MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **56** (2006) 479-484

ISSN 0340 - 6253

DNA Computing of Directed Line-Graphs

Shiying Wang and Jun Yuan*
School of Mathematical Sciences, Shanxi University, Taiyuan 030006
People's Republic of China

(Received December 2, 2005)

Abstract

Two given integers $k \geq 2$ and $1 \leq i \leq k$. A set $\mathcal{S}$ consists of all k-long oligonucleotides. All of these oligonuleotides are written 5' to 3'. A directed graph $D(k,i)$ (called DNA graph) is defined on $S \subseteq \mathcal{S}$ as follows: each oligonucleotide from S becomes a point, two points from first point to second one 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 this paper, it is obtained that a digraph is a DNA graph if and only if it is a directed line-graph. Moreover, we present an useful equivalence relation and give an efficient algorithm to find the equivalence class with respect to the equivalence relation in a DNA graph.

1 Introduction

As it is known DNA (deoxyribonucleic acid) is a double helix in which the two coiled strands (chains) are composed each of only different 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 pairs of complementary bases, which are A - T and G - C. It follows that knowing one chain, the other (complementary) can be easily reconstructed. DNA computing must not be confused with biocomputing. Usually, biocomputing means everything that computer scientists can do to help biologists to study genes. For example, algorithms and data structures have been developed to investigate the properties of the sequences of nucleotides in DNA or RNA, and those of amino acids in the primary structure of a protein. In DNA computing, instead, molecular biology is suggested to solve a problem for computer scientists. There are many reasons to investigate DNA computing. As known, Hamilton Path Problem is an NP-Complete one. Adleman's experiment[1] showed that DNA can be used to solve the hamiltonian path problem and bio-steps are O(n), where n is the number of the points of the directed graph. In 1997, Qi Ouyang et al[2] gave the DNA solution of the Maximal Clique Problem. In 2000, Qinghua Liu et al[3] presented the DNA computing on surfaces. In 2002, Adleman's group[4] solved an instance of 20-variable 3-SAT(satisfiability) problem experimentally, and this is the largest instance of the SAT problem solved experimentally by DNA computing. In 2002, Yachun Liu et al[5] proposed the DNA solution of a graph coloring problem. In 2002, Shiying Wang[6]

*This work is supported by Shanxi Province Science Foundation(20041002) and the National Natural Science Foundation of China (10471081)

gave the DNA computing of bipartite graphs for maximum matching. In 2003, G. Rozenberget et al[7] published the DNA computing by blocking. In 2005, Shiying Wang et al[8] give the DNA solution of integer linear programming. A class of graphs is called directed line-graphs. It and DNA graphs are close. In this paper, it is obtained that a digraph is a DNA graph if and only if it is a directed line-graph. Moreover, we present an useful equivalence relation and give an efficient algorithm to find the equivalence class with respect to the equivalence relation in a DNA graph.

2 DNA Graph

A DNA sequence in molecular biology is a sequence of characters from the set {A, T, G, C}. One of the methods of recognition of the primary structure of DNA sequences is sequencing by hybridization. This method consists of two phases: biochemical and computational. In its computational phase, the first approach to this problem based on graph theory, has been described by Blazewicz et al[9]. A directed graph is built from *Spectrum* (a set of some l-long oligonucleotides) as follows: each oligonucleotide from *Spectrum* becomes a point, two points are connected by an arc if the $(l-1)$ rightmost nucleotides of the first point overlap with the $(l-1)$ leftmost nucleotides of the second one. For more information on the biological background, we refer the reader to Blazewicz et al[9]. On the other hand, there are many NP-complete problems in graph theory. In DNA computing, the points of a graph must be labeled by DNA sequences. If the points of a graph are just oligonucleotides, then it is convenient in DNA computing. Thus we give the following.

