de Bruijn graph $B(k, \alpha)$ is a digraph which has vertices labelled by words of length k over a certain alphabet of cardinality α (there are α^k vertices in such a graph) such that

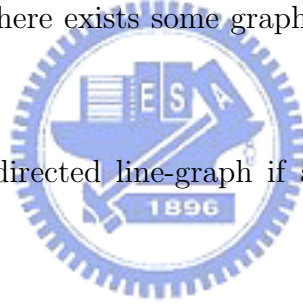
$$(x, y) \in E(B(k, \alpha)) \Leftrightarrow (l_2(x), \dots, l_k(x)) = (l_1(y), \dots, l_{k-1}(y)).$$

In fact, S_k^α is the set of induced subgraphs of $B(k, \alpha)$. Notice that if D can be (α, k) -labelled and has α^k vertices, then D is the de Bruijn graph $B(k, \alpha)$. In 2002, Jacek et al.[3] prove that we can recognize de Bruijn graph in polynomial time.

Definition 1.7. The adjoint $L(D)$ of a digraph D is the 1-graph with vertex set $E(D)$ such that there is an arc from a vertex x to a vertex y in $L(D)$ if and only if the head of the arc x in D is the tail of the tail of the arc y in D .

A graph D' is an adjoint if there exists some graph D such that D' is the adjoint of D .

Definition 1.8. A graph is a directed line-graph if and only if it is the adjoint of a 1-graph.



We give some notations in the following:

Notations Let D be a digraph. For $x \in V(D)$, let $\Gamma^+(x) = \{y \in V(D) | (x, y) \in E(D)\}$, $\Gamma^-(x) = \{y \in V(D) | (y, x) \in E(D)\}$. The outdegree of x , denoted by $d^+(x)$, is the number of vertices in $\Gamma^+(x)$, i.e. $d^+(x) = |\Gamma^+(x)|$. The indegree of x , denoted by $d^-(x)$, is the number of vertices in $\Gamma^-(x)$, i.e. $d^-(x) = |\Gamma^-(x)|$. The minimum outdegree (minimum indegree) of D is $\delta^+(D) = \min\{d^+(x) | x \in V(D)\}$ ($\delta^-(D) = \min\{d^-(x) | x \in V(D)\}$). The minimum semidegree of D is $\delta^0(D) = \min\{\delta^+(D), \delta^-(D)\}$. The maximum outdegree (maximum indegree) of D is $\Delta^+(D) = \max\{d^+(x) | x \in V(D)\}$ ($\Delta^-(D) = \max\{d^-(x) | x \in V(D)\}$). The maximum semidegree of D is $\Delta^0(D) = \max\{\Delta^+(D), \Delta^-(D)\}$.

2 DNA Graphs

2.1 Characterization of directed line-graph

In this section, we have directed line-graph can be recognized in polynomial time.

Theorem 2.1. [2] *Let H be the adjoint of graph G . Then there is an Eulerian path/circuit in G if and only if there is a Hamiltonian path/circuit in H .*

Since directed line-graphs are special cases of adjoint, we have the following corollary:

Corollary 2.2. [2] *Let H be the directed line-graph of a 1-graph G . Then there is an Eulerian path/circuit in G if and only if there is a Hamiltonian path/circuit in H .*

Since finding Eulerian path/circuit in a graph can be done in polynomial time, finding Hamiltonian path/circuit in an adjoint also can be done in polynomial time.

Theorem 2.3. [1] *A 1-graph H is the adjoint of a graph if and only if the following holds for any pair $x, y \in V(H)$:*

$$\Gamma^+(x) \cap \Gamma^+(y) \neq \emptyset \Rightarrow \Gamma^+(x) = \Gamma^+(y).$$

By definition 1.6 and 1.7, a directed line-graph is an adjoint but an adjoint is not necessary a directed line-graph. As an example, one can easily check that the graphs G_1 , G_2 and G_3 of Figure 1 are adjoints but not directed line-graphs.

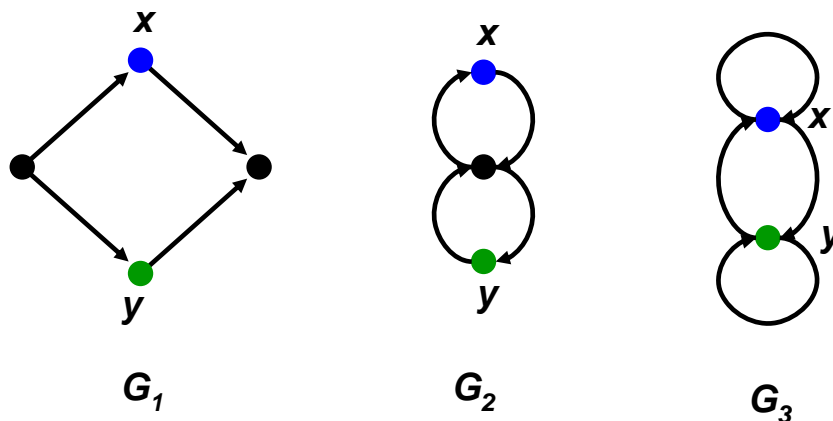


Figure 1: The graphs G_1 , G_2 and G_3 .

The next theorem characterizes which adjoints are directed line-graphs and we give an alternative proof in the following:

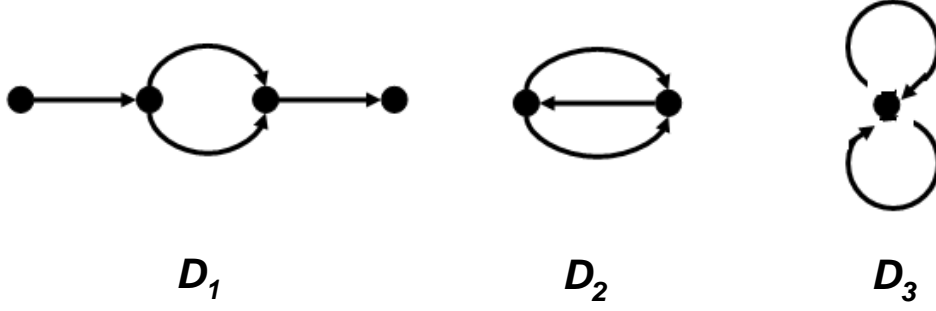


Figure 2: The graphs D_1 , D_2 and D_3 .

Theorem 2.4. [2] *An adjoint is a directed line-graph if and only if it contains none of the digraphs G_1 , G_2 and G_3 as its subgraph.*

Proof. (\Rightarrow) Assume H is the directed line-graph of a 1-graph G . Suppose that H contains one of G_1 , G_2 and G_3 as its subgraph. It is easy to check that G must contain one of D_1 , D_2 and D_3 as its subgraph. It contradicts to that G is a 1-graph. Hence, if an adjoint is a directed line-graph, then it contains none of the digraphs G_1 , G_2 and G_3 as its subgraph.

(\Leftarrow) Let H be the adjoint of a graph G and assume that H contains none of G_1 , G_2 and G_3 as its subgraph.

Case 1 : If G is a 1-graph, then the proof is completed.

Case 2 : If G is not a 1-graph, then we only need to construct a 1-graph G' such that H is also the adjoint of G' . This is done in the following way: We first set G' equal to G . Then, as long as G' is not a 1-graph, we consider any pair x, y of vertices in G' with at least two parallel arcs linking x to y . Since G_3 is not a subgraph of H , these two vertices x and y are distinct. Moreover, since G_1 and G_2 are not subgraphs of H , $\Gamma^-(x) = \phi$ or $\Gamma^+(y) = \phi$. Therefore, we can apply the following changes to G' , where e_1, e_2, \dots, e_p ($p > 1$) are the parallel arcs from x to y :

if $\Gamma^-(x) = \phi$ then

replace x by x_1, \dots, x_p and each arc e_i by an arc (x_i, y) , $i = 1, \dots, p$;

replace each arc (x, z) , with $z \neq y$, by an arc (x_i, z) for some i ;

else ($\Gamma^+(y) = \phi$)

replace y by y_1, \dots, y_p and each arc e_i by an arc (x, y_i) , $i = 1, \dots, p$;

replace each arc (z, y) , with $z \neq x$, by an arc (z, y_i) for some i ;

After these changes, H is still the adjoint of G' . Indeed, the above changes do not disconnect two arcs of G' that formed a path. Moreover, the number of parallel arcs is strictly decreased; thus after a finite number of steps, the graph G' will be the 1-graph we are looking for. The proof is completed. ■

Corollary 2.5. [2] *A 1-graph H is a directed line-graph if and only if the following holds for any pair $x, y \in V(H)$*

$$\Gamma^+(x) \cap \Gamma^+(y) \neq \phi \Rightarrow (\Gamma^+(x) = \Gamma^+(y) \wedge \Gamma^-(x) \cap \Gamma^-(y) = \phi).$$

We give an alternative proof in the following:

Proof. (\Rightarrow) Since the graph H is a directed line-graph, it is also an adjoint and therefore, by Theorem 2.3. $\Gamma^+(x) \cap \Gamma^+(y) \neq \phi$ already implies $\Gamma^+(x) = \Gamma^+(y)$. Suppose, on the contrary. It is easy to check that if for a pair $x, y \in V(H)$ we have $\Gamma^-(x) \cap \Gamma^-(y) \neq \phi$, then the graph must contains at least one of G_1, G_2 and G_3 as its subgraphs. It contradicts to Theorem 2.4.

