DISCRETE
APPLIED
MATHEMATICS

ELSEVIER Discrete Applied Mathematics 98 (1999) 1–19

On some properties of DNA graphs

J. Blazewicz[a], A. Hertz[b,*], D. Kobler[b], D. de Werra[b,1]

[a] *Instytut Informatyki, Politechnika Poznanka, Poznan, Poland*
[b] *Département de Mathématiques, EPFL, MA (Ecublens), CH-1015 Lausanne, Switzerland*

Received 10 July 1997; revised 13 November 1998; accepted 7 December 1998

Abstract

Molecular biology which aims to study DNA and protein structure and functions, has stimulated research in different scientific disciplines, discrete mathematics being one of them. One of the problems considered is that of recognition of DNA primary structure. It is known that some methods for solving this problem may be reduced (in their computational part) to graph-theoretic problems involving labeled graphs. Each vertex in such graphs has a label of length k written with an alphabet of size α, where k and α are two parameters. This paper is concerned with studying properties of these graphs (referred to as DNA graphs). More precisely, we give recognition algorithms and compare graphs labeled with different values of k and α. © 1999 Elsevier Science B.V. All rights reserved.

Keywords: DNA graphs; Protein structure and functions; Recognition algorithms

1. Introduction

It is widely believed that a discovery of a *Deoxyribonucleic acid* (DNA) structure by Watson and Crick [12] has reshaped a structure of modern biology. As a result molecular biology has emerged as a clearly defined research area. It appeared, however, that studying DNA structure and functions is impossible without help from other research disciplines. Because of the discrete nature of DNA (at least on its information carrying – genetic level), discrete mathematics appeared to be of special value for developing tools useful for solving particular problems of molecular biology.

One of the most challenging issues in the above area is to read (recognize) a structure of human genome, being a DNA chain composed of 3×10^9 pairs of nucleotides

---

[*] Corresponding author.
E-mail address: Alain.Hertz@epfl.ch (A. Hertz)

[1] The research of this author was supported by the CNR (Italy) while he was visiting the Department of Electronics, Informatics and Systematics at the University of Bologna in April 97. This support is gratefully acknowledged.

PII: S0166-218X(99)00109-2

(bases). It should be stressed that only four different types of nucleotides are distinguished. One of the methods used for reading DNA chains is *sequencing by hybridization* [2,4,5,8,10,12], where chains of several hundreds of nucleotides can be read with the current technology [9]. The method (as described later) 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. It appeared that one of the approaches used in the second phase may be based on graph theory [8,10]. More specifically, it uses labeled graphs where either vertices or arcs are labeled by particular DNA subchains. In such graphs either Hamiltonian or Eulerian paths corresponding to the DNA chains, are looked for. We will refer to these graphs as DNA *graphs*. The aim of this paper is to study the properties of DNA graphs which may be useful from the viewpoint of DNA chain recognition.

The organization of the paper is as follows. Section 2 describes some basic notions from molecular biology, relevant to the considered topic. Classes of graphs such as adjoints and directed line-graphs are defined and a particular labeling of directed graphs is considered. A characterization of directed line-graphs is given in Section 3. Results about the considered labelings are proved in Sections 4 and 5. A slight variation on the definition of the labelings is studied in Section 6. We conclude in Section 7 with some open problems.

2. Sequencing DNA chains and graph theory

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

As we mentioned, one of the methods of recognition of the primary structure of DNA (i.e. a sequence of nucleotides) is sequencing by hybridization. Its biochemical phase is based on the property of single-stranded acids to form a complex with a complementary strand of nucleic acid. All short fragments of nucleic acids (*oligonucleotides*) of length l (a library composed of 4^l subchains) are used in the hybridization experiment and thus, the formation of the complex indicates the occurrence of a sequence complementary to the oligonucleotide in the DNA chain. It is detected by a nuclear or spectroscopic detector. As a result of the experiment one gets a set (called *Spectrum*) of all l-long oligonucleotides which are known to hybridize with the investigated DNA sequence N of length n (i.e. they are substrings of string N). In case of ideal data (when no l-long oligonucleotide appears more than once in the sequence) we have thus $|Spectrum| = n - l + 1$ (We will not consider here experiments with errors).

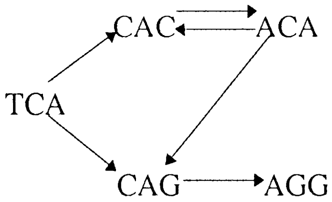

Fig. 1. The graph *H* for the example *TCACAGG*.

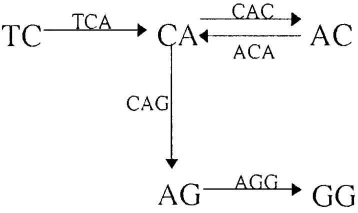

Fig. 2. The graph *G* for the example *TCACAGG*.

Now comes the computational phase, where for a given *Spectrum* one should reconstruct an unknown sequence *N*. The first approach to this problem based on graph theory, has been described by Lysov et al. [6]. They have proposed to formulate the problem of finding original sequence *N* as the problem of looking for a Hamiltonian path in a special graph. A directed graph *H* is built from *Spectrum* as follows: each oligonucleotide from *Spectrum* becomes a vertex, two vertices are connected by an arc if the $l-1$ rightmost nucleotides of the first vertex overlap with the $l-1$ leftmost nucleotides of the second one. A Hamiltonian path found in this graph corresponds to a proper sequence of elements of the *Spectrum*, i.e. a possible solution.

To illustrate this procedure let us consider the original sequence TCACAGG of length $n=7$. After the hybridization with oligonucleotides of length $l=3$ we get full *Spectrum* $\{\mathrm{TCA}, \mathrm{CAC}, \mathrm{CAG}, \mathrm{ACA}, \mathrm{AGG}\}$. Graph *H* constructed by this method is as shown in Fig. 1. The only Hamiltonian path in this graph is $TCA \rightarrow CAC \rightarrow ACA \rightarrow CAG \rightarrow AGG$ from which the original sequence can be read.

The above approach, however, leads to an exponential-time algorithm since looking for a Hamiltonian path is in general strongly NP-complete [6]. Fortunately, Pevzner has observed that in this particular case one can treat graph *H* as a directed line graph of a certain original graph *G*. Now, graph *H* can be transformed into graph *G*, each vertex of *H* corresponding to an arc of *G* (the set of arcs in the new graph corresponds, in fact, to the *Spectrum*). The arc connects vertices labeled as $l-1$ left and $l-1$ right nucleotides of the oligonucleotide corresponding to this arc. As a result of the transformation one gets the new graph in which a Eulerian path is looked for. This reduces the complexity of the algorithm solving the DNA sequencing problem since finding a Eulerian path can be done in polynomial time. Coming back to our example we have the graph *G* given in Fig. 2.