Definition 2.1. Two given integers $k \geq 2$ and $1 \leq i \leq k$. A set $\mathcal{S}$ consists of all k-long oligonucleotides. All of the oligonuleotides are written 5' to 3'. A directed graph $D(k,i)$ (called *DNA graph*) is defined on $S \subseteq \mathcal{S}$ as follows: each oligonucleotide from S becomes a point, two points from first point to second one are connected by an arc if the i rightmost nucleotides of the first point overlap with the i leftmost nucleotides of the second one.

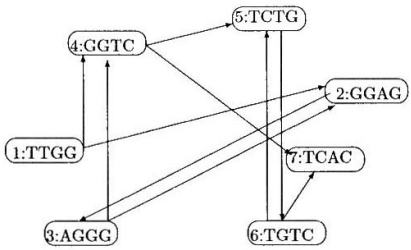

Figure 2.1

Definition 2.1 has extended the definition of DNA graphs in [9]. For example, let S=[TTGG,GGAG, AGGG,GGTC,TCTG,TGTC,TCAC]. Then the DNA graph $D(4,3)$ which is exactly the DNA graph in

[9] is an empty digraph and the DNA graph $D(4,2)$ is shown in Figure 2.1. In fact, $D(4,2)$ is the digraph in which Adleman's experiment solves the hamiltonian path problem [1]. The points of a DNA graph are oligonucleotides, i.e., the points are labeled by DNA sequences, and the arcs are determined by the points. Therefore DNA graphs are better in DNA computing.

A directed graph, or digraph, D consists of a set of points $V(D)$ and a set of ordered pairs of points $E(D)$ called arcs. For $x \in V(D)$, let $\Gamma^+(x) = \{y \in V(D) : (x,y) \in E(D)\}$.

Definition 2.2. A digraph is a *p-graph* if given any ordered pair x, y of points (x possibly equal to y), there are at most p parallel arcs from x to y.

By the definition of DNA graphs, a DNA graph is a 1-graph.

Definition 2.3. 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 *$(k,i;\alpha)$-labeled* if it is possible to assign a label $(l_1(x), \cdots, l_k(x))$ of length k to each point x of D such that

(a) $l_j(x) \in \{1, \cdots, \alpha\}$, for every $j \in \{1, \cdots, k\}$;

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

Definition 2.4. Given three integers $k \geq 2, 1 \leq i \leq k$ and $\alpha \geq 1$, $\mathcal{S}^{\alpha}_{k,i}$ is the class of 1-graphs that can be $(k,i;\alpha)$-labeled.

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

Definition 2.5. A digraph D is a *DNA-graph* if and only if there are k, i ($k \geq 2, 1 \leq i \leq k$) such that $D \in \mathcal{S}^4_{k,i}$.

DNA graphs are defined by the biological background. Its points are oligonucleotides, and its arcs are determined by the points. The complete properties of two oligonucleotides corresponding to the same DNA sequences may be different. Therefore the labels of two different points of a DNA graph allow to be the same. Definition 2.3 is consistent with the concept of DNA graphs (Definition 2.1).

Definition 2.6. The *directed line-graph* D' of a digraph D is the 1-graph with point set $E(D)$ and such that there is an arc from a point x to a point y in D' if and only if the head of the arc x in D is the tail of the arc y in D.

A digraph D' is a directed line-graph if there exists some digraph D such that D' is the directed line-graph of D.

Let s be an oligonucleotide of length $2r$. Decompose s into two strands, each of length r, $s = s_1 s_2$. Thus s_1 (resp. s_2) can be viewed as the first (resp. second) half of s.

The notations and definitions not defined here can be found in [10].

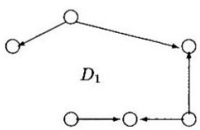

Figure 2.2

A semiwalk in a digraph is a finite non-null sequence $v_0, a_1, v_1, \cdots, a_k, v_k$, whose terms are alternately points and arcs, such that, for $i = 1, 2, \cdots, k$, the arc a_i has head v_i and tail v_{i-1} or head v_{i-1} and tail v_i. A semitrail is a semiwalk whose arcs are distinct. A semitrail is called *a friend trail* if its each internal point satisfies outdegree=2 or indegree=2. A friend trail is closed if it starts and terminates at a same point. For example, D_1 of Figure 2.2 is a friend trail.