(\Leftarrow) By Theorem 2.3, we know that the graph must be an adjoint. Moreover, since all three graphs G_1, G_2 and G_3 there is a pair x, y such that $\Gamma^+(x) \cap \Gamma^+(y) \neq \phi$ and $\Gamma^-(x) \cap \Gamma^-(y) \neq \phi$, the given graph can not have G_1, G_2 and G_3 as its subgraph. Hence by Theorem 2.4 we have the graph is a directed line-graph. ■

It follows from Corollary 2.5 that recognizing directed line-graphs can be done in $O(n^3)$ time.

2.2 Some properties of the classes S_k^∞

In this section, we will only consider 1-graphs. Moreover, we use S_k^∞ to denote the class of 1-graphs H for which there exists an integer $\alpha > 0$ such that H can be (k, α) -labelled.

Theorem 2.6. [2] *Let $k \geq 2$ be an integer, G be a graph belonging to S_k^∞ and H be its directed line-graph. Then H belongs to S_{k+1}^∞ .*

We give an alternative proof in the following:

Proof. Let $G \in S_k^\infty$ and H be its directed line-graph, then by definition, there exists

an integer $\alpha \in \mathbb{N}$ and a mapping $l : x \rightarrow l(x) = (l_1(x), \dots, l_k(x))$ from $V(G)$ to \mathbb{Z}_α^k such that

$$(x, y) \in E(G) \Leftrightarrow (l_2(x), \dots, l_k(x)) = (l_1(y), \dots, l_{k-1}(y)).$$

We assign a new mapping $l' : (x, y) \rightarrow l'(x, y) = (l'_1(x, y), \dots, l'_{k+1}(x, y)) = (l_1(x), l_2(x), \dots, l_k(x), l_k(y))$ from $V(H) = E(G)$ to \mathbb{Z}_α^{k+1} .

Claim: l' is a $(k+1, \infty)$ -labeling of H .

(1). Since $G \in S_k^\infty$, it follows that all labels in H are different.

(2). Let $v_a = (x_1, x_2)$ and $v_b = (x_3, x_4)$ be two vertices of H . It remains to prove that

$$(v_a, v_b) \in E(H) \Leftrightarrow (l'_2(v_a), \dots, l'_{k+1}(v_a)) = (l'_1(v_b), \dots, l'_k(v_b)).$$

Since $(x_1, x_2), (x_3, x_4) \in E(G)$, $(l_2(x_1), \dots, l_k(x_1)) = (l_1(x_2), \dots, l_{k-1}(x_2))$ and $(l_2(x_3), \dots, l_k(x_3)) = (l_1(x_4), \dots, l_{k-1}(x_4))$. We now have the following equivalences:

$$\begin{aligned} & (v_a, v_b) \in E(H) \\ \Leftrightarrow & x_2 = x_3 \\ \Leftrightarrow & l'(v_b) = (l'_1(v_b), \dots, l'_{k+1}(v_b)) = (l_1(x_3), \dots, l_k(x_3), l_k(x_4)) \\ & = (l_1(x_2), \dots, l_k(x_2), l_k(x_4)) = (l_2(x_1), \dots, l_k(x_1), l_k(x_2), l_k(x_4)). \\ & l'(v_a) = (l'_1(v_a), \dots, l'_{k+1}(v_a)) = (l_1(x_1), \dots, l_k(x_1), l_k(x_2)). \\ \Leftrightarrow & (l'_2(v_a), \dots, l'_{k+1}(v_a)) = (l'_1(v_b), \dots, l'_k(v_b)). \end{aligned}$$

Hence, by the above argument, the proof is completed. ■

Theorem 2.7. [2] *A graph is a directed line-graph of a 1-graph if and only if it belongs to S_2^∞ .*

We give an alternative proof in the following:

Proof. (\Rightarrow) Let H be a directed line-graph of a 1-graph G . Without loss of generality, assume $V(G) = \{0, 1, 2, \dots, |V(G)| - 1\}$. Then each vertex x of H corresponds to an arc (i, j) of G where $i, j \in V(G)$. Consider the mapping $l : x \rightarrow l(x) = (i, j)$ from $V(H)$ to $\mathbb{Z}_{|V(G)|}^2$. Since G is a 1-graph, all labels of l are different. Hence l is a $(2, |V(G)|)$ -labeling of H and $H \in S_2^\infty$.

(\Leftarrow) Let $H \in S_2^\infty$. Then there exists $\alpha \in \mathbb{N}$ and a mapping $l : x \rightarrow l(x)$ from $V(H)$ to \mathbb{Z}_α^2 . We now construct a graph G as follows:

(a). Let $V(G) = \mathbb{Z}_\alpha$.

(b). There is an arc from a vertex i to a vertex j in $G \Leftrightarrow$ there is a vertex v with label $l(v) = (x, y)$ in H .

Hence G is a 1-graph since all labels of H are different, and it follows from the construction that H is the directed line-graph of G . ■

Theorem 2.8. [2] *Let $k > 2$ be an integer. Then $S_k^\infty \subsetneq S_d^\infty$ for $d = 2, 3, \dots, k - 1$.*

We give an alternative proof in the following:

Proof. It suffices to prove that $S_k^\infty \subsetneq S_{k-1}^\infty$ for $k > 2$. We prove $S_k^\infty \subseteq S_{k-1}^\infty$ first.

Let H be a digraph in S_k^∞ . By definition of S_k^∞ , there exists an integer α such that $H \in S_k^\alpha$. Let l be a (α, k) -labeling of H and φ be an isomorphism from \mathbb{Z}_α^2 to \mathbb{Z}_{α^2} . We assign a new mapping $l' : x \rightarrow l'(x) = (l'_1(x), \dots, l'_{k-1}(x))$ from $V(H)$ to $\mathbb{Z}_{\alpha^2}^{k-1}$ by $l'_i(x) = \varphi(l_i(x), l_{i+1}(x))$, $i = 1, \dots, k - 1$. It is easy to verify that l' is a $(k - 1, \alpha^2)$ -labeling of H . Hence $H \in S_{k-1}^\infty$. Therefore, $S_k^\infty \subseteq S_{k-1}^\infty$.

Second, we show that this inclusion is strict by giving an example in the following. Since we give the labels in Figure 3, $H \in S_3^\infty$. Suppose $H \in S_k^\infty$ for some integer $k \geq 2$.

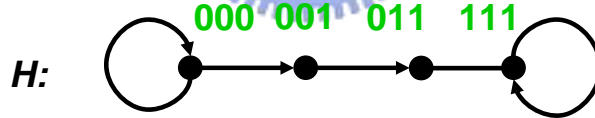


Figure 3: $H \in S_3^\infty$ but $H \notin S_k^\infty$ for $k \geq 4$.

Since the distance between two loops is 3, $k \leq 3$. Hence $H \notin S_k^\infty$ for $k \geq 4$. Therefore, this inclusion is strict. We have the proof. ■

In [2], they give an algorithm, called PROPAGATION ALGORITHM, that when giving a graph H and an integer $k \geq 2$ this algorithm can determine whether H belongs to S_k^∞ or not. If $H \in S_k^\infty$, then the algorithm produced an (∞, k) -labeling of H .

PROPAGATION ALGORITHM:

1. set $l_i(v) = 0$ for each vertex v in H and for all $i = 1, \dots, k$; set $\alpha := 0$;
2. **while** there exists a vertex v in H with a label component equal to 0 **do**

- set $\alpha := \alpha + 1$;
- choose a label component $l_q(v)$ equal to 0 and fix $l_q(v) := \alpha$;
- determine the set L containing all pairs (v, i) such that $l_i(v) = 0$ and either v has a successor w with $l_{i-1}(w) = \alpha$ or v has a predecessor w with $l_{i+1}(w) = \alpha$;
- while** $L \neq \phi$ **do**
- choose any pair (v, i) in L , set $l_i(v) := \alpha$ and update L ;
- end while.**
- end while.**
3. **if** two vertices have the same label **then** STOP: $H \notin S_k^\infty$;
 4. **if** no arc is linking vertex v to vertex w in H while $(l_2(v), \dots, l_k(v)) = (l_1(w), \dots, l_{k-1}(w))$ **then** STOP: $H \notin S_k^\infty$;
 5. STOP: a (∞, k) -labeling of H has been determined.

The complexity of PROPAGATION ALGORITHM was modified to $O(n^k \log(nk))$ in [3] where $n = |H|$. Moreover, they formulate an algorithm which answers the question whether a given graph H is a directed de Bruijn graph and the complexity of this algorithm is $O(n^2 \log^2 n)$ where $n = |H|$. Therefore, we can correctly recognize directed de Bruijn graphs in polynomial time.