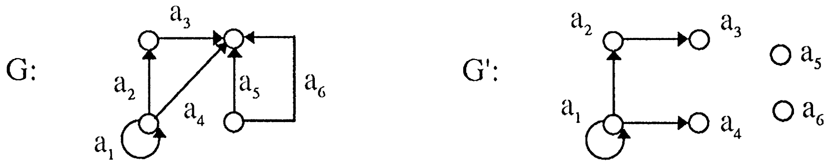

Fig. 3. A graph G and its adjoint G'.

The Eulerian path is $TC \rightarrow CA \rightarrow AC \rightarrow CA \rightarrow AG \rightarrow GG$ from which the same original sequence can be obtained.

The above approach raised some interesting questions in graph theory itself. They are concerned with the above class of labeled graphs which will be referred to as *DNA graphs* in the following. Specifically, one is interested in the characterization and recognition of these labeled graphs as well as in finding conditions for which the above transformation is possible. In the paper, these issues will be studied for unbounded and bounded alphabets used for graph labeling. Before doing this, we will set up the subject more formally in terms of graph theory. The definitions not given here can be found in [3]. Note that by *graph*, we mean *directed graph*. The following definitions will be used.

Definitions (*Berge* [3]). A graph is a p-graph if given any ordered pair x, y of vertices (x possibly equal to y), there are at most p parallel arcs from x to y.

The *adjoint* $G' = (V, U)$ of a graph $G = (X, V)$ is the 1-graph with vertex set V and such that there is an arc from a vertex x to a vertex y in G' if and only if the terminal endpoint of the arc x in G is the initial endpoint of arc y in G.

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

An example of a graph G and its adjoint G' is given in Fig. 3.

Notation. Let $G = (V, U)$ and $x \in V$, then $N^+(x) = \{y \in V \mid (x, y) \in U\}$ and $N^-(x) = \{y \in V \mid (y, x) \in U\}$.

Definition 1. A graph is a *directed line-graph* (or *line digraph* for some authors [7]) if and only if it is the adjoint of a 1-graph.

Definition 2. Let $k > 1$ and $\alpha > 0$ be two integers. We say that a 1-graph $H = (V, U)$ can be (α, k)-*labeled* if it is possible to assign a label $(l_1(x), \ldots, l_k(x))$ of length k to each vertex x of H such that

1. $l_i(x) \in \{1, \ldots, \alpha\} \quad \forall i \; \forall x \in V$;
2. all labels are different, that is $(l_1(x), \ldots, l_k(x)) \neq (l_1(y), \ldots, l_k(y))$ if $x \neq y$;
3. $(x, y) \in U \Leftrightarrow (l_2(x), \ldots, l_k(x)) = (l_1(y), \ldots, l_{k-1}(y))$.

Definition 3. Given two integers $k > 1$ and $\alpha > 0$, $\mathscr{L}_k^\alpha$ is the class of 1-graphs that can be (α, k)-labeled.

Notice that $\mathscr{L}_k^{\alpha} \subseteq \mathscr{L}_k^{\beta}\ \forall \beta > \alpha$ since $\{1,\ldots,\alpha\} \subset \{1,\ldots,\beta\}$. Given an integer $k > 1$, we denote by $\mathscr{L}_k^{\infty}$ the set of 1-graphs H for which there exists an integer $\alpha > 0$ such that H can be (α, k)-labeled. Since the graphs considered have a finite number n of vertices and each vertex gets k label components, a graph belonging to $\mathscr{L}_k^{\infty}$ also belongs to $\mathscr{L}_k^{nk}$. Thus for a graph in $\mathscr{L}_k^{\infty}$, we know that it can be (∞, k)-labeled with label components $l_i(x) \in \{1,\ldots,nk\}\ \forall i\ \forall x \in V$.

As DNA uses only four letters, we consider the special case where $\alpha = 4$. For this special case, all label components will be chosen in the set $\{A, C, G, T\}$ instead of $\{1, 2, 3, 4\}$.

Definition 4. A graph H is a DNA-*graph* if and only if $\exists k > 1$ such that $H \in \mathscr{L}_k^4$.

3. Characterization of directed line-graphs

In this section, it is shown that directed line-graphs can be recognized in polynomial time on the basis of a characterization involving forbidden partial subgraphs.

The following theorem explains why adjoints are interesting:

Theorem 1. *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.*

The proof is immediate from the definitions.

Since directed line-graphs are special cases of adjoints, we get the following corollary:

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

Since the problem of finding a Eulerian path/circuit in a graph (if any) can be solved in polynomial time, it follows that the problem of finding a Hamiltonian path/circuit in an adjoint is also polynomially solvable. It may therefore be interesting to know if adjoints can be recognized in polynomial time. The following theorem, whose proof can be found in [3], shows that the answer is positive:

Theorem 2 (Berge [3]). *A* 1*-graph* $H = (V, U)$ *is the adjoint of a graph if and only if the following holds for any pair* x, y *of vertices in* V:

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

The above theorem implies that if each vertex x of H is split into two new vertices x' and x'', and if each arc (x, y) is replaced by an arc (x', y''), then one gets a collection of vertex-disjoint complete bipartite graphs and isolated vertices. An example of this is given in Fig. 4.

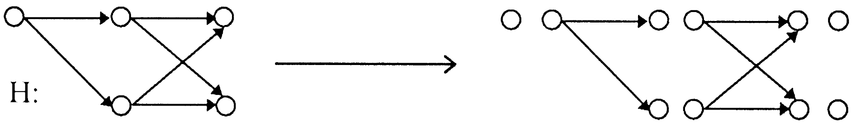

Fig. 4. An adjoint H and the result of its decomposition.

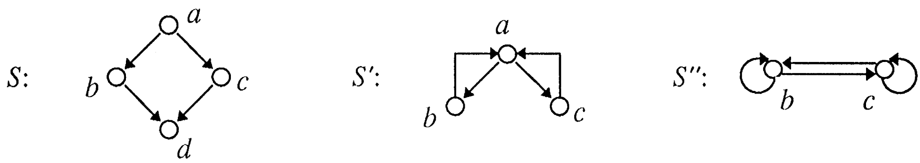

Fig. 5. The graphs S, S' and S''.

Notice that the above characterization is essentially the Monge condition for flow problems [1].

By definition, an adjoint is not necessarily a directed line-graph. As an example, one can easily check that the graphs S, S' and S'' of Fig. 5 are adjoints but not directed line-graphs:

The next theorem characterizes which adjoints are directed line-graphs.

Theorem 3. *An adjoint is a directed line-graph if and only if it contains neither of the graphs S, S' or S'' of Fig.* 5 *as partial subgraph.*

Proof. ($\Rightarrow$) Assume H is the directed line-graph of a 1-graph G. Suppose that H contains S or S' as a partial subgraph. Then the arcs b and c of G, corresponding to the vertices b and c of S or S', must have the same initial endpoint since the arcs (a,b) and (a,c) belong to S or S'; this common initial endpoint is the terminal endpoint of the arc a of G. But as the arcs (b,d) and (c,d) belong to S and arcs (b,a) and (c,a) belong to S', the arcs b and c of G must also have the same terminal endpoint. This is in contradiction with the fact that G is a 1-graph. Similarly, suppose that H contains S'' as a (partial) subgraph. Since there is a loop on vertices b and c of H, the arcs b and c of G must be loops themselves. These loops must be on the same vertex of G in order to have S'' in H, and G is no longer a 1-graph.