For every two $e_1, e_2 \in E(D)$, we define a relation $e_1 \sim e_2$ when e_1 and e_2 have the same head or tail or are in a closed friend trail of length 4. Clearly, we have the following.

(a) $e \sim e$ for any $e \in E(D)$;

(b) $e_1 \sim e_2 \Longrightarrow e_2 \sim e_1$.

Theorem 2.7. "$\sim$" is an equivalence relation if and only if the following holds

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

Proof. *(only if).* Suppose that "$\sim$" is an equivalence relation. Assume, on the contrary, that there exists a pair $x, y \in V(D)$ such that $\Gamma^+(x) \cap \Gamma^+(y) \neq \emptyset$ and $\Gamma^+(x) \neq \Gamma^+(y)$. Without loss of generality, let $\Gamma^+(x) - \Gamma^+(y) \neq \emptyset$. Let $x' \in \Gamma^+(x) - \Gamma^+(y)$ and $y' \in \Gamma^+(x) \cap \Gamma^+(y)$. $(x, x') \sim (x, y')$ and $(x, y') \sim (y, y')$, therefore $(x, x') \sim (y, y')$. But arcs $(x, x'), (y, y')$ do not have the same head, tail and they are not in a closed friend trail of length 4. This is a contradiction.

(if). Suppose that $\Gamma^+(x) \cap \Gamma^+(y) \neq \emptyset \Longrightarrow \Gamma^+(x) = \Gamma^+(y)$. Suppose that $(x_1, x_1') \sim (x_2, x_2')$ and $(x_2, x_2') \sim (x_3, x_3')$, where $(x_1, x_1'), (x_2, x_2'), (x_3, x_3') \in E(D)$. There are the following cases to treat.

Case 1: (x_1, x_1') and (x_2, x_2') $(x_1 = x_2)$ have the same tail.

If (x_1, x_2') and (x_3, x_3') $(x_1 = x_3)$ have the same tail, then (x_1, x_1') and (x_3, x_3') have the same tail. If (x_1, x_2') and (x_3, x_3') $(x_2' = x_3')$ have the same head, then by hypothesis ($\Gamma^+(x) \cap \Gamma^+(y) \neq \emptyset \Longrightarrow \Gamma^+(x) = \Gamma^+(y)$.) there exists an arc (x_3, x_1') and so (x_1, x_1') and (x_3, x_3') are in a closed friend trail of length 4. If (x_1, x_2') and (x_3, x_3') are in a closed friend trail of length 4, then by hypothesis (x_1, x_1') and (x_3, x_3') are in a closed friend trail of length 4.

Case 2: (x_1, x_1') and (x_2, x_2') $(x_1' = x_2')$ have the same head.

If (x_2, x_1') and (x_3, x_3') $(x_1' = x_3')$ have the same head, then (x_1, x_1') and (x_3, x_3') have the same head. If (x_2, x_1') and (x_3, x_3') $(x_2 = x_3)$ have the same tail, then by hypothesis (x_1, x_1') and (x_3, x_3') are in a closed friend trail of length 4. If (x_2, x_1') and (x_3, x_3') are in a closed friend trail of length 4, then by hypothesis (x_1, x_1') and (x_3, x_3') are in a closed friend trail of length 4.

Case 3: (x_1, x_1') and (x_2, x_2') are in a closed friend trail of length 4.

If (x_2, x_2') and (x_3, x_3') $(x_3 = x_2)$ have the same tail, then by hypothesis (x_1, x_1') and (x_3, x_3') are in a closed friend trail of length 4. If (x_2, x_2') and (x_3, x_3') $(x_3' = x_2')$ have the same head, then by hypothesis (x_1, x_1') and (x_3, x_3') are in a closed friend trail of length 4. If (x_2, x_2') and (x_3, x_3') are in a closed friend trail of length 4, then by hypothesis (x_1, x_1') and (x_3, x_3') are in a closed friend trail of length 4. Therefore " $\sim$ " is an equivalence relation on $E(D)$. The proof is complete. □