1. count vertices which have a loop—the number of such vertices is the cardinality α' of the alphabet;
2. count all vertices of the graph—the number n of all vertices is used to establish the length k of a label: $k = \frac{\log n}{\log \alpha'} = \log_{\alpha'} n$;
if k is not an integer larger than 1
then STOP: H is not a directed de Bruijn graph;
3. apply Propagation Algorithm;
4. **if** PROPAGATION ALGORITHM ended with an $(\alpha'; k)$ -labeling of H
(that is, if it stopped at Step 5. with $\alpha = \alpha'$)
then STOP: H is a de Bruijn graph;

else STOP: H is not a de Bruijn graph.

2.3 Some properties of the classes S_k^α

In the previous section, we have studied the case where there is no upper bound for the size of the alphabet used for the label components. In the case of DNA graphs, all label components must be chosen in the set \mathbb{Z}_4 . Notice first that by definition of S_k^α , we have $S_k^\alpha \subseteq S_k^\beta$ for all $\beta \geq \alpha$. It follows from Theorem 2.8. that $S_k^\alpha \subset S_2^\infty$ for any $k > 2$ and $\alpha > 0$. Moreover, if a graph D with n vertices belongs to S_k^∞ , then it also belongs to S_k^{nk} . In fact, this last property can be improved as stated in the following Theorem.

Theorem 2.9. [2] *If $D \in S_k^\infty$ then $D \in S_k^{n+p(k-1)}$ where n is the number of vertices and p the number of connected components of the underlying undirected graph.*

A question that naturally arises is the following one: knowing that a graph D is in S_k^∞ , which is the smallest integer α such that D is in S_k^α ? This number will be denoted by $\alpha_k(D)$. It has been shown in the proof of Theorem 2.8. that $\alpha_{k-1}(D) \leq \alpha_k^2(D)$. Hence we get the following proposition:

Proposition 2.10. [2] *If $D \in S_k^\infty$, then $D \notin S_k^\alpha$ for all $\alpha < \lceil \sqrt{\alpha_{k-1}(D)} \rceil$.*

We do not know any polynomial algorithm for determining $\alpha_k(D)$. However, if $k = 2$ the problem can be solved in polynomial time as shown below.

Theorem 2.11. [2] *Let $D \in S_2^\infty$. The problem of determining $\alpha_2(D)$ can be solved in polynomial time.*

In [2], they give some open problems. Some of these problems has been solved in [3] and [5].

- Given a graph $D \in S_2^\infty$, the largest integer L such that $D \in S_L^\infty$ can be determined in polynomial time in [3]. Moreover, they prove that $L(D) = 2n$ is a threshold value for which the following is true: $D \in S_{L(D)}^\infty \Leftrightarrow D \in S_k^\infty$ for all $k \geq 2$ where $n = |D|$.
- In [4], they show that it is NP-hard to decide whether

- $D \in S_k^\alpha$, for any fixed $k \geq 3$, with D and α as the input;
- $D \in S_k^\alpha$, for any fixed $\alpha \geq 3$, with D and k as the input;
- $D \in S_\infty^\alpha = \bigcup_{k=1}^{\infty} S_k^\alpha$, for any fixed $\alpha \geq 3$, with D as the input.



3 DNA labelled Graphs

3.1 The relationship between DNA labelled graphs and DNA graphs

By definition of DNA labelled graphs and DNA graphs, we have that $S_k^4 \subseteq S_{k,k-1}^4$. This implies the following:

Theorem 3.1. [9] *If a digraph D is a DNA graph, then D is a DNA labelled graph.*

We can easily check that the digraphs G_1 , G_2 and G_3 in Figure 1 are DNA labelled graphs. Moreover, we will prove that G_1 , G_2 and G_3 are not DNA graphs in the following theorem:

Theorem 3.2. [9] *Let D be a DNA graph. Then D contains none of the digraphs G_1 , G_2 and G_3 as its subgraph.*

Proof. [9] Let D be a DNA graph. By definition 1.5, there exist an integer $k \geq 2$ such that $D \in S_k^4$. Let $l : x \rightarrow l(x) = (l_1(x), \dots, l_k(x))$ be a $(k, 4)$ -labeling of D . This implies that all labels of l are different. Suppose, on the contrary, that D contains at least one of the digraphs G_1 , G_2 and G_3 as its subgraph. Without loss of generality, assume that D contains G_1 as its subgraph. Consider the point x, y shown in Figure 1. Since $\Gamma^+(x) \cap \Gamma^+(y) \neq \emptyset$ and $\Gamma^-(x) \cap \Gamma^-(y) \neq \emptyset$, $(l_2(x), \dots, l_k(x)) = (l_2(y), \dots, l_k(y))$ and $(l_1(x), \dots, l_{k-1}(x)) = (l_1(y), \dots, l_{k-1}(y))$, respectively. It follows that $l(x) = l(y)$. It is contrary to that all labels of l are different. Therefore, D contains none of the digraphs G_1 , G_2 and G_3 as its subgraph. The proof is completed. ■

By Theorem 3.2., we know that none of the digraphs G_1 , G_2 and G_3 is a DNA graph. Combining Theorem 3.1. and Theorem 3.2., we can conclude that $S_k^4 \subsetneq S_{k,k-1}^4$. That is a DNA graph is a DNA labelled graph but a DNA labelled graph is not necessary a DNA graph. The next theorem characterizes which DNA labelled graphs under some conditions are DNA graphs.

Theorem 3.3. [9] *Let $k \geq 2$ and $\alpha \geq 1$ be two integers and D be a digraph in $S_{k,k-1}^4$ with $\delta^0(D) \geq 1$. Then D belongs to S_k^4 if and only if it contains none of the digraphs G_1 , G_2 and G_3 as its subgraph.*

We give an alternative proof in the following:

Proof. (\Rightarrow) If $D \in S_k^4$, then D is a DNA graph. The necessity follows from Theorem 3.2.

(\Leftarrow) Assume $D \in S_{k,k-1}^4$ with $\delta^0(D) \geq 1$ and D contains none of the digraphs G_1 , G_2 and G_3 as its subgraph. Let l be a $(k, k-1; 4)$ -labeling of D . It is enough to prove that all labels are different. Suppose, on the contrary, that there exist two distinct points $x, y \in V(D)$ such that $l(x) = l(y)$. Then $\Gamma^+(x) = \Gamma^+(y)$ and $\Gamma^-(x) = \Gamma^-(y)$. Since $\delta^0(D) \geq 1$, we have $\Gamma^+(x) \neq \phi$ and $\Gamma^-(x) \neq \phi$.

Case 1: If $x \in \Gamma^+(x)$. Since $\Gamma^+(x) = \Gamma^+(y)$, $x \in \Gamma^+(y)$. Moreover, since $x \in \Gamma^+(x) \cap \Gamma^+(y)$ and $l(x) = l(y)$, $x, y \in x \in \Gamma^+(x) \cap \Gamma^+(y)$. This implies that G_3 is a subgraph of D , a contradiction.

Case 2: If $x \notin \Gamma^+(x)$. Since $\delta^0(D) \geq 1$, let $u, v \in V(D)$, $v \in \Gamma^+(x)$ and $u \in \Gamma^-(x)$.

Subcase 1: If $v \neq u$, then G_1 is a subgraph of D , a contradiction.

Subcase 2: If $v = u$, then G_2 is a subgraph of D , a contradiction.

The proof is completed. ■

The following example shows that there exist DNA labelled graphs which contain none of the digraphs G_1 , G_2 and G_3 as the partial subgraph which are not DNA graphs.

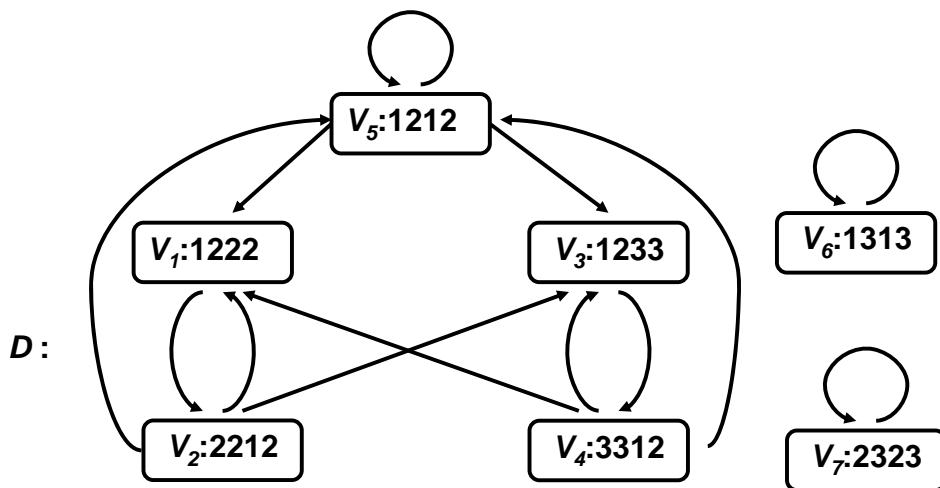


Figure 4: D is a DNA labelled graph but not a DNA graph.