($\Leftarrow$) Let H be the adjoint of a graph G and assume that H contains neither S nor S' nor S'' as partial subgraph. If G is a 1-graph, the proof is completed. Hence, assuming that G is not a 1-graph, 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 S'' is not a subgraph of H, these two vertices x and y are distinct. Moreover, we have $N^-(x)=\emptyset$ or $N^+(y)=\emptyset$. Indeed, if this is not the case, then there is in G' an arc a entering x and an arc d leaving y. Let b and c be two parallel arcs linking x to y. If $a \neq d$, the arcs a, b, c and d form the partial subgraph S in the adjoint H of G', a contradiction. If $a = d$, the arcs a, b and c form the partial subgraph S' in H, also a contradiction.

Therefore, we have $N^-(x)=\emptyset$ or $N^+(y)=\emptyset$ and we can apply the following changes to G', where $e_1,\dots,e_p$ $(p > 1)$ are the parallel arcs from x to y:

```
if $N^-(x) = \emptyset$ 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 $(*N^+(y) = \emptyset*)$
    replace $y$ by $y_1,\dots,y_p$ and each arc $e_i$ by an arc $(x, y_i)$ $i = 1,\dots,p$;
    replace an 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. □

Corollary 2. *A* 1-*graph is a directed line-graph if and only if the following holds for any pair* x, y *of vertices*:

$$N^+(x) \cap N^+(y) \neq \emptyset \;\Rightarrow\; (N^+(x) = N^+(y) \text{ and } N^-(x) \cap N^-(y) = \emptyset).$$

Proof. ($\Rightarrow$) Since the graph is a directed line-graph, it is also an adjoint and therefore, by Theorem 2, $N^+(x) \cap N^+(y) \neq \emptyset$ already implies $N^+(x)=N^+(y)$. It is easy to check that if for a pair of vertices x, y we have $N^+(x)=N^+(y) \neq \emptyset$ and $N^-(x) \cap N^-(y) \neq \emptyset$, then the graph must contain S, S' or S'' of Fig. 5 as partial subgraph, contradicting Theorem 3.

($\Leftarrow$) By Theorem 2, we know that the graph must be an adjoint. Moreover, since in all three graphs S, S' and S'' there is a pair of vertices b and c such that $N^+(b) \cap N^+(c) \neq \emptyset$ and $N^-(b) \cap N^-(c) \neq \emptyset$, the given graph cannot have S, S' or S'' as partial subgraph and is therefore a directed line-graph. □

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

4. Classes $\mathscr{L}_k^\infty$

In the next sections we will only consider 1-graphs. In order to simplify the reading, we will use the term ‘graph’ to mean ‘1-graph’ when no confusion occurs.

In this section, we shall consider classes $\mathscr{L}_k^\infty$, that is without any upper bound on the values (size of the alphabet) of the label components.

Theorem 4. *Let G be a graph belonging to* $\mathscr{L}_k^\infty$ *with* $k>1$, *and let H be its directed line-graph. Then H belongs to* $\mathscr{L}_{k+1}^\infty$.

Proof. Consider any (∞,k)-labeling of G, and any arc $(x_i,\ x_j)$ in G. Let $(l_1(x_i),l_2(x_i),\ldots,l_k(x_i))$ and $(l_1(x_j),\ldots,l_{k-1}(x_j),l_k(x_j))$ be the labels assigned to the vertices x_i and x_j. We assign the label $(l_1(x_i),l_2(x_i),\ldots,l_k(x_i),l_k(x_j))$ to vertex $v=(x_i,\ x_j)$ in H. We prove now that this is a $(\infty,k+1)$-labeling of H. First notice that each label has length $k+1$. Moreover, since G is a graph belonging to $\mathscr{L}_k^\infty$, it follows that all labels in H are different. It remains to prove that (v_a,v_b) is an arc in H if and only if the k last label components of v_a are equal to the k first label components of v_b.

Let $v_a=(x_p,\ x_q)$ and $v_b=(x_r,\ x_s)$ be two vertices of H. Since x_q is a successor of x_p in G, we know that $(l_2(x_p),\ldots,l_k(x_p))=(l_1(x_q),\ldots,l_{k-1}(x_q))$. We now have the following equivalences:

(v_a,v_b) is an arc of H

$\Leftrightarrow$ vertices x_q and x_r are the same

$\Leftrightarrow$ the label $(l_1(x_r),l_2(x_r),\ldots,l_k(x_r),l_k(x_s))$ of v_b is equal to $(l_1(x_q),l_2(x_q),\ldots,l_{k-1}(x_q),l_k(x_q),l_k(x_s))$, which is equal to $(l_2(x_p),\ldots,l_k(x_p),l_k(x_q),l_k(x_s))$ according to the above remark

$\Leftrightarrow$ the k last label components of the label $(l_1(x_p),l_2(x_p),\ldots,l_k(x_p),l_k(x_q))$ of v_a are the same as the k first label components of v_b. □

Theorem 5. *A graph is a directed line-graph if and only if it belongs to* $\mathscr{L}_2^\infty$.

Proof. ($\Rightarrow$) Let H be a directed line-graph of a graph G. Each vertex v corresponds to an arc $(x_i,\ x_j)$ of G. It is easy to verify that, by assigning label (i,j) to vertex $v=(x_i,\ x_j)$, one gets a $(\infty,2)$-labeling of H, where all labels are different since G does not contain parallel arcs.

($\Leftarrow$) Consider a $(\infty,2)$-labeling of a graph $H\in\mathscr{L}_2^\infty$. Without loss of generality, we may assume that all label components belong to the set $A=\{1,\ldots,\alpha\}$ where $\alpha\leqslant 2n$. We now construct a graph $G=(A,V)$ in the following way: there is an arc from a vertex i to a vertex j in G if and only if there is a vertex with label $(i,\ j)$ in H. 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 6. *Let k be an integer* >2. *Then* $\mathscr{L}_k^\infty\subset\mathscr{L}_d^\infty$ *for* $d=2,\ldots,k-1$.

Proof. It is enough to prove that $\mathscr{L}_k^\infty\subset\mathscr{L}_{k-1}^\infty$ for $k>2$.

Let H be a graph in $\mathscr{L}_k^\infty$ and consider any $(\infty,\ k)$-labeling of H. Without loss of generality, we may assume that all label components belong to $A=\{1,\ldots,\alpha\}$, with $\alpha\leqslant nk$. Let ϕ be an isomorphism from $A\times A$ to $A'=\{1,\ldots,\alpha^2\}$. A $(\infty,\ k-1)$-labeling of H can be constructed in the following way, where all label components are chosen in the set A': we transform each label $(l_1(v),\ldots,l_k(v))$ of a vertex

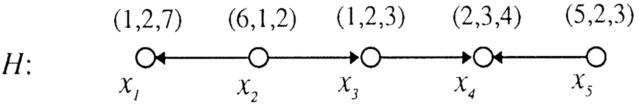

Fig. 6. The graph H is a counterexample of the converse of Theorem 4.

v into the label $(m_1(v),\dots,m_{k-1}(v))$ where $m_i(x)=\phi(l_i(v),l_{i+1}(v))$. Hence H belongs to $\mathscr{L}^{\infty}_{k-1}$.

Up to this point, we have proved that $\mathscr{L}^{\infty}_{k}\subseteq\mathscr{L}^{\infty}_{k-1}$. We now show that this inclusion is strict. Let H be a graph with vertex set $\{a,\ x_1,\ x_2,\dots,\ x_{k-1},\ y_1,\ y_2,\dots,y_{k-1},b\}$ and made of two paths $(a,\ x_1,\ x_2,\dots,x_{k-1},b)$ and $(a,y_1,y_2,\dots,y_{k-1},b)$ linking a to b. This graph can be $(\infty,k-1)$-labeled by assigning the following labels to the vertices of H:

$$l(a)=(1,2,\dots,k-1),$$

$$l(x_i)=(i+1,i+2,\dots,k-1,2k-1,k,k+1,\dots,k+i-2),$$

$$l(y_i)=(i+1,i+2,\dots,k-1,2k,k,k+1,\dots,k+i-2),$$

$$l(b)=(k,k+1,\dots,2k-2).$$

However, $H\notin\mathscr{L}^{\infty}_{k}$. Indeed, assume H can be $(\infty,\ k)$-labeled. Since the $k-1$ first label components of a vertex x_i are equal to the $k-1$ last label components of its predecessor, it follows that the $k-i$ first label components of a vertex x_i are equal to the $k-i$ last label components of vertex a. Also, the i last label components of x_i are equal to the i first label components of vertex b. The same applies to vertex y_i. Hence vertices x_i and y_i have the same label, which implies that there should be an arc from x_i to y_{i+1} for $1\leqslant i\leqslant k-2$, a contradiction. □

Notice that the converse of Theorem 4 is not true. Indeed, a $(7,3)$-labeling of a graph H is represented in Fig. 6, hence $H\in\mathscr{L}^{\infty}_{3}$. By Theorem 6, we know that H also belongs to $\mathscr{L}^{\infty}_{2}$, which means, by Theorem 5, that H is a directed line-graph of a graph G. Let a be the terminal endpoint of arc x_2 in G, and let b be the initial endpoint of arc x_5 in G. Vertices a and b have a common successor, which is the terminal endpoint of both arcs x_3 and x_5. However, the terminal endpoint of arc x_1 is a successor of a, but not of b. By Theorem 2, G is not an adjoint, hence not a directed line-graph. It follows from Theorem 5 that $G\notin\mathscr{L}^{\infty}_{2}$. In summary, H is the directed line-graph of G with $H\in\mathscr{L}^{\infty}_{3}$ and $G\notin\mathscr{L}^{\infty}_{2}$.

Given a graph H and an integer $k>1$, we now give an algorithm, called PROPAGATION ALGORITHM, that determines whether or not H belongs to $\mathscr{L}^{\infty}_{k}$. If $H\in\mathscr{L}^{\infty}_{k}$, then the algorithm produces a (∞,k)-labeling of H.

PROPAGATION ALGORITHM:

```
1. set $l_i(v) = 0$ for each vertex $v$ in $H$ and for all $i = 1, \ldots, k$; 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 $L \neq \emptyset$ do
           choose any pair $(v, i)$ in $L$, set $l_i(v) := \alpha$ and update $L$;
3. if two vertices have the same label then STOP: $H \notin \mathscr{L}_k^{\infty}$;
4. if no arc is linking vertex $v$ to vertex $w$ in $H$ while $(l_2(v), \ldots, l_k(v)) = (l_1(w), \ldots,
   l_{k-1}(w))$ then STOP: $H \notin \mathscr{L}_k^{\infty}$;
5. STOP: a $(\infty, k)$-labeling of $H$ has been determined.
```

Theorem 7. PROPAGATION ALGORITHM works correctly and has $O(n^2k)$ complexity.

Proof. At each iteration of step (2), a number α is propagated along the arcs in order to satisfy the condition: (v, w) is an arc of $H \Rightarrow (l_2(v), \ldots, l_k(v)) = (l_1(w), \ldots, l_{k-1}(w))$. Thus the assignments done in (2) are all necessary. If two vertices get the same label, then the algorithm has to be stopped since this is not allowed. It remains to check that $(l_2(v), \ldots, l_k(v)) = (l_1(w), \ldots, l_{k-1}(w)) \Rightarrow (v, w)$ is an arc of H. If this is not the case, the algorithm stops at step (4). This in particular occurs when H is not an adjoint.

Step (2) can be performed in $O((n + m)k)$. Indeed, consider the undirected graph H' obtained from H as follows: for each vertex v of H, put k vertices $v_1, \ldots, v_k$ in H'; for each arc (v, w) of H, put $k - 1$ edges $(v_2, w_1), \ldots, (v_k, w_{k-1})$ in H'. Performing step (2) is equivalent to associating a different label to every connected component of H'. This can be done using Tarjan's algorithm [11]. Since steps (3) and (4) have $O(n^2k)$ complexity, it follows that the overall complexity of PROPAGATION ALGORITHM is $O(n^2k)$. □

While Theorem 5 proves that graphs belonging to $\mathscr{L}_2^{\infty}$ can be recognized in polynomial time, Theorem 7 proves that recognition of graphs in $\mathscr{L}_k^{\infty}$ can be performed in $O(n^2k)$ time. The proof of Theorem 5 (the 'if' part), combined with the PROPAGATION ALGORITHM, shows that, given a directed line-graph H, it is easy to find a graph G such that H is the directed line-graph of G.

5. Classes $\mathscr{L}_k^{\alpha}$

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 $\{1,2,3,4\} \equiv \{A,C,G,T\}$. Notice first that by definition of $\mathscr{L}_k^\alpha$, we have $\mathscr{L}_k^\alpha \subseteq \mathscr{L}_k^{\alpha'}$ for all $\alpha' \geqslant \alpha$. It follows from Theorem 6 that $\mathscr{L}_k^\alpha \subset \mathscr{L}_2^\infty$ for any $k > 2$ and $\alpha > 0$. Moreover, as already mentioned in Section 2, if a graph H with n vertices belongs to $\mathscr{L}_k^\infty$, then it also belongs to $\mathscr{L}_k^{nk}$. In fact, this last property can be improved as stated in Theorem 8.

Theorem 8. *If* $H \in \mathscr{L}_k^\infty$ *then* $H \in \mathscr{L}_k^{n+p(k-1)}$ *where* n *is the number of vertices and* p *the number of connected components of the underlying undirected graph.*

Proof. Assume that $H = (V,U) \in \mathscr{L}_k^\infty$. We will consider a sequence $H_0,\dots,H_{|U|}$ of graphs, where $H_0 = (V,\emptyset)$ and H_i is obtained from H_{i-1} by adding an arc of H not already in H_{i-1}. Notice that $H_{|U|} = H$. These arcs are added in a particular order. If there is an arc $(v,w) \in U$ such that exactly one of its endpoints is isolated in H_{i-1}, than such an arc is added to H_{i-1} for getting H_i. If this is not the case, then we look for an arc $(v,w) \in U$ such that both endpoints are isolated in H_{i-1}. If such an arc exists, we add it to H_{i-1} for getting H_i. Otherwise, any arc is chosen for generating H_i.

Since H contains p connected components, it follows that H_{n-p} is a maximal forest in H. We now prove by induction that the number of distinct values in any (∞,k)-labeling of H_i is at most equal to $nk-(k-1)i$ and that this bound is sharp for $i \leqslant n-p$. The result is clear for $i=0$ since it is sufficient to give distinct values to all label components. Suppose now that the result is true for $i-1 < n-p$, and let $(v,w) \in U$ be the arc added to H_{i-1} for getting H_i. Since at least one of its endpoints is isolated in H_{i-1}, we consider two cases. If v is isolated, the $k-1$ last label components of v can no longer be different from the $k-1$ first label components of w. Otherwise, w is isolated and the $k-1$ first label components of w can no longer be different from the $k-1$ last label components of v. Hence the upper bound has to be decreased by exactly $k-1$ units and remains sharp.