Theorem 2.8. A 1-graph D is a DNA graph if and only if the following holds for any pair $x, y \in V(D)$:

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

Proof. *(only if).* Suppose that D is a DNA graph and $\Gamma^+(x) \cap \Gamma^+(y) \neq \emptyset$. Then there are k, i $(k \geq 2, 1 \leq i \leq k)$ such that $D \in \mathcal{S}^4_{k,i}$. Since $\Gamma^+(x) \cap \Gamma^+(y) \neq \emptyset$, there exists a point $z \in \Gamma^+(x) \cap \Gamma^+(y)$. Therefore $(l_{k-i+1}(x), \cdots, l_k(x)) = (l_1(z), \cdots, l_i(z))$ and $(l_{k-i+1}(y), \cdots, l_k(y)) = (l_1(z), \cdots, l_i(z))$.

Without loss of generality, let $w \in \Gamma^+(x)$ and $w \notin \Gamma^+(y)$. Then $(l_{k-i+1}(x), \cdots, l_k(x)) = (l_1(w), \cdots, l_i(w))$ and hence $(l_{k-i+1}(y), \cdots, l_k(y)) = (l_1(w), \cdots, l_i(w))$. By definition of the DNA graph, (y, w) is an arc, a contradiction.

(if). Let the following hold for any pair $x, y \in V(D)$:

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

By Theorem 2.7, there exists a partition of $E(D)$ into nonempty subsets $E_1, E_2, \cdots, E_n$. Let n_1 denote the number of isolated points and n_2 denote the number of points which only have out-degrees or in-degrees in D. Choose that 4^m is greater than or equal to $n + 2n_1 + n_2$. In an equivalence class, we assign a label $(l_1, \cdots, l_m)$ to the second half of the label of the tail of each arc and the first half of the label of the head of each arc, where $l_i \in \{A,T,G,C\}$ $(1 \leq i \leq m)$. In different equivalence classes, we assign different labels. For the points which only have out-degrees, assign a different label $(l_1, \cdots, l_m)$ to the first half of the label of each point; for the points which only have in-degrees, assign a different label $(l_1, \cdots, l_m)$ to the second half of the label of each point. For the isolated points, assign a different label $(l_1, \cdots, l_m)$ to the first half and the second half of the label of each point, respectively.

Every arc in D belongs to an equivalence class. By the above label, $(x, y) \in E(D) \Longrightarrow$ the second half of the label of x is equal to the first half of the label of y. Suppose that the second half of the label of x is equal to the first half of the label of y and $(x, y) \notin E(D)$. Then by the above label, $deg^+(x) \neq 0$ and $deg^-(y) \neq 0$. Let $(x, x'), (y', y) \in E(D)$. Then by the above label (x, x') and (y', y) belong to an equivalence class. (x, x') and (y', y) do not have the same head or tail because $(x, y) \notin E(D)$. Clearly, (x, x') and (y', y) are not in a closed friend trail of length 4. This is a contradiction. Therefore, if the second half of the label of x is equal to the first half of the label of y, then $(x, y) \in E(D)$. Thus, there exists a $(2m, m; 4)$-labeling of D. The proof is complete. □

Algorithm 2.9.

Let D be a directed graph. Input $E(D)$. Output $S = \{E_1, \cdots, E_n\}$.

Step 0. $S := \emptyset$, $i := 1$ and $E := E(D)$.

Step 1. If $E = \emptyset$, then stop.

Step 2. (Find E_i.)

(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 of which is the same to the head of a given arc. $F^{(2)}$ is also an arc subset of $E(D)$, the tail of each arc of which is the same to the tail of a given arc.)

(2) For any $e \in F^{(2)}$, find $F^{(1)}$ and obtain $\bigcup F^{(1)}$ and $F^{(1)} := \bigcup F^{(1)}$.

(3) $E_i := F^{(1)} \cup F^{(2)}$, $S := \{E_1, \cdots, E_i\}$, $E := E - E_i, i := i + 1$ and return Step 1.