Example 3.4. The graph D shown in Figure 4 is a DNA labelled graph. It is easy to

verify that D contains none of the digraphs G_1 , G_2 and G_3 as its subgraph. Suppose D is a DNA graph. By definition 1.5, there exists some integer $k \geq 2$ such that $D \in S_k^4$. Therefore, there exist a mapping $l : x \rightarrow l(x) = (l_1(x), \dots, l_k(x))$ from $V(D)$ to \mathbb{Z}_4^k such that :

- (a). l is a $(k, k - 1; 4)$ -labeling;
- (b). all labels are different; that is $l(x) \neq l(y)$ if $x \neq y$.

Since $(v_1, v_2), (v_2, v_1) \in E(D)$ and $(v_1, v_1) \notin E(D)$, there exists two distinct integers $a, b \in \mathbb{Z}_4$ such that

$$\begin{cases} l_1(v_1) = l_3(v_1) = \dots = l_2(v_2) = l_4(v_2) = \dots = a \\ l_2(v_1) = l_4(v_1) = \dots = l_1(v_2) = l_3(v_2) = \dots = b \end{cases} \quad (3.1)$$

Similarly, there exists two distinct integers $c, d \in \mathbb{Z}_4$ such that

$$\begin{cases} l_1(v_3) = l_3(v_3) = \dots = l_2(v_4) = l_4(v_4) = \dots = c \\ l_2(v_3) = l_4(v_3) = \dots = l_1(v_4) = l_3(v_4) = \dots = d \end{cases} \quad (3.2)$$

Suppose $k \geq 3$. Since $(v_5, v_1), (v_5, v_3) \in E(D)$, we have

$$\begin{cases} (l_2(v_5), \dots, l_k(v_5)) = (l_1(v_1), \dots, l_{k-1}(v_1)) \\ (l_2(v_5), \dots, l_k(v_5)) = (l_1(v_3), \dots, l_{k-1}(v_3)) \end{cases}$$

Hence $l_1(v_1) = l_1(v_3)$ and $l_2(v_1) = l_2(v_3)$. Combining this with (3.1) and (3.2) we can conclude that $l(v_1) = l(v_3)$. This is contrary to (b) in definition 1.4. So $2 \leq k < 3$. Assume $k = 2$. Clearly, the point with loop must have the label $l(v_i) = (l_1(v_i), l_2(v_i))$ satisfying $l_1(v_i) = l_2(v_i)$ where $i = 6, 7$. Without loss of generality, assume $l(v_6) = (0, 0)$ and $l(v_7) = (1, 1)$. Since v_6 and v_7 are isolated points with loop, $l(v_j)$ must belong to the set $S = \{(2, 2), (2, 3), (3, 2)\}$ where $j = 1, 2, 3, 4, 5$. It is easy to verify that $D \notin S_2^4$. Therefore, $D \notin S_k^4$ for all $k \geq 2$. Hence D is a DNA labelled graph but not a DNA graph.

3.2 Some properties of DNA labelled graphs

We will give some properties of DNA labelled graphs in this section. And use these properties to prove that every graph in $S_{k,i}^\alpha$ is a DNA labelled graph where $k, i, \alpha \in \mathbb{N}$ satisfying $k \geq 2$, $1 \leq i \leq k$ and $\alpha \geq 1$.

Theorem 3.5. [9] *Let $k \geq 2$ and $1 \leq i \leq k$ be two integers. If $k \geq 2i - 1$, then $S_{k,i}^4 \subseteq S_{k+a,i}^4$ for any $a \in \mathbb{N}$.*

We give an alternative proof in the following:

Proof. It is enough to prove that $S_{k,i}^4 \subseteq S_{k+1,i}^4$. Let $D \in S_{k,i}^4$, then there exist a mapping $l : x \rightarrow l(x) = (l_1(x), \dots, l_k(x))$ from $V(D)$ to \mathbb{Z}_4^k such that

$$(x, y) \in E(D) \Leftrightarrow (l_{k-i+1}(x), \dots, l_k(x)) = (l_1(y), \dots, l_i(y)).$$

We assign a new mapping

$$l' : x \rightarrow l'(x) = (l'_1(x), \dots, l'_{k+1}(x)) = (l_1(x), \dots, l_{k-i+1}(x), l_{k-i+1}(x), \dots, l_k(x))$$

from $V(D)$ to \mathbb{Z}_4^{k+1} . Claim: l' is a $(k+1, i; 4)$ -labeling of D in the following. By definition of l' , we have

$$\begin{cases} l'_j(x) = l_j(x), & j \in \{1, 2, \dots, k-i+1\} \\ l'_j(x) = l_{j-1}(x), & j \in \{k-i+2, \dots, k+1\} \end{cases}$$

Since $k \geq 2i-1$, $k-i+1 \geq i$. It follows that

$$\begin{aligned} (x, y) \in E(D) &\Leftrightarrow (l_{k-i+1}(x), \dots, l_k(x)) = (l_1(y), \dots, l_i(y)). \\ &\Leftrightarrow (l'_{k-i+2}(x), \dots, l'_{k+1}(x)) = (l'_1(y), \dots, l'_i(y)). \end{aligned}$$

Therefore, l' is a $(k+1, i; 4)$ -labeling of D which implies that $D \in S_{k+1,i}^4$. The proof is completed. ■

Theorem 3.6. [9] *Let $k \geq 2$ and $1 \leq i \leq k$ be two integers. Then $S_{k,i}^4 \subseteq S_{2i,i}^4$. Furthermore, $S_{k,i}^4 = S_{2i,i}^4$ when $k \geq 2i$.*

We give an alternative proof in the following:

Proof. Let $D \in S_{k,i}^4$, then there exist a mapping $l : x \rightarrow l(x) = (l_1(x), \dots, l_k(x))$ from $V(D)$ to \mathbb{Z}_4^k such that

$$(x, y) \in E(D) \Leftrightarrow (l_{k-i+1}(x), \dots, l_k(x)) = (l_1(y), \dots, l_i(y)).$$

We assign a new mapping

$$l' : x \rightarrow l'(x) = (l'_1(x), \dots, l'_{2i}(x)) = (l_1(x), \dots, l_i(x), l_{k-i+1}(x), \dots, l_k(x)).$$

It is easy to see that

$$\begin{aligned} (x, y) \in E(D) &\Leftrightarrow (l_{k-i+1}(x), \dots, l_k(x)) = (l_1(y), \dots, l_i(y)). \\ &\Leftrightarrow (l'_{i+1}(x), \dots, l'_{2i}(x)) = (l'_1(y), \dots, l'_i(y)). \end{aligned}$$

Therefore, l' is a $(2i, i; 4)$ -labeling of D which implies that $D \in S_{2i,i}^4$. Hence $S_{k,i}^4 \subseteq S_{2i,i}^4$.

If $k = 2i$, then it is trivial that $S_{k,i}^4 = S_{2i,i}^4$. Let $k > 2i$. Since $2i > 2i - 1$, by Theorem 3.5. we have $S_{2i,i}^4 \subseteq S_{2i+a,i}^4$ for all $a \in \mathbb{N}$. Hence $S_{2i,i}^4 \subseteq S_{k,i}^4$. Combining the above argument we have $S_{k,i}^4 = S_{2i,i}^4$ when $k \geq 2i$. \blacksquare

It is possible that $S_{k,i}^4 \subsetneq S_{2i,i}^4$ when $i \leq k < 2i$. For example, the digraph D we give in Figure 4 is in $S_{4,2}^4$. Suppose $D \in S_{2,2}^4$, and l be a $(2, 2; 4)$ -labeling of D . Since $(v_5, v_1) \in E(D)$, we have $(l_1(v_5), l_2(v_5)) = (l_1(v_1), l_2(v_1))$. So $(v_1, v_5) \in E(D)$, a contradiction implying that $D \notin S_{2,2}^4$. Suppose $D \in S_{3,2}^4$ and l be a $(3, 2; 4)$ -labeling of D . Without loss of generality, assume $l(v_6) = (0, 0, 0)$, $l(v_7) = (1, 1, 1)$ and $l(v_5) = (2, 2, 2)$. It is easy to verify that $l(v_1) = l(v_2) = l(v_3) = l(v_4) = (2, 2, 2)$. This implies that $(v_3, v_4), (v_4, v_3) \in E(D)$, a contradiction. Hence $D \notin S_{3,2}^4$. Therefore both $S_{3,2}^4$ and $S_{2,2}^4$ are the proper subset of $S_{4,2}^4$.

By Theorem 3.6. and Definition 1.3, we immediately have the following corollary.