This proves that the number of distinct values in any (∞, k)-labeling of H_{n-p} is at most equal to $nk-(k-1)(n-p) = n+p(k-1)$. Since adding arcs to H_{n-p} for getting $H_{|U|} = H$ can only reduce this bound, we have $H \in \mathscr{L}_k^{n+p(k-1)}$. $\square$

Notice that the end of the proof of Theorem 8 shows that in any (α,k)-labeling of a graph H with n vertices and p connected components, at least $\alpha-(n+p(k-1))$ available values are not used.

It is proved in Theorem 6 that $\mathscr{L}_k^\infty \subset \mathscr{L}_{k-1}^\infty$. However, this is no longer the case when the maximum value of the label components is fixed to an integer $\alpha > 0$. For example, it is easy to check that the circuit C_3 on three vertices belongs to $\mathscr{L}_3^2$ but not to $\mathscr{L}_2^2$. The following observation shows that the situation is even worse. Consider a graph H and assume that there are three integers k_1,k_2,α such that $k_2 > k_1+1$, $\alpha > 0$, $H \in \mathscr{L}_{k_1}^\alpha$ and $H \in \mathscr{L}_{k_2}^\alpha$. It may happen that $H \notin \mathscr{L}_k^\alpha$ with $k_1 < k < k_2$. An example is given in Fig. 7 where a (2,4)-labeling and a (2,6)-labeling of a graph H are represented. For proving that this graph does not belong to $\mathscr{L}_5^2$, we first propagate value 1 starting

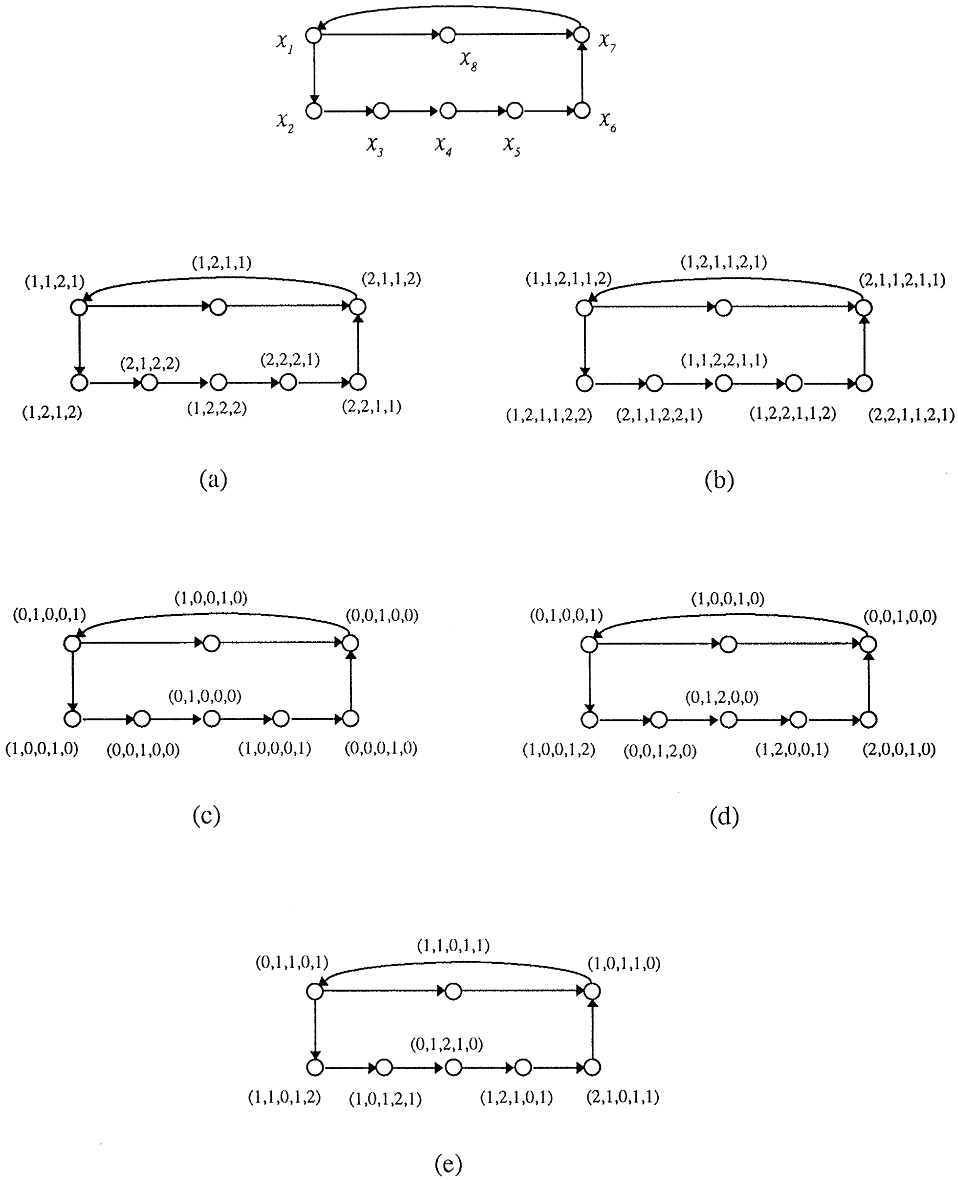

Fig. 7. A graph in $\mathscr{L}_4^2$ and $\mathscr{L}_6^2$, but not in $\mathscr{L}_5^2$.

from the last component of vertex x_1. The result of this propagation is given in Fig. 7(c). Since x_6 and x_8 have a common successor and must receive different labels, it follows that their first label components must be different, which means that the first label component of x_6 must be equal to 2. By propagating this value, we get the partial labeling represented in Fig. 7(d). Again, since x_2 and x_8 have a common predecessor, this means that their last label components must be different. Hence the last label component of x_8 must be equal to 1. The result of its propagation is given in Fig. 7(e). Finally, since there is no loop on x_8, its middle label component must be equal to 2 and we get a labeling with the same label on x_3 and x_5 which is of course not allowed.

A question that naturally arises is the following one: knowing that a graph H is in $\mathscr{L}_k^{\infty}$, which is the smallest integer α such that H is in $\mathscr{L}_k^{\alpha}$? This number will be denoted $\alpha_k(H)$. It has been shown in the proof of Theorem 6 that $\alpha_{k-1}(H) \leqslant \alpha_k^2(H)$. Hence we get the following property:

Property. If $H \in \mathscr{L}_k^{\infty}$, then $H \notin \mathscr{L}_k^{\alpha}\,\forall \alpha < \lceil \sqrt{\alpha_{k-1}(H)}\,\rceil$.

As already proved, $\alpha_k(H)$ exists and is such that $1 \leqslant \alpha_k(H) \leqslant n + p(k-1)$ where n is the number of vertices in H and p is its number of connected components. The PROPAGATION ALGORITHM described in Section 4 delivers an integer β such that $H \in \mathscr{L}_k^{\beta}$. However β can be strictly larger than $\alpha_k(H)$. For example, it is easy to check that $\beta = 4$ and $\alpha_k(H) = 2$ if H is the circuit on four vertices and $k = 3$.

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

Theorem 9. *Let $H \in \mathscr{L}_2^{\infty}$, and consider the induced subgraph H_1 obtained from H by removing all isolated vertices without a loop. The problem of determining $\alpha_2(H_1)$ can be solved in polynomial time.*

Proof. Let us first apply PROPAGATION ALGORITHM to H_1 in order to determine an upper bound β for $\alpha_2(H_1)$. The numbers used in this $(\beta, 2)$-labeling l can be partitioned into three sets:

$S=$ the numbers that only appears as first component of the labels;
$T=$ the numbers that only appear as last component of the labels;
$I = \{1, \ldots, \beta\} \setminus (S \cup T)$.

Notice that the value of the first label component of a vertex v belongs to S if and only if v is a source in H_1. Also, the value of the last label component of a vertex v belongs to T if and only if v is a sink. Moreover, an integer i belongs to I if and only if there exists an arc (v, w) where $l_2(v) = l_1(w) = i$.

The PROPAGATION ALGORITHM is not necessarily optimal for the following reason. Each time the main loop of step (2) is performed, a new integer is considered. However the same integer, or even a smaller one, could perhaps also be used. This means that some integers in the set $\{1, \ldots, \beta\}$ can be replaced by others, without losing the fact that we have a labeling. The possible replacements are defined in the following claim.