It is easy to verify that the algorithm 2.9 is a polynomial-time one. We have the following.

Theorem 2.10. The output $S = \{E_1, \cdots, E_n\}$ of Algorithm 2.9 is a partition of $E(D)$ for a given DNA graph D. Moreover, for any $i \in \{1, \cdots, n\}$, E_i is an equivalence class under the equivalence relation "$\sim$" mentioned in Theorem 2.7.

Proof. Clearly, it is sufficient to prove that E_i is an equivalence class under the equivalence relation "$\sim$". Let e_i be the arc chosen in Step 2 (0) of Algorithm 2.9 for E_i, $i = 1, \cdots, n$. It is easy to see that $e_j \notin E_i$ if $j > i$.

Let $e \in E_i$. If e and e_i have the same tail or head, then it follows that $e_i \sim e$ from the definition of $\sim$. Suppose there exists an arc, say e', such that e_i and e' have the same tail, and e and e' have the same

head. Since $\Gamma^+(x) \cap \Gamma^+(y) \neq \emptyset \Rightarrow \Gamma^+(x) = \Gamma^+(y)$, we know that e and e_i are in a closed friend trail of length 4, which implies $e_i \sim e$. Therefore every pair of arcs in E_i is equivalent.

Without loss of generality, assume $j > i$. Suppose that $e_i \sim e_j$. If e_i and e_j have the same tail or head, then $e_j \in E_i$ by Algorithm 2.9, a contradiction. Hence, e_i and e_j are in a closed friend trail of length 4. Then there exists an arc e such that e_i and e have the same tail, and e_j and e have the same head. This implies $e_j \in E_i$, a contradiction again. Therefore, e_i and e_j are not equivalent and so $e^{'}$ and $e^{''}$ are not equivalent for any $e^{'} \in E_i, e^{''} \in E_j$. The proof is complete. □

Theorem 2.11(Berge [9]). A 1-graph D is a directed line-graph if and only if the following holds for any pair $x, y \in V(D)$:

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

By Theorem 2.8 and 2.11, we have the following.

Corollary 2.12. A digraph is a DNA graph if and only if it is a directed line-graph.

Acknowledgements

The authors would like to thank Prof. Guo Xiaofeng for many useful suggestions. The authors are appreciative of the suggestions given by the referee.

References

[1] Leonard M. Adleman, *Molecular computation of solutions to combinatorial problems*, Science 266(1994) 1021-1024.

[2] Qi Ouyang, Peter D. Kaplan, Shumao Liu and Albert Libchaber, *DNA solution of the maximal clique problem*, Science 278(1997) 446-449.

[3] Qinghua Liu, Liman Wang, Anthony G. Frutos, Anne E. Condon, Robert M. Corn, Lloyd M. Smith, *DNA computing on surfaces*, Nature 403(2000) 175-179

[4] Ravinderjit S. Braich, Nickolas Chelyapov, Cliff Johnson, Paul W. K. Rothemund, Leonard Adleman, *Solution of a 20-variable 3-SAT problem on a DNA computer*, Science 296(2002) 499-502.

[5] Yachun Liu, Jin Xu, Linqiang Pan and Shiying Wang, *DNA solution of a graph coloring problem*, Journal of Chemical Information and Computer Sciences 42(2002) 524-528.

[6] Shiying Wang, *DNA computing of bipartite graphs for maximum matching*, Journal of Mathematical Chemistry 31(3)(2002) 271-279.

[7] G. Rozenberg, H. Spaink, *DNA computing by blocking*, Theoretical Computer Science 292(2003) 653-665.

[8] Shiying Wang, Aiming Yang, *DNA solution of integer linear programming*, Applied Mathematics and Computation 170(1)(2005) 626-632.

[9] J. Blazewicz, A. Hertz, D. Kobler and D. de Werra , *On some properties of DNA graphs*, Discrete Applied Mathematics 98 (1999) 1-19.

[10] L. Lovasz and M. D. Plummer, *Matching Theory*, Elsevier Science Publishing Company, Inc., New York, 1986.