Corollary 3.7. [9] *A digraph D is a DNA labelled graph if and only if there exists a positive integer i such that $D \in S_{2i,i}^4$.*

Theorem 3.8. [9] *Let $k \geq 2$, $1 \leq i \leq k$ and $a \geq 1$ be three integers. If $D \in S_{k+a,i+a}^4$, then there exists a digraph D' such that D is a spanning subgraph of D' .*

We give an alternative proof in the following:

Proof. It is enough to prove that if $D \in S_{k+1,i+1}^4$, then there exists a digraph D' such that D is a spanning subgraph of D' . Let $D \in S_{k+1,i+1}^4$ and l be a $(k+1, i+1; 4)$ -labeling of D . We construct a digraph D' as follows:

(a). $V(D') = V(D)$;

(b). $(x, y) \in E(D) \Leftrightarrow l_{k-i+j}(x) + l_{k-i+j+1}(x) = l_j(y) + l_{j+1}(y) \pmod{4}$ for any $j \in \{1, \dots, i\}$.

If $(x, y) \in E(D)$, then $l_{k-i+j}(x) = l_j(y)$ for any $j \in \{1, \dots, i\}$. This implies that $l_{k-i+j}(x) + l_{k-i+j+1}(x) = l_j(y) + l_{j+1}(y) \pmod{4}$ for any $j \in \{1, \dots, i\}$ and hence $(x, y) \in E(D')$.

Therefore, D is a spanning subgraph of D' . Now, we need to claim $D \in S_{k,i}^4$. Assign a new mapping $l' : x \rightarrow l'(x) = (l'_1(x), \dots, l'_k(x))$ from $V(D')$ to \mathbb{Z}_4^k such that $l'_j(x) =$

$l_j(x) + l_{j+1}(x) \pmod{4}$ for all $j \in \{1, 2, \dots, k\}$. It is easy to see that

$$\begin{aligned} (x, y) \in E(D') &\Leftrightarrow (l_{k-i+1}(x) + l_{k-i+2}(x), \dots, l_k(x) + l_{k+1}(x)) \\ &= (l_1(y) + l_2(y), \dots, l_i(y) + l_{i+1}(y)). \\ &\Leftrightarrow (l'_{k-i+1}(x), \dots, l'_k(x)) = (l'_1(y), \dots, l'_i(y)). \end{aligned}$$

Therefore, l' is a $(k, i; 4)$ -labeling of D' which implies that $D' \in S_{k,i}^4$. The proof is completed. \blacksquare

Theorem 3.9. [9] *Let $k \geq 2$, $1 \leq i \leq k$ and $m \geq 1$ be three integers. Then $S_{km,im}^4 = S_{k,i}^{4m}$.*

We give an alternative proof in the following:

Proof. The main technique of this proof is using quaternary transformation. Let $\psi : p \rightarrow \varphi(p) = (p_1, p_2, \dots, p_m)$ be a quaternary bijection from \mathbb{Z}_{4^m} to \mathbb{Z}_4^m , $D \in S_{k,i}^{4m}$ and l be a $(k, i; 4^m)$ -labeling of D . We assign a new mapping $l' : x \rightarrow l'(x) = (l'_1(x), \dots, l'_{km}(x))$ from $V(D)$ to \mathbb{Z}_4^{km} by $(l'_{(j-1) \times m+1}(x), \dots, l'_{j \times m}(x)) = \psi(l_j(x))$. for any $j \in \{1, 2, \dots, k\}$. It is easy to verify that l' is a $(km, im; 4)$ -labeling of D . Hence $D \in S_{k,i}^{4m}$. Therefore, $S_{k,i}^{4m} \subseteq S_{km,im}^4$. Similarly, we can use another quaternary bijection from \mathbb{Z}_4^m to \mathbb{Z}_{4^m} to prove that $S_{km,im}^4 \subseteq S_{k,i}^{4m}$. Therefore, $S_{km,im}^4 = S_{k,i}^{4m}$. \blacksquare

Corollary 3.10. [9] *Let $k \geq 2$ and $1 \leq i \leq k$ be two integers. Then $S_{k,i}^\alpha \subseteq S_{km,im}^4$ for any $\alpha, m \in \mathbb{N}$ satisfying $4^m \geq \alpha$.*

Proof. [9] Since $\alpha \leq 4^m$, $S_{k,i}^\alpha \subseteq S_{k,i}^{4m}$. It follows that $S_{k,i}^\alpha \subseteq S_{km,im}^4$ from Theorem 3.9. \blacksquare

Corollary 3.11. [9] *Let $m \in \mathbb{N}$. Then $S_{km,im}^4 \subseteq S_{k(m+a),i(m+a)}^4$ for any $\alpha \in \mathbb{N}$.*

Proof. [9] $S_{km,im}^4 = S_{k,i}^{4m} \subseteq S_{k,i}^{4^{m+a}} = S_{k(m+a),i(m+a)}^4$. \blacksquare

Corollary 3.12. [9] *For integers k, i, α satisfying $k \geq 2$, $1 \leq i \leq k$ and $\alpha \geq 1$, every graph in $S_{k,i}^\alpha$ is a DNA labelled graph.*

Proof. [9] Let $D \in S_{k,i}^\alpha$. Choose an integer m such that $4^m \leq \alpha$. By corollary 3.10., $D \in S_{km,im}^4$. Hence D is a DNA labelled graph. \blacksquare

By the above argument, we have $\bigcup_{k=2}^{\infty} \bigcup_{i=1}^k S_{k,i}^4 = \bigcup_{\alpha=1}^{\infty} \left(\bigcup_{k=2}^{\infty} \bigcup_{i=1}^k S_{k,i}^\alpha \right)$.

3.3 The relationship between DNA labelled graphs and adjoints

Lemma 3.13. [9] *A digraph D is the adjoint of some digraph H if and only if $D \in S_{2,1}^{|V(H)|}$.*

Proof. The proof is similar to the proof of Theorem 2.7. ■

Theorem 3.14. [9] *A digraph D is the adjoint of some digraph H if and only if $D \in S_{2m,m}^4$, where $4^m \geq |V(H)|$.*

We give an alternative proof in the following:

Proof. (\Rightarrow) If D is the adjoint of some digraph H , then by Lemma 3.13. $D \in S_{2,1}^{|V(H)|}$. Since $4^m \geq |V(H)|$, $D \in S_{2,1}^{4^m}$. It follows that $D \in S_{2m,m}^4$ from Corollary 3.10.

(\Leftarrow) Suppose $D \in S_{2m,m}^4$. By Corollary 3.10. we have $D \in S_{2,1}^{4^m}$. Hence by Lemma 3.13, there exists a digraph H with $|V(H)| = 4^m$ such that D is the adjoint H . The proof is completed. ■

By Corollary 3.7. and Theorem 3.14., we have the following Theorem:

Theorem 3.15. [8] *The digraph D is a DNA labelled graph if and only if D is an adjoint of some graph H .*

Moreover, by Theorem 2.3. and Theorem 3.15., we have recognizing DNA labelled graphs can be done in polynomial-time.

3.4 An equivalence relation of DNA labelled graphs

We start with a very useful definition. Let D be a given digraph. We define a relation \sim (called a friend relation) on $E(D)$ as follows. For every two arcs $e_1 = (x_1, y_1)$, $e_2 = (x_2, y_2)$ in D , $e_1 \sim e_2$ if $x_1 = x_2$ or $y_1 = y_2$ or $(x_2, y_1), (x_1, y_2) \in E(D)$. Clearly, we have the following:

- (a) $e \sim e$ for any $e \in E(D)$;
- (b) $e_1 \sim e_2 \Rightarrow e_2 \sim e_1$.

Theorem 3.16. [8] *Let D be a DNA labelled graph. Then the friend relation is an equivalence relation on $E(D)$.*

The following algorithm will be used.

Algorithm 3.[8]

Input: A digraph $D = (V(D), E(D))$.

Output: $S = \{E_1, \dots, E_n\}$ and n .

Step 0. Set $S := \phi$, $n := 0$ and $E := E(D)$.

Step 1. If $E = \phi$, then stop; Otherwise $n := n + 1$.

Step 2. (Find E_n .)

(0) Let $e \in E$.

(1) Find $F^{(2)}$ for e . ($F^{(2)}$ contains e . $F^{(1)}$ is an arc subset of $E(D)$, the head of each arc in which is the same to the head of the given arc. $F^{(2)}$ is also an arc subset of $E(D)$, the tail of each arc in which is the same to the tail of the given arc.)

(2) For every $e \in F^{(2)}$, find $F^{(1)}$ and set $F^{(1)} := \bigcup_{e \in F^{(2)}} F^{(1)}$.

(3) Set $E_n := F^{(1)} \cup F^{(2)}$, $S := \{E_1, \dots, E_n\}$, $E := E - E_n$ and go to Step 1.

It is easy to verify that Algorithm 3 is a polynomial-time one. We have the following:

Theorem 3.17. [8] *The output $S = \{E_1, \dots, E_n\}$ of Algorithm 3 is a partition of $E(D)$ for a given DNA labelled graph D . Moreover, for any $i \in \{1, \dots, n\}$, E_i is an equivalence class under the friend relation.*

Let D be a DNA labelled graph with $E(D) \neq \phi$ and let $\{E_1, \dots, E_n\}$ be the output of Algorithm 3 for D . For $i = 1, \dots, n$, let

$$A_i = \{x \in V(D) : \text{there exists a point } y \text{ such that } (x, y) \in E_i\},$$

$$B_i = \{x \in V(D) : \text{there exists a point } y \text{ such that } (y, x) \in E_i\}.$$

For two sets A, B , let $A \times B = \{(a, b) | a \in A, b \in B\}$. Therefore, we have the following lemma immediately.