Claim. *Consider a $(\beta, 2)$-labeling of H_1 and let i and j be two distinct integers in $\{1, \ldots, \beta\}$. If the labeling obtained by replacing all occurrences of i by j is also a $(\beta, 2)$-labeling, then either $\{i, j\} \subseteq S$ or else $\{i, j\} \subseteq T$.*

Proof. Assume $i \in S$ and $j \notin S$. As described above, there exists a source v such that $l_1(v) = i$ and a vertex w (possibly equal to v) such that $l_2(w) = j$. If i is replaced by j, then there must exist an arc (w, v). Hence v is not a source, a contradiction.

Similarly, if $i \in T$ and $j \notin T$ then there exists a sink v such that $l_2(v) = i$ and a vertex w (possibly equal to v) such that $l_1(w) = j$. If i is replaced by j, then there must exist an arc (v, w). Hence v is not a sink, a contradiction.

Finally, if both i and j belong to I, there must be two vertices v and w (not necessarily distinct) with $l_1(v) = i$ and $l_2(w) = j$. If i is replaced by j, then there must exist an arc (w, v). But if such an arc exists then PROPAGATION ALGORITHM would have given the same value $\min\{i, j\}$ to both $l_1(v)$ and $l_2(w)$. Since i and j are supposed to be distinct, we get a contradiction.

Hence the only possible cases are $\{i, j\} \subseteq S$ and $\{i, j\} \subseteq T$, and this concludes the proof of the claim. □

Let H_1^s be the subgraph of H_1 induced by the sources and their successors. It follows from Theorem 2 that H_1^s is the union of node-disjoint complete bipartite graphs. Sources having a common successor must have different values for their first label component (since otherwise two vertices in H_1 would have the same label). However, two sources belonging to two different connected component of H_1^s can have the same value for their first label component. Let V_s be the largest set of sources having a common successor and let D_1 be the set of values of the first label components of the vertices in V_s. We now modify the $(\beta, 2)$-labeling in the following way. Given any connected component C of H_1^s, we change the value of the first label components of the sources in C by using different values chosen in D_1.

Similarly, let V_T be the largest set of sinks having a common predecessor and let D_2 be the set of values of the last label components of the vertices in V_T. The value of the last label components of the sinks are chosen in D_2 in such a way that no two sinks having a common predecessor receive the same value.

It follows from the above observation that $|D_1|$ is the minimum number of different values needed for the first label components of the sources in any labeling of H_1. In other words, one can use $|D_1|$ instead of $|S|$ different values for the first label components of the sources. Similarly, $|D_2|$ is the minimum number of different values needed for the last label components of the sinks in any labeling of H_1.

Notice that the values used for the first label components of the sources must be different from the values used for the last label components of the sinks. Moreover, according to the claim, none of the $|D_1| + |D_2|$ values can be replaced by a value in I. Also, since H_1 does not contain any isolated vertex without a loop, at most one component of a label of a vertex has been modified when reducing the number of different values in the $(\beta, 2)$-labeling. Therefore all labels of the new labeling are different.

It follows that $\alpha_2(H_1) = |I| + |D_1| + |D_2|$. This number can be computed in polynomial time and this concludes the proof. □

In the previous theorem, isolated vertices without a loop have been ignored. We now also take these vertices into account.

Theorem 10. *Let $H \in \mathscr{L}_2^\infty$. The problem of determining $\alpha_2(H)$ can be solved in polynomial time.*

Proof. Let $H = (V, U)$ be a graph in $\mathscr{L}_2^\infty$ and let V' be the subset of vertices in H that are not the endpoint of at least one arc. Let H_1 be the subgraph of H induced by $V \backslash V'$. Consider a $(\alpha_2(H_1), 2)$-labeling of H_1. Let S, T and I be the same sets as those used in the proof of Theorem 9. Notice that $|S|$ is equal to the maximum number of sources in H having a common successor. Also $|T|$ is equal to the maximum number of sinks in H having a common predecessor.

The value of the first label component of a vertex in V' cannot belong either to I or to T. Indeed, if this was the case, there would be an arc in H entering this vertex. Similarly, the value of the last label component of a vertex in V' cannot belong either to I or to S.