Lemma 3.18. [9] *Let D be a DNA labelled graph and E_i be an equivalence class under the friend relation. Then $E_i = A_i \times B_i$.*

Theorem 3.19. [9] *Let D be a DNA labelled graph without loops and isolated points. Then D has exactly one equivalence class under the friend relation if and only if there exists a partition (A, B) of $V(D)$ such that $E(D) = A \times B$.*

Now, we regard a point and a loop as a path and a cycle respectively.

Theorem 3.20. [9] *Let D be a DNA labelled graph. Then D has $|E(D)|$ equivalence classes under the friend relation if and only if every component of D is a path or a cycle.*

Let D be a DNA labelled graph. Denote the set $\{1, \dots, n\}$ by I , where n is the output of Algorithm 3 for D . We have the following:

Theorem 3.21. [9] *Let D be a DNA labelled graph with $\delta^0(D) \leq 1$. Then $\{A_i\}_{i \in I}$ and $\{B_i\}_{i \in I}$ are two partitions of $V(D)$ such that $E(D) = \bigcup_{i \in I} A_i \times B_i$.*

Let D be a DNA labelled graph. By Section 3.2 we have there exists a positive integer α such that $D \in S_{2,1}^\alpha$. Clearly, the fact that $D \in S_{2,1}^\alpha$ implies that $D \in S_{2,1}^\beta$ for any integer $\beta \geq \alpha$. A question that naturally arises is the following one: knowing that D is a DNA labelled graph, which is the smallest integer α such that $D \in S_{2,1}^\alpha$? This number will be denoted by $\alpha(D)$.

Theorem 3.22. [9] *Let D be a DNA labelled graph and let n be the output of Algorithm 3 for D . Then*

$$\alpha(D) = \begin{cases} n & \text{if } \delta^0(D) \geq 1; \\ n + 1 & \text{if } \delta^0(D) = 0, \text{ and } \delta^+(D) \geq 1 \text{ or } \delta^-(D) \geq 1; \\ n + 2 & \text{if } \delta^0(D) = 0, \text{ and } \delta^-(D) = 0. \end{cases}$$

Let D be a DNA labelled graph. By Corollary 3.7., there exists a positive integer i such that $D \in S_{2i,i}^4$. Moreover, by Corollary 3.11., if $D \in S_{2i,i}^4$, then $D \in S_{2m,m}^4$ for any integer $m \geq i$. We use $i(D)$ to denote the smallest integer i such that $D \in S_{2i,i}^4$.

Theorem 3.23. [9] *Let D be a DNA labelled graph. Then $i(D) = \lceil \log_4 \alpha(D) \rceil$.*

Theorem 3.21. and 3.22. imply the following.

Corollary 3.24. [9] *Let D be a DNA labelled graph and let n be the output of Algorithm 3 for D . Then*

$$i(D) = \begin{cases} \lceil \log_4 n \rceil & \text{if } \delta^0(D) \geq 1; \\ \lceil \log_4(n + 1) \rceil & \text{if } \delta^0(D) = 0, \text{ and } \delta^+(D) \geq 1 \text{ or } \delta^-(D) \geq 1; \\ \lceil \log_4(n + 2) \rceil & \text{if } \delta^0(D) = 0, \text{ and } \delta^-(D) = 0. \end{cases}$$

4 Main Results

Lemma 4.1. *If D is a DNA graph, then $\Delta^0(D) \leq 4$.*

Proof. Since D is a DNA graph, let l be a $(k, 4)$ -labelling of D . Suppose $\Delta^0(D) \geq 5$. W.L.O.G. assume $\Delta^+(D) \geq 5$, then there exists a vertex x in D such that $d^+(x) \geq 5$. Let $\Gamma^+(x) = \{v_1, v_2, v_3, v_4, v_5, \dots\}$. By the definition of DNA graph, we have $(l_2(x), \dots, l_k(x)) = (l_1(v_i), \dots, l_{k-1}(v_i))$ and $l_k(v_i) \in \mathbb{Z}_4$ where $i = 1, 2, 3, 4, 5$. By the Pigeonhole Principle, we have $l(v_i) = l(v_j)$ where $v_i, v_j \in \Gamma^+(x)$. This is contrary to that D is a DNA graph. The proof is complete. ■

In [9], they give an open problem:

Open Problem Give a characterization of DNA labelled graphs which are not DNA graphs.

The following main results is aimed at this open problem. We start from Theorem 3.3. Recall the Theorem 3.3. The proof of Theorem 3.3. is not very hard. But if we want to omit the conditions $D \in S_{k,k-1}^4$ or $\delta^0(D) \geq 1$ of D , then the characterization will be difficult. That is, there exist infinite graphs such that there graphs are DNA labelled graphs but not DNA graphs.

- It is difficult to characterize DNA labelled graphs which are not DNA graphs when we omitting the condition $D \in S_{k,k-1}^4$ in Theorem 3.3.

The graph we shown in Figure 4 is the example that omitting the condition $D \in S_{k,k-1}^4$ in Theorem 3.3. That is, D is a DNA labelled graph with $\delta^0(D) \geq 1$ but D is not a DNA graph.

- It is difficult to characterize DNA labelled graphs which are not DNA graphs when we omitting the condition $\delta^0(D) \geq 1$ in Theorem 3.3.

First, we define a graph D_i in the following. Assume $i \geq 2$ be an integer. Let D_i be a digraph with $V(D_i) = \{v_1, \dots, v_i, v_{i+1}\}$ and $E(D) = \{(v_1, v_1), (v_{i+1}, v_{i+1})\} \cup \{(v_j, v_{j+1}) | j = 1, 2, \dots, i\}$. Suppose $D_i \in S_k^4$ for some integer $k \geq 2$. Since the distance between the loops v_1 and v_{i+1} is $d(v_1, v_{i+1}) = i$, $k \leq i$. If we add enough

isolated points without loop to each D_i , then the new graph D will be an example which belonging to $S_{i,i-1}^4$ with $\delta^0(D) = 0$ but not belonging to S_i^4 . Therefore, there exists infinite graphs such that these graphs belong to $S_{k,k-1}^4$ for some integer $k \geq 2$ but not belong to S_k^4 .

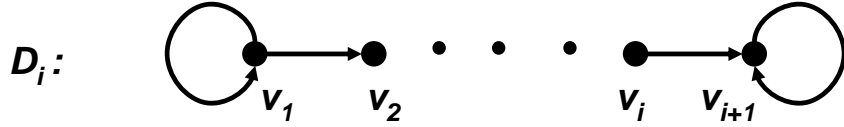


Figure 5: The graph D_i .

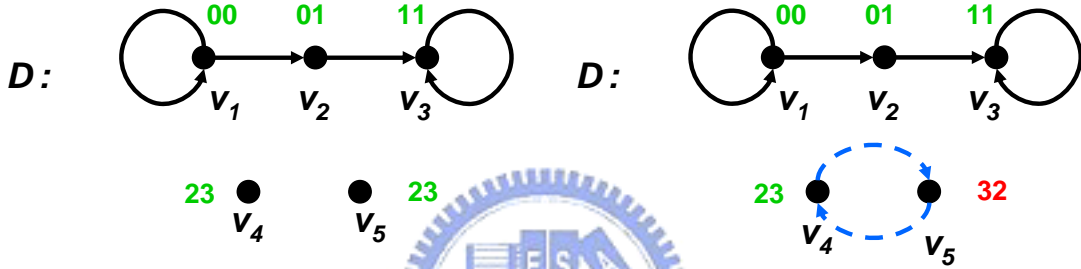


Figure 6: $D \in S_{2,1}^4$ but $D \notin S_2^4$.

We take D_2 as an example in the following.

Example 1: Let D be the graph that we add two isolated points without loop to D_2 . It is easy to verify that $D \in S_{2,1}^4$ (We shown the labels in Figure 6). Suppose $D \in S_2^4$. Without loss of generality, let $l(v_1) = (0, 0)$, $l(v_2) = (0, 1)$ and $l(v_3) = (1, 1)$. Since v_4 and v_5 are isolated points without loop in D , only can use $S = \{(2, 3), (3, 2)\}$ to label them, a contradiction. Hence $D \notin S_2^4$.

If we only consider the graph D with $|D| \leq 6$, then omit the condition $\delta^0(D) \geq 1$ of D is allowed.

Theorem 4.2. Let $k \geq 3$ be an integer and let D be a digraph in $S_{k,k-1}^4$ with $|D| \leq 6$ and D contains none of G_1, G_2 , and G_3 as its subgraph. Then D belongs to S_k^4 if and only if $\Delta^0(D) \leq 4$.