If the label of a vertex in V' belongs to $S \times T$, then this label is different from the label of any vertex in $V \backslash V'$. This follows from the fact that no vertex in H_1 is both a source and a sink.

Consider now a $(\alpha, 2)$-labeling of H and let n_1 (resp. n_2) denote the number of values that only appear as first (resp. last) component of a label. Hence, we have $\alpha = |I| + n_1 + n_2$. Since $|I|$ is fixed, the smallest α is obtained by minimizing $n_1 + n_2$. Notice that $n_1 \geqslant |S|$ and $n_2 \geqslant |T|$. Also, $n_1 * n_2 \geqslant |V'|$ since every vertex in V' must receive a different label. It is not difficult to check that the minimum value of α (that is $\alpha_2(H)$) is equal to $n_1 + n_2$, where n_1 and n_2 are computed as follows:

$$\text{if } |S| \geqslant |T| \text{ then } n_1 = \max\{\lceil\sqrt{|V'|}\rceil, |S|\} \text{ and } n_2 = \max\left\{\left\lceil\frac{|V'|}{n_1}\right\rceil, |T|\right\}$$

$$\text{else } n_2 = \max\{\lceil\sqrt{|V'|}\rceil, |T|\} \text{ and } n_1 = \max\left\{\left\lceil\frac{|V'|}{n_2}\right\rceil, |S|\right\},$$

It follows from Theorem 9 that $|I|, |S|$ and $|T|$ can be computed in polynomial time. Hence this is also the case for n_1, n_2 and $\alpha_2(H)$, and this concludes the proof. □

It may be useful to summarize the procedure which, given a graph $H \in \mathscr{L}_2^\infty$, determines $\alpha_2(H)$ in polynomial time:

(1) apply PROPAGATION ALGORITHM on H_1 (see Theorem 9) to determine I;

(2) set $|S|$ equal to the maximum number of sources in H_1 having a common successor, and $|T|$ equal to the maximum number of sinks in H_1 having a common predecessor;

(3) compute n_1 and n_2 as described in Theorem 10;

(4) set $\alpha_2(H) = |I| + n_1 + n_2$.

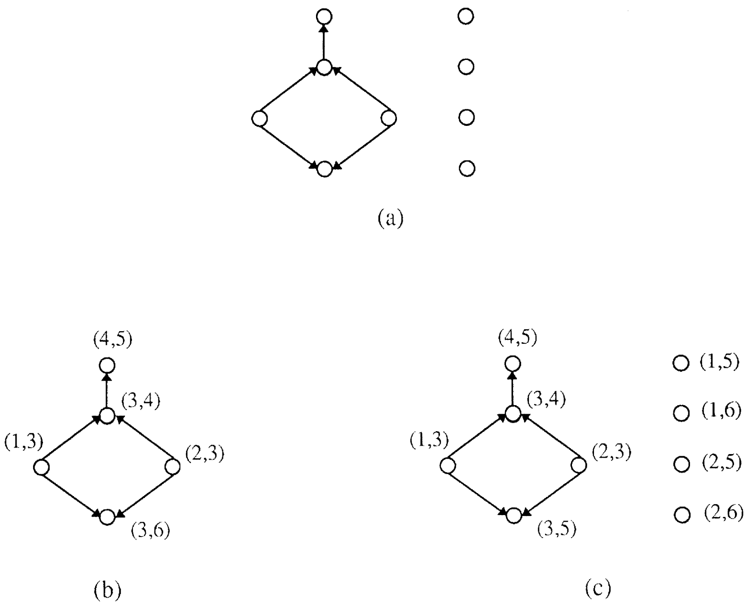

Fig. 8. An illustration of the determination of $\alpha_2(H)$.

This procedure is illustrated in Fig. 8 the graph H is described in (a), and the $(6,2)$-labeling of H_1 obtained by means of PROPAGATION ALGORITHM is represented in (b). Since $|I|=|S|=2$ and $|T|=1$, we have $n_1=n_2=2$ and $\alpha_2(H)=6$. A $(6,2)$-labeling of H is illustrated in (c). Notice that in the optimal labeling of H given here, both sinks have the same value for their last label component.

6. A relaxation

In Definition 2 it is imposed that all vertices must get different labels. This constraint was motivated by a biological background, as explained in Section 2. We now present results where this constraint is relaxed. Proofs are not given since they are similar to those given in the previous sections.

Definition 2′. Let $k > 1$ and $\alpha > 0$ be two integers. We say that a 1-graph $H=(V,U)$ can be (α,k)-*free-labeled* if it is possible to assign a label $(l_1(x),\ldots,l_k(x))$ of length k to each vertex x of H such that

1. $l_i(x) \in \{1,\ldots,\alpha\}\ \forall i\ \forall x \in V$;
2. $(x,y) \in U \Leftrightarrow (l_2(x),\ldots,l_k(x)) = (l_1(y),\ldots,l_{k-1}(y))$.

Definition 3′. Given two integers $K > 1$ and $\alpha > 0$, $\widetilde{\mathscr{L}}_k^{\alpha}$ is the class of 1-graph that can be (α,k)-free-labeled.

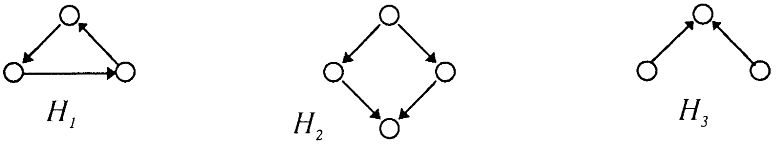

Fig. 9. Graphs H_1, H_2 and H_3.

Obviously, $\mathscr{L}_k^\alpha \subset \widetilde{\mathscr{L}}_k^\alpha$ for all $k > 1$ and $\alpha > 0$. Notice that Theorem 4 is no longer valid. Indeed, consider the graph S of Fig. 5 in which we add an arc from d to a. It is not difficult to check that, given any integer $k > 1$, this graph belongs to $\widetilde{\mathscr{L}}_k^\infty$, while its directed line-graph does not belong to $\widetilde{\mathscr{L}}_{k+1}^\infty$. The statement of Theorems 5 and 6 are modified as follows:

Theorem 5′. *A graph is an adjoint if and only if it belongs to* $\widetilde{\mathscr{L}}_2^\infty$.

Theorem 6′. *Let k be an integer* > 2. *Then* $\widetilde{\mathscr{L}}_k^\infty \subset \widetilde{\mathscr{L}}_d^\infty$ *for* $d = 2, \ldots, k-1$.

If we remove step (3) in PROPAGATION ALGORITHM Theorem 7 remains valid. Moreover, the graph in Fig. 7 is still an example of a graph that can be $(\infty, 2)$-free-labeled but not $(\infty, 3)$-free-labeled.

Given a graph H in $\widetilde{\mathscr{L}}_k^\infty$, let $\tilde{\alpha}_k(H)$ denote the smallest integer such that H belongs to $\widetilde{\mathscr{L}}_k^{\tilde{\alpha}_k(H)}$. If we replace $\alpha_k(H)$ by $\tilde{\alpha}_k(H)$ and $\mathscr{L}_i^\alpha$ by $\widetilde{\mathscr{L}}_i^\alpha$, all results of Section 5 are still valid. The proof of Theorems 9 and 10 is however a little easier, since the sets S and T used in the proofs can be reduced to singletons. More precisely, $|D_1| = |D_2| = n_1 = n_2 = 1$ and a graph in $\widetilde{\mathscr{L}}_2^\infty$ needs only $|I| + 2$ numbers to be free-labeled.