Proof. (\Rightarrow) Let $D \in S_k^4$, then D is a DNA graph by definition. The necessity follows from Lemma 4.1.

(\Leftarrow) Let $\Delta^0(D) \leq 4$ and l be a $(k, k-1; 4)$ -labelling of D .

Claim: $D \in S_k^4$.

It's sufficient to prove that all labels are different. If $\delta^0(D) \geq 1$, then by Theorem 3.3. the proof is completed. Hence, assuming that $\delta^0(D) = 0$. Suppose $D \notin S_k^4$ and minimize the repetition of labels, then there exists two distinct vertices $x, y \in V(D)$ such that $l(x) = l(y)$. Hence $\Gamma^+(x) = \Gamma^+(y)$ and $\Gamma^-(x) = \Gamma^-(y)$. Since D contains no G_1, G_2 , and G_3 as its subgraph, $\Gamma^+(x) = \phi$ or $\Gamma^-(x) = \phi$.

Case 1: If $\Gamma^+(x) = \phi$ and $\Gamma^-(x) \neq \phi$.

Subcase 1: Let $v \in \Gamma^-(x)$ and $v \notin \Gamma^+(v)$.

Without loss of generality, let $l(v) = (l_1(v), \dots, l_k(v)) = (0, 0, \dots, 0, 1)$ and $l(x) = (l_1(x), \dots, l_k(x)) = (0, \dots, 0, 1, 0) = (l_1(y), \dots, l_k(y)) = l(y)$. Then there must exist a point v_1 whose label is $l(v_1) = (l_1(v_1), \dots, l_k(v_1)) = (0, \dots, 0, 1, 1)$ or $(0, \dots, 0, 1, 1, i)$ where $i \in \mathbb{Z}_4$. Otherwise, we can change the label of y to $l(y) = (0, \dots, 0, 1, 1)$. This is contrary to that $l(x) = l(y)$. Similarly, there must exist a point v_2 whose label is $l(v_2) = (l_1(v_2), \dots, l_k(v_2)) = (0, \dots, 0, 1, 2)$ or $(0, \dots, 0, 1, 2, j)$ where $j \in \mathbb{Z}_4$. If $|D| \leq 5$, then we can always change the label of y to $l(y) = (0, \dots, 0, 1, 3)$ such that $l(x) \neq l(y)$. Hence all labels are different, we are done. If $|D| = 6$, then there must exist a point v_3 whose label is $l(v_3) = (l_1(v_3), \dots, l_k(v_3)) = (0, \dots, 0, 1, 3)$ or $(0, \dots, 0, 1, 3, p)$ where $p \in \mathbb{Z}_4$. Hence $V(D) = \{v, x, y, v_1, v_2, v_3\}$.

(1). If $k \geq 4$, then $l(v_i) = (0, \dots, 0, 1, i) \forall i = 1, 2, 3$. Suppose, on the contrary, that there exist i such that $l(v_i) = (0, \dots, 0, 1, i, *)$ where $* \in \mathbb{Z}_4$ and $i = 1, 2, 3$. Then $\Gamma^+(v_i) = \phi$ and $\Gamma^-(v_i) = \phi$. We can change the label of v_i and y to

$$\begin{cases} l(v_i) = (1, \dots, 1, 2, i + 1(\text{mod}4), * + 1(\text{mod}4)) & i = 1, 2, 3 \\ l(y) = (0, \dots, 0, 1, i) & \text{for some } i \end{cases}$$

such that l preserves arcs and nonarcs. This is contrary to that $l(x) = l(y)$. Hence $l(v_i) = (0, \dots, 0, 1, i) \forall i = 1, 2, 3$. Therefore $\Gamma^+(v) = \{x, y, v_1, v_2, v_3\}$. This is contrary to that $\Delta^0(D) \leq 4$. Thus, in this case we have $D \in S_k^4$.

(2). If $k = 3$, then $l(v_i) = (0, 1, i) \forall i = 1, 2, 3$. Suppose, on the contrary, that there exist i such that $l(v_i) = (1, i, *)$ where $* \in \mathbb{Z}_4$ and $i = 1, 2, 3$. Since (v_1, v_2) and (v_1, v_3) might

belong to $E(D)$, we change the label of v_i and y in the following way:

$$\begin{cases} l(v_i) = (2, i + 1(\bmod 4), * + 1(\bmod 4)) & \text{if } i = 1, 2 \\ l(v_3) = (2, 0, 2) \\ l(y) = (0, 1, i) \end{cases} \quad \text{for some } i$$

such that l preserves arcs and nonarcs. This is contrary to that $l(x) = l(y)$. Hence $l(v_i) = (0, 1, i), i = 1, 2, 3$. Therefor $\Gamma^+(v) = \{x, y, v_1, v_2, v_3\}$. This is contrary to that $\Delta^0(D) < 5$. Thus, in this case we have $D \in S_k^4$.

Subcase 2: Let $v \in \Gamma^-(x)$ and $v \in \Gamma^+(v)$.

Without loss of generality, let $l(v) = (l_1(v), \dots, l_k(v)) = (0, 0, \dots, 0, 0)$ and $l(x) = (l_1(x), \dots, l_k(x)) = (0, \dots, 0, 1) = (l_1(y), \dots, l_k(y)) = l(y)$. There must exists a point v_1 whose label is $l(v_1) = (l_1(v_1), \dots, l_k(v_1)) = (0, \dots, 0, 2)$ or $(0, \dots, 0, 2, i)$ where $i \in \mathbb{Z}_4$. Otherwise, we can change the label of y to $l(y) = (0, \dots, 0, 2)$. This is contrary to that $l(x) = l(y)$. If $|D| \leq 4$, W.L.O.G. we can assume $V(D) = \{v, x, y, v_1\}$, then we can always change the label of y to $l(y) = (0, \dots, 0, 3)$ such that $l(x) \neq l(y)$. Hence all labels are different, we are done. If $|D| = 5$, there must exists a point v_2 whose label is $l(v_2) = (l_1(v_2), \dots, l_k(v_2)) = (0, \dots, 0, 3)$ or $(0, \dots, 0, 3, j)$ where $j \in \mathbb{Z}_4$. By the similar argument in Subcase1(1), we have $\Gamma^+(v) = \{v, x, y, v_1, v_2\}$. This is contrary to that $\Delta^0(D) < 5$. Thus, in this case we have $D \in S_k^4$. Otherwise, we can change the label of v_i and y to

$$\begin{cases} l(v_i) = (0, \dots, 0, i + 1(\bmod 4), *) & i = 1, 2, \text{ and } * \in \mathbb{Z}_4 \\ l(y) = (0, \dots, 0, i + 1) \end{cases} \quad \text{for some } i$$

such that l preserves arcs and nonarcs. This is contrary to that $l(x) = l(y)$. If $|D| = 6$, then let $V(D) = \{v, x, y, v_1, v_2, u\}$.

(1). If $\Gamma^+(u) = \phi = \Gamma^-(u)$, then let $l(u) = (l_1(u), \dots, l_k(u)) = (3, \dots, 3, 1)$. By the similar argument, we have $\Gamma^+(v) = \{v, x, y, v_1, v_2\}$. This is contrary to that $\Delta^0(D) \leq 4$. Thus, in this case we have $D \in S_k^4$.

(2). If $\Gamma^+(u) \neq \phi$.

(2.a). If $\{v, x, y\} \subseteq \Gamma^+(u)$, then let $l(u) = (l_1(u), \dots, l_k(u)) = (0, \dots, 0, *)$ where $* \in \{1, 2, 3\}$. Since we minimize the repetition of labels, $* \neq 0$. Otherwise $l(u) = l(v)$. By the similar argument, we have $\Gamma^+(v) = \{v, x, y, v_1, v_2\} = \Gamma^+(u)$. This is contrary to that $\Delta^0(D) \leq 4$. Thus, in this case we have $D \in S_k^4$.

(2.b). If $\{v, x, y\} \not\subseteq \Gamma^+(u)$, then $\varphi v_i, i = 1, 2$ such that $l(v_i) = (0, \dots, 0, i + 1, *)$ and $v_i \in \Gamma^+(u)$. Otherwise, $\Gamma^+(u) = \phi$. W.L.O.G. let $l(v_1) \in \Gamma^+(u)$ and $l(u) = (l_1(u), \dots, l_k(u)) = (p, 0, \dots, 0, 2)$ where $p \in \mathbb{Z}_4$. If $p \neq 0$, then we can change the label of u, v_1 and y to

$$\begin{cases} l(u) = (p + 1, 1, \dots, 1, 3) \\ l(v_1) = (1, \dots, 1, 3, * + 1) \\ l(y) = (0, \dots, 0, 2) \end{cases}$$

such that l preserves arcs and nonarcs. This is contrary to that $l(x) = l(y)$. Hence $p = 0$. Therefor $l(v_2) = (0, \dots, 0, 3)$. Otherwise, we can change the label of v_2 and y to

$$\begin{cases} l(v_2) = (1, \dots, 1, 0, * + 1) \\ l(y) = (0, \dots, 0, 3) \end{cases}$$

such that l preserves arcs and nonarcs. This is contrary to that $l(x) = l(y)$. Hence $\Gamma^+(v) = \{v, x, y, u, v_2\}$. This is contrary to that $\Delta^0(D) < 5$. Thus, in this case we have $D \in S_k^4$.

(3). If $\Gamma^-(u) \neq \phi$.

(3.a). If $v \in \Gamma^-(u)$. Since we minimize the repetition of labels, $l(u) \neq l(x) = l(y)$. Let $l(u) = (l_1(u), \dots, l_k(u)) = (0, \dots, 0, *)$ where $* \in \{2, 3\}$. W.L.O.G. let $l(u) = (0, \dots, 0, 2)$, then $l(v_1) = (0, \dots, 0, 2, *')$ where $*' \in \mathbb{Z}_4$. Then $l(v_2) = (0, \dots, 0, 3)$. Otherwise, let $l(v_2) = (0, \dots, 0, 3, *'')$, then we can change the label of v_2 and y to

$$\begin{cases} l(v_2) = (1, \dots, 1, 0, *'' + 1) \\ l(y) = (0, \dots, 0, 3) \end{cases}$$

such that l preserves arcs and nonarcs. This is contrary to that $l(x) = l(y)$. Hence $\Gamma^+(v) = \{v, x, y, u, v_2\}$. This is contrary to that $\Delta^0(D) < 5$. Thus, in this case we have $D \in S_k^4$.

(3.b). If $v \notin \Gamma^-(u)$. Since $\Gamma^+(x) = \Gamma^+(y) = \phi, x, y \notin \Gamma^-(u)$. Hence one of v_1 and v_2 will belongs to $\Gamma^-(u)$. Otherwise, $\Gamma^-(u) = \phi$. W.L.O.G. let $v_1 \in \Gamma^-(u)$. Then $l(v_1) = (0, \dots, 0, 2), l(u) = (0, \dots, 0, 2, *)$ and $l(v_2) = (0, \dots, 0, 3)$. Otherwise, we can change the label of u, v_1, v_2 and y such that l preserves arcs and nonarcs and $l(x) \neq l(y)$. This is contrary to that $l(x) = l(y)$. Hence $\Gamma^+(v) = \{v, x, y, v_1, v_2\}$. This is contrary to that $\Delta^0(D) \leq 4$. Thus, in this case we have $D \in S_k^4$.

Case 2: If $\Gamma^+(x) \neq \phi$ and $\Gamma^-(x) = \phi$.

It is similar to case1, hence $D \in S_k^4$.

Case 3: If $\Gamma^+(x) = \phi$ and $\Gamma^-(x) = \phi$, then x, y are isolated points without loop. Suppose $|D| = 6$ and $k = 3$. Let $V(D) = \{x, y, v_1, v_2, v_3, v_4\}$. Since $k = 3$, there are $4^3 = 64$ different labels. Consider $\{x, v_1, v_2, v_3, v_4\}$. There are $64 - 5 - 4 \times 5 \times 2 - 4 = 15$ different labels such that if we change the label of y to one of these fifteen labels, then y is still an isolated point without loop and $l(x) \neq l(y)$. Hence $D \in S_k^4$. If $|D| < 6$ and $k > 3$, then we can use the same argument to find a different label for y such that all labels of D are different. Hence $D \in S_k^4$. The proof is completed. ■

If we consider the cases $|D| \geq 6$, then there must be quite a few isolated vertices without loop, the proof will be very tedious. If we only consider the case that D is weakly connected, then there still exist infinite graphs such that these graphs satisfying the sufficient condition of Theorem 4.2. but not belonging to S_k^4 for all $k \geq 2$.

We give examples in the following:

Example 2: See Figure 7. It is easy to verify that D contains none of G_1, G_2, G_3 as its subgraphs and $\Delta^0(D) \leq 4$. We shown the labels in Figure 6. Hence $D \in S_{2,1}^4$. Since D_2 is an induced subgraph of D , suppose $D \in S_k^2$ then $k = 2$. We can easily check that no matter how we change the labels of D either there exists two different vertices with the same labels or there exists two vertices x, y such that $l_2(x) = l_1(y)$ but $(x, y) \notin E(D)$. Hence $D \notin S_2^4$. Therefore, D is not a DNA graph.

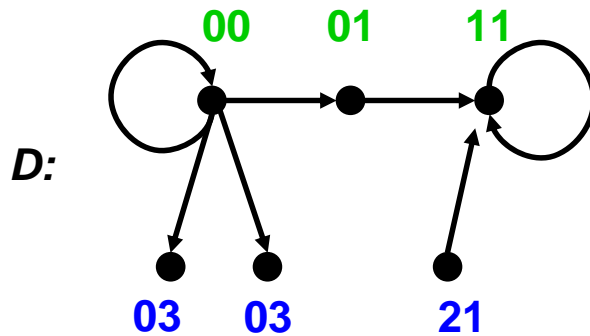


Figure 7: D is weakly connected and satisfying the sufficient condition of Theorem 4.2, but D is not a DNA graph.

For each $D_i, i \geq 3$, we construct a new graph D as follows: first, add two vertices to $\Gamma^+(v_1)$ and two vertices to $\Gamma^-(v_{i+1})$ to form a new graph D'_i . Second, if $d^+(x) = 1$, then

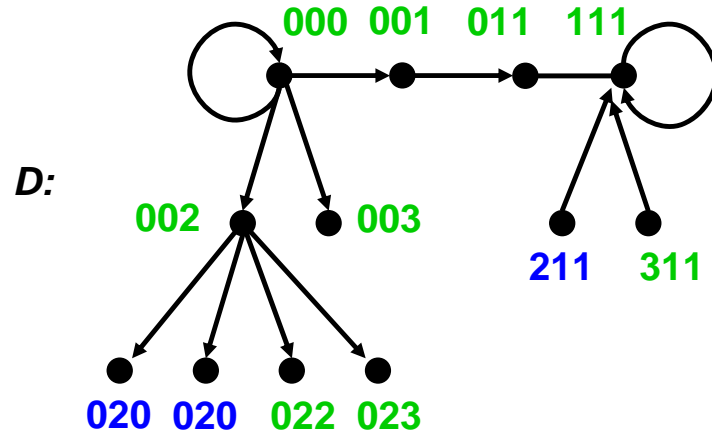


Figure 8: D is weakly connected and satisfying the sufficient condition of Theorem 4.2, but D is not a DNA graph.

add four vertices to $\Gamma^-(x)$. If $d^-(x) = 1$, then add four vertices to $\Gamma^+(x)$. Update D'_i and repeat the second step enough times, we will get D . Figure 8 is the example where D_3 is its induced subgraph and use the same argument as Example 2, we have the graph shown in Figure 7 is weakly connected and satisfying the sufficient condition of Theorem 4.2, but not a DNA graph.



5 Concluding Remarks

Through this study, we have the following three remarks.

1. If D is a DNA graph, then $\Delta^0(D) \leq 4$.
2. It is difficult to characterize DNA labelled graphs which are not DNA graphs when we omit the condition $D \in S_{k,k-1}^4$ or $\delta^0(D) \geq 1$ in Theorem 3.3.
3. It is difficult to characterize DNA labelled graphs which are not DNA graphs when considering D is weakly connected.

Therefore, for future study, we might have to find some more criterions (on graph structures) in order to settle this problem.



References

- [1] C. Berge, Graphes, Dunod, Paris, 1970.
- [2] J. Blazewicz, A. Hertz, D. Kobler and D. de Werra, On some properties of DNA graphs. *Discrete Appl. Math.*, **98**: 1 – 19(1999).
- [3] B. Jacek, F. Piotr, K. Marta and K. Daniel, On the recognition of de Bruijn graphs and their induced subgraphs. *Discrete Math.*, **245**: 81 – 92(2002).
- [4] Y.P. Lysov, V.L. Florentiev, A.A. Khorlyn, K.R. Khrapko, V.V. Shick and A.D. Mirzabekov, Determination of the nucleotide sequence of DNA using hybridization with oligonucleotides. *A new method, Dokl. Acad. Sci. USSR*: (303) 1508 – 1511(1988).
- [5] R. Pendavingh, P. Schuurman and G.J. Woeginger, Recognizing DNA graphs is difficult. *Discrete Appl. Math.*, **127**: 85 – 94(2003).
- [6] P.A. Pevzner, l -Tuple DNA sequencing: computer analysis. *J. Biomol. Struct. Dyn.* (7): 63 – 73(1989).
- [7] E.M. Southern, U. Maskos and J.K. Elder, Analyzing and comparing nucleic acid sequences by hybridization to arrays of oligonucleotides: evaluation using experimental models. *Genomics*, (13): 1008 – 1017(1992).
- [8] S. Wang and J. Yuan, DNA Computing of directed line-graphs. *MATCH Commun Math Comput Chem*, **56**(3): 479 – 484(2006).
- [9] S. Wang, J. Yuan and S. Lin, DNA labelled graphs with DNA computing. *Science in China Series A: Mathematics*, **51**(3): 437 – 452(2008).