Let $\widetilde{\mathscr{L}}_k^\infty$ be the set of graphs for which there exists a $(\infty,\ k)$-free-labeling where at least two vertices have equal labels. It follows that $\widetilde{\mathscr{L}}_k^\infty = \hat{\mathscr{L}}_k^\infty \cup \mathscr{L}_k^\infty$. Notice that $\hat{\mathscr{L}}_k^\infty \cap \mathscr{L}_k^\infty \neq \emptyset$, $\hat{\mathscr{L}}_k^\infty \nsubseteq \hat{\mathscr{L}}_k^\infty$ and $\mathscr{L}_k^\infty \nsubseteq \hat{\mathscr{L}}_k^\infty$. Indeed, graphs H_1, H_2 and H_3 in Fig. 9 illustrate these three properties where:

- $H_1 \in \mathscr{L}_k^\infty$ and $H_1 \notin \hat{\mathscr{L}}_k^\infty$ $\forall k$,
- $H_2 \in \hat{\mathscr{L}}_k^\infty$ and $H_2 \notin \mathscr{L}_k^\infty$ $\forall k$,
- $H_3 \in \mathscr{L}_k^\infty$ and $H_3 \in \hat{\mathscr{L}}_k^\infty$ $\forall k$.

7. Open questions

We conclude with open questions.

Open question 1. Given a graph $H \in \mathscr{L}_2^\infty$, determine the largest integer L such that $H \in \mathscr{L}_L^\infty$.

This number L is not necessarily finite. For example, a circuit can be $(\infty,\ k)$-labeled for any integer $k > 1$. In such a case, we say that H belongs to $\mathscr{L}_\infty^\infty$. We conjecture that there exists a threshold value $L(H)$ for which we have $H \in \mathscr{L}^{\infty}_{L(H)}$ if and only if $H \in \mathscr{L}^{\infty}_{\infty}$.

Notice that it follows from Theorems 5 and 10 that graphs in $\mathscr{L}^{\alpha}_{2}$ can be recognized in polynomial time for any $\alpha > 0$. We have no such result when exchanging the roles of k and α.

Open question 2. Given an integer $k > 1$ and a graph H, is it possible to recognize whether H belongs to $\mathscr{L}^{2}_{k}$ in polynomial time (where k is not considered as a constant)?

This would allow to determine whether $\alpha_k(H)$ is equal to 2 or not. The question of the determination of $\alpha_k(H)$ is more difficult.

Open question 3. Given an integers $k > 1$ and a graph H, is it possible to determine $\alpha_k(H)$ in polynomial time (where k is not considered as a constant)?

Open question 4. Given two integers $k > 2$ and $\alpha > 1$, and a graph H, is it possible to recognize whether H belongs to $\mathscr{L}^{\alpha}_{k}$ in polynomial time (where k is not considered as a constant)?

In particular, if $\alpha = 4$ answering this question would allow recognition of DNA graphs. Finally, if we know that a graph H belongs to some class $\mathscr{L}^{\alpha}_{k}$, it could be interesting to determine if this also the case for different lengths of the labels.

Open question 5. Given two integers $k > 1$ and $\alpha > 1$, and a graph $H \in \mathscr{L}^{\alpha}_{k}$, determine all integers k' such that $H \in \mathscr{L}^{\alpha}_{k'}$.

Acknowledgements

The authors thank Marta Kasprzak and an anonymous referee for their useful comments that helped to improve this paper.

References

[1] I. Adler, A.J. Hoffman, R. Shamir, Monge and feasibility sequences in general flow problems, Discrete Appl. Math. 44 (1993) 21–38.

[2] W. Bains, G.C. Smith, A novel method for nucleic acid sequence determination, J. Theoret. Biol. 135 (1988) 303–307.

[3] C. Berge, Graphes, Dunod, Paris, 1970.

[4] J. Blazewicz, J. Kaczmarek, M. Kasprzak, W.T. Markiewicz, J. Weglarz, Sequential and parallel algorithms for DNA sequencing, CABIOS 13 (2) (1997) 151–158.

[5] R. Drmanac, I. Labat, I. Brukner, R. Crkvenjakov, Sequencing of megabase plus DNA by hybridization: theory of the method, Genomics 4 (1989) 114–128.

[6] Yu.P. Lysov, V.L. Florentiev, A.A. Khorlyn, K.R. Khrapko, V.V. Shick, A.D. Mirzabekov, Determination of the nucleotide sequence of DNA using hybridization with oligonucleotides. A new method, Dokl. Acad. Sci. USSR 303 (1988) 1508–1511.

[7] O. Melnikov, R. Tyshkevich, V. Yemelichev, V. Sarvanov, Lectures on Graph Theory, BI Wissenschftsverlag, 1994.

[8] P.A. Pevzner, l-Tuple DNA sequencing: computer analysis, J. Biomol. Struct. Dyn. 7 (1989) 63–73.

[9] P.A. Pevzner, R.J. Lipshutz, Towards DNA sequencing chips, Symposium on Mathematical Foundations of Computer Science, Lecture Notes in Computer Science, vol. 841, Springer, Berlin, 1994, pp. 143–158.

[10] E.M. Southern, U. Maskos, J.K. Elder, Analyzing and comparing nucleic acid sequences by hybridization to arrays of oligonucleotides: evaluation using experimental models, Genomics 13 (1992) 1008–1017.

[11] R. Tarjan, Depth-first search and linear graph algorithms, SIAM J. Comput. 1 (1972) 146–160.

[12] J.D. Watson, F.H.C. Crick, A structure for deoxyribose nucleic acid, Nature 173 (1953) 737–